Towards Robust and Composable Disentangled Representations for Music and Speech

Yin-Jyun Luo

Supervised by Simon Dixon and Sebastian Ewert

School of Electronic Engineering and Computer Science

September 2025

Submitted in partial fulfillment of the requirements
of the Degree of Doctor of Philosophy

This work was supported by Spotify and the UKRI CDT in AI and Music under grant EP/S022694/1

Statement of Originality

I, Yin-Jyun Luo, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by other, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that Queen Mary University of London has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature: Yin-Jyun Luo

Date: September 1 2025

Details of collaboration and publications:
Each core chapter of this thesis corresponds to a peer-reviewed publication led by the author. The main research tasks—including problem formulation, model design, implementation, experiments, and manuscript drafting—were primarily carried out by the author. For Chapters 3, 4, and 5, Sebastian Ewert contributed technical feedback and architectural suggestions, while Simon Dixon provided overall guidance and sup-

port in shaping the research direction. Both also assisted in revising the manuscripts. Chapter 6 was completed during an industrial internship, where the author was responsible for the core research and implementation. Other co-authors contributed through topic framing, methodological input, and feedback on evaluation and writing. In addition, ChatGPT was used as a proofreading tool to improve grammar and clarity; all ideas, methods, and intellectual contributions remain the author's own responsibility.

1. Chapter 3 is based on: Y.-J. Luo, S. Ewert, and S. Dixon (2022). "Towards Robust Unsupervised Disentanglement of Sequential Data – A Case Study Using Music Audio". In: *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*.

2. Chapter 4 is based on: Y.-J. Luo, S. Ewert, and S. Dixon (2024). "Unsupervised Pitch–Timbre Disentanglement of Musical Instruments Using a Jacobian Disentangled Sequential Autoencoder". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

3. Chapter 5 is based on: Y.-J. Luo and S. Dixon (2024). "Posterior Variance-Parameterised Gaussian Dropout: Improving Disentangled Sequential Autoencoders for Zero-Shot Voice Conversion". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

4. Chapter 6 is based on: Y.-J. Luo, K. W. Cheuk, W. Choi, W.-H. Liao, K. Toyama, T. Uesaka, K. Saito, C.-H. Lai, Y. Takida, S. Dixon, and Y. Mitsufuji (2024). "Disentangling Multi-Instrument Music Audio for Source-Level Pitch and Timbre Manipulation". In: *Proceedings of the NeurIPS 2024 Workshop on Audio Imagination: AI-Driven Speech, Music, and Sound Generation*. (This paper underwent a peer review process organised by the workshop programme committee, although NeurIPS workshops are non-archival and are not included in the main conference proceedings.)

Abstract

Disentangled representation learning seeks to uncover latent factors of data that are distinct and semantically significant, thereby facilitating controllable generation of novel data. In sequential domains such as music and speech, disentangled sequential autoencoders (DSAEs) model each sequence with dynamic, frame-level latents and a global, sequence-level latent. This design separates local and global generative attributes—such as pitch and timbre in music, or content and speaker identity in speech. However, standard DSAEs are prone to posterior collapse of the global latent, with performance sensitive to architecture and hyperparameters.

This thesis introduces methods to improve the robustness of unsupervised disentanglement in DSAEs. First, a two-stage framework (TS-DSAE) is proposed: it learns informative prior distributions for global factors, then encourages disentanglement between global and local factors through auxiliary regularisations. This mitigates collapse without adversarial losses or data augmentation. To balance reconstruction quality with disentanglement, a teacher–student model (J-DSAE) extends this framework by regularising output variation with respect to changes in the global latent. The effectiveness of both TS-DSAE and J-DSAE is demonstrated on pitch–timbre disentanglement for monophonic instruments, although both approaches are in principle applicable to other sequential settings.

As an alternative to auxiliary objectives, Gaussian dropout is applied to local latents. Here, the variance of the Gaussian noise is parameterised by that of the global latent, dynamically constraining the information capacity of local latents. This method is evaluated on zero-shot voice conversion of speech, yet its general formulation does not rely on speech-specific assumptions.

Finally, a preliminary study addresses disentanglement of multi-instrument mixtures using a supervised approach. Each instrument is represented by a source-level latent, which further disentangles source-specific pitch and timbre. This hierarchical and modular design enables synthesis of mixtures with arbitrary pitch–timbre combinations.

Together, these contributions advance robust unsupervised disentanglement of sequential data, with music and speech serving as case studies, and lay the foundation for composable representation learning of multi-instrument mixtures.

Acknowledgements

It has been quite a journey! First and foremost, this thesis would not have been possible without the support of my supervisors, Simon and Sebastian. I was lucky to be given complete intellectual freedom, able to fully dedicate myself to my own research proposal. I'm grateful to Simon for being the North Star who always helped me reorient when I was stuck and had lost sight of the big picture. I cannot thank Sebastian enough for his insight—from low-level technical advice to research philosophy: don't just throw random darts at a wall and see which sticks. I felt fortunate to be able to speak openly with both of them when I was frustrated or second-guessing myself. They always cleared the fog and showed me a valid path forward.

My exploration of disentangled representation learning began when I first enrolled as a PhD student under Dorien and Kat at the Singapore University of Technology and Design (SUTD), before transitioning to Queen Mary University of London (QMUL). They also gave me the intellectual freedom to pursue my curiosity, and their steadfast support nudged me to where I am today—a position I could only have dreamed of back then. Leaving SUTD for QMUL was a tough call, and I couldn't have done it without their backing. Emmanouil, who later became my independent assessor at QMUL, gave me tremendous feedback when I applied for the PhD programme, helping me refine my original proposal, and continued to provide insightful advice throughout my journey.

Before embarking on my PhD, I was lucky to develop research independently as a Research Assistant at Academia Sinica. My supervisor, Li, laid the foundation for my research skills and introduced me to the community of music information retrieval (MIR). I also have to thank Tai-Shih, my Master's supervisor, for letting me

freely explore MIR, which gave me the confidence to pursue research and fuelled my curiosity. I'm equally grateful to Yi-Hsuan, who introduced me to Li and facilitated our collaboration. I learnt a great deal from my Master's thesis committee, led by Tai-Shih, including Li and Yi-Hsuan. Their feedback and support shaped my career path in ways I couldn't be happier about.

I'm also grateful to my mentors during internships. At Stability AI, Xubo gave me the opportunity to work on industry-relevant topics, which made me much better prepared for job hunting. Also at Stability, Julian was an inspiring mentor with a unique perspective. He's a researcher, a manager, a friend—and I've always loved his humour.

At Sony AI, Kin Wai—also my closest colleague back at SUTD—inspired me deeply with his disciplined and persistent approach to problem-solving, which I admire greatly. Woosung and Wei-Hsiang consistently offered constructive feedback, always raising great first-principles questions. My internship wouldn't have been complete without Koichi and my peers Ayano, Junyoung, Shang-Fu, and Yixiao. Working (and hanging out) with them was so much fun, and I especially enjoyed our philosophical chats, which helped me get through tough times.

To my colleagues and friends at QMUL: I admire Ben, Chin-Yun, and Christopher for their relentless passion and determination to tackle difficult questions. I was impressed by Jinhua's adaptability in pivoting research topics in this volatile, fast-moving field. I also appreciate Nicolas for his creative and thoughtful perspective on research. Each of their traits has influenced how I approach ideas and problems.

Special thanks to friends outside QMUL who added colour to my PhD journey: Keitaro, with whom I had countless spontaneous research discussions, and who shared my enthusiasm for disentangled representation learning. My internship in Japan was even more joyful thanks to him. Maciek made the transition from SUTD to QMUL so much easier, and I can't thank him enough for the fun times and the memories we created throughout our PhD journey. Although I never had the chance to know Wei-Ning personally, as a fellow Taiwanese his work "Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data" profoundly shaped my research taste and inspired me to set out on the journey of unsupervised

disentangled representation learning. His contribution has been a constant source of motivation, and I owe much of my own direction to the path he opened.

I wouldn't have discovered MIR—or switched fields—if it weren't for my brother, Yin-Tsung. This whole journey was possible only because he planted the seed of intellectual curiosity and adventurous spirit in me, pointing me towards MIR before I even started my Master's. And of course, to my beloved mother, Shu-Min—your unconditional love and support have been the bedrock of all my achievements. My partner, Sheau-Jwu, has been the best companion I could ask for, always there with emotional support through the ebbs and flows.

Finally, this work was generously sponsored by Spotify. Thank you for your support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Modern machine learning systems for audio aim not only to analyse signals, but also to manipulate and generate them in a controlled and interpretable manner. In domains such as music and speech, this requires the ability to isolate and recombine meaningful attributes of the signal.

In music, a fundamental challenge is to modify specific aspects of a recording—such as the instrument timbre or melodic content—without affecting others. For instance, one may wish to render a melody played by a violin using the timbre of a trumpet, or recombine elements from multiple recordings to create novel musical material. Beyond note-level transformations, such capabilities form a foundation for higher-level operations, including the manipulation of motifs, phrases, and structural patterns in musical compositions.

Similarly, in speech processing, tasks such as voice conversion require altering speaker identity while preserving the underlying linguistic content. Achieving this reliably, particularly in zero-shot settings where speakers are unseen during training, remains a significant challenge.

Despite their differences, these problems share a common structure: the underlying signal is governed by multiple factors of variation that operate at different temporal scales. Attributes such as instrument timbre or speaker identity are relatively stable

over time, while others such as pitch or linguistic content vary rapidly. The ability to separate and control these factors is therefore central to both applications.

These challenges can be addressed through representation learning (Bengio et al., 2013a), which seeks to transform high-dimensional raw observations, such as images and audio waveforms, into a low-dimensional, compact latent space. The objective is to develop a neural network capable of learning these representations, thereby facilitating downstream tasks such as classification and generation. Classification tasks aim to predict specific targets of interest given the representation, whereas generative tasks focus on sampling realistic data from, and manipulating particular factors of variation within, the latent space.

In order to address these challenges effectively, it is important to learn representations that capture perceptually salient and semantically meaningful factors of variation; for example, the shape and size of an object in an image or the vocal characteristics and spoken content in a speech utterance. Classification models stack multiple layers of neural networks such that layers further away from the input data capture semantic features that are invariant to low-level details. For generative tasks, capturing these semantically meaningful features allows one to sample novel data with control over them. For instance, with separate representations encoding speaker identity and spoken content in a factorised manner, one can fix the representation of the spoken content while replacing that of the speaker identity within the latent space, thereby achieving voice conversion by conditioning a neural network decoder on the two representations. Factorisation and invariance to other factors of variation are key properties that facilitate downstream tasks, which are the main goal of disentangled representation learning (DRL) (X. Wang et al., 2024).

DRL aims to learn representations that are structured in a way that isolates each factor of variation. Using the running example of speaker–content representations, given a common text transcription spoken by two speakers, disentangled representations extracted from the two utterances would ideally differ only in the speaker feature and share the same content feature. Conversely, having a person speak two different text scripts should result in a common speaker feature and different content features. For discriminative tasks, such representations are expected to be robust to

spurious correlations between factors caused by biased sampling of available training data. For generative tasks, they allow for explicit control over the disentangled factors and could potentially generalise to factor combinations that are unseen during training.

DRL underpins compositionality (Greff et al., 2020): the ability to encode sensory data using compact and reusable concepts, and to synthesise new data from these concepts. In a visual scene composed of multiple physical objects, simply learning factorised representations of attributes such as shape and size is insufficient to disambiguate which attributes belong to which object. An additional level of abstraction that captures the concept of an object is required to resolve this ambiguity. Similarly, in an audio recording of multiple musical instruments, each instrument must first be bound to an entity, which can then be described by independent pitch and timbre components, in order to disambiguate which attributes belong to which instrument. By learning disentangled representations across different levels of abstraction, where attribute descriptors such as pitch and timbre are shared by entities such as instruments relevant to higher-level cognition, a model can compose novel visual or acoustic scenes of multiple objects, each with their own attributes. This leads to composable representations, enabling generalisation to novel combinations of conceptual representations.

Inductive biases are crucial to learning structured latent representations (Tschannen et al., 2018; Locatello et al., 2019b). In sequential domains such as music and speech, realising such structured and compositional representations requires inductive biases that reflect the temporal organisation of the data. To address disentanglement in these settings, this thesis builds upon the disentangled sequential autoencoder (DSAE) framework (Y. Li et al., 2018; Han et al., 2021; Bai et al., 2021; Naiman et al., 2023; Berman et al., 2024; Simon et al., 2024), which introduces an architectural bias reflecting global–local factor separation over time. The underlying factors of variation in sequential data typically operate at different rates of change along the time axis. In speech, one can assume that the vocal characteristics of a speaker are fixed, while the spoken content changes frequently over time. For single-instrument audio recordings, the overall timbre of the instrument remains largely unchanged,

while the played melody evolves frame by frame. This prior knowledge can be used to inform the architectural design of a model to disentangle these factors operating at distinct rates of change. For example, a simple architectural bias is to configure a time-invariant vector to represent the static factor (e.g., speaker or instrument identity), while a time-varying sequence of vectors encodes the dynamic factor (e.g., linguistic content or melody). While this inductive bias provides a principled structure for representation learning, it does not by itself guarantee that the model will utilise the global latent effectively during training.

Despite the simplicity of this global–local formulation, disentanglement remains inconsistent in practice due to a lack of explicit incentives for the model to allocate information to the global latent. This imbalance is largely attributed to the disparity in information capacity: the global representation, being a single vector, is often overshadowed by the temporally rich local representation. As a result, the global latent tends to become underutilised, a phenomenon commonly referred to as posterior collapse (Y. Wang et al., 2021; X. Chen et al., 2017; Alemi et al., 2018; Bowman et al., 2016; Serban et al., 2017; Fraccaro et al., 2016). More broadly, unsupervised approaches to disentanglement are fragile and highly sensitive to configurations of hyperparameters (Locatello et al., 2019b).

To counteract the collapse of the global latent in the DSAE framework, prior work has introduced additional constraints informed by domain-specific knowledge. One prominent strategy is to construct contrastive pairs—such that two data examples share or differ in a specific factor—and impose similarity or dissimilarity objectives on their latent representations (Zhu et al., 2020; Bai et al., 2021; Naiman et al., 2023). These methods can indeed encourage the global latent to encode discriminative information. However, they rely on modality-specific transformations or heuristics to create positive and negative pairs. For example, in speech, it is possible to apply voice conversion to manipulate speaker identity while keeping content fixed (Bai et al., 2021), but different transformations are required to modify timbre while preserving pitch in musical audio. Moreover, these additional objectives introduce further hyperparameters and training complexity, often requiring dataset-specific tuning.

Taken together, these limitations highlight the need for methods that can reli-

ably learn disentangled representations in sequential data, while remaining robust to model configuration and applicable across domains. This leads to the central research question of the thesis: *How can robust disentanglement of global and local factors in sequential data be achieved without relying on modality-specific priors or supervision?* Here, robustness is defined as the ability to consistently obtain disentangled representations across a broad range of architectures and hyperparameter settings.

This thesis addresses this question through the lens of DSAEs, using pitch–timbre disentanglement (PTD) in music and voice conversion (VC) in speech as primary case studies. While most existing work on PTD is limited to single-instrument recordings (Engel et al., 2017; Bitton et al., 2018; Luo et al., 2019; Y. Wu et al., 2023; Esling et al., 2018; Luo et al., 2020; Tanaka et al., 2021; Cífka et al., 2021; Tanaka et al., 2022; Tanaka et al., 2025; Demerlé et al., 2024), this thesis further explores a more common and complex setting—mixtures of multiple instruments playing distinct pitches. This scenario introduces a two-level disentanglement problem: first separating the sources, and then factorising each into pitch and timbre components. Such a setup enables compositional representations, where instrument identities and musical content can be independently manipulated or recombined to form novel mixtures. In this thesis, this direction is explored through a supervised framework in Chapter 6, complementing the unsupervised approaches developed for single-instrument settings in Chapters 3 to 5.

In summary, a main part of this thesis develops unsupervised, robust, and generalisable frameworks for disentangling latent representations from sequential audio data, and separately scales the task of PTD to tackle the case of multiple-instrument mixtures using a supervised method. It addresses the limitations of current approaches in both unsupervised and supervised settings and demonstrates their effectiveness across a variety of music and speech tasks.

1.2 Contributions

The main contributions of this thesis are as follows:

1. **A robust two-stage training strategy for unsupervised disentangle-**

ment of sequential data. This thesis introduces a general two-stage training strategy (TS-DSAE) to address the problem of posterior collapse in DSAEs. The strategy separates the disentanglement of the global latent from the reconstruction of sequential data, thereby mitigating posterior collapse without introducing adversarial objectives (which add additional optimisation challenges) or modality-specific data augmentation (which limits generalisability across domains). Using music audio as a case study, the method demonstrates consistent disentanglement performance across a range of hyperparameters, latent dimensionalities, and model architectures, establishing a robust foundation for unsupervised representation learning of pitch and timbre.

2. **Jacobian supervision for mitigating the trade-off of reconstruction and disentanglement.** A teacher–student framework (J-DSAE) is proposed to address the trade-off between reconstruction quality and disentanglement. Building on TS-DSAE, outputs of the model from the first training stage—expected to capture disentangled representations—are used to constrain the outputs of the model in the second stage. Unlike prior methods, the regularisation is imposed in the observation space rather than the latent space. This approach is shown to improve audio fidelity while preserving the separation of pitch and timbre factors in the task of unsupervised PTD for musical instrument audio.

3. **Posterior variance-parameterised Gaussian dropout for balancing latent information.** Orthogonal to TS-DSAE and J-DSAE, a simple yet effective regularisation technique—posterior variance-parameterised Gaussian dropout (pvpGD)—is introduced to dynamically adjust the capacity of the local latent space based on the uncertainty of the global latent. Applied to a range of existing voice conversion models that disentangle speaker identity and linguistic content, this technique consistently improves global latent utilisation without requiring supervision or additional hyperparameter tuning.

4. **DisMix: a supervised framework for pitch–timbre disentanglement of musical mixtures.** To extend PTD beyond single-instrument recordings,

DisMix is introduced as a two-level disentanglement framework for learning per-source pitch and timbre representations from multi-instrument mixtures. DisMix supports modular manipulation and composition by treating pitch and timbre as independently controllable attributes for each source. Two implementations are proposed: a lightweight autoencoder serving as a proof of concept, and a latent diffusion model capable of handling realistic orchestral mixtures.

5. **Empirical analysis across modalities, configurations, and tasks.** The proposed methods are evaluated across synthetic and real-world music and speech datasets, covering a diverse range of model architectures, latent dimensionalities, and hyperparameter settings. The analysis identifies key factors influencing disentanglement robustness, demonstrates the effectiveness of modality-agnostic, adversarial-training-free regularisation strategies, and showcases applications enabled by composable latent representations.

1.3 Overview

The remainder of this thesis is structured as follows. Chapters 3 to 6 are based on peer-reviewed publications, each addressing a specific technical challenge in learning robust and composable disentangled representations for music and speech.

All core chapters are based on first-authored publications. The ideation, implementation, experimentation, and writing were led by the author of this thesis. For Chapters 3, 4, and 5, Sebastian Ewert contributed through technical feedback and architectural insights, while Simon Dixon provided high-level guidance and helped shape the research direction. Both contributed to paper revisions. Chapter 6 is the outcome of an industrial internship, where the author led the core research and implementation. All co-authors contributed through topic framing, methodological suggestions, and feedback on evaluation and writing.

- **Chapter 2** reviews relevant literature on representation learning, variational autoencoders, disentanglement in sequential data, and applications to music and speech. It provides the theoretical and empirical context that motivates the approaches developed in this thesis.

- **Chapter 3** introduces the TS-DSAE framework for robust unsupervised disentanglement, based on Luo et al. (2022). The method addresses posterior collapse in DSAEs by decoupling the learning of global and local factors and is evaluated on monophonic music datasets.

- **Chapter 4** presents the Jacobian-regularised teacher–student framework, proposed in Luo et al. (2024c). Building on TS-DSAE, the model imposes Jacobian constraints in the observation space to improve reconstruction while preserving pitch–timbre disentanglement in musical instrument recordings.

- **Chapter 5** proposes posterior variance-parameterised Gaussian dropout, introduced in Luo et al. (2024b). This technique dynamically adjusts the local latent capacity based on uncertainty in the global latent and improves speaker–content disentanglement in zero-shot voice conversion tasks.

- **Chapter 6** presents DisMix, a supervised two-level disentanglement framework for multi-instrument music mixtures, published by Luo et al. (2024a). The method supports modular manipulation and composition using per-source pitch and timbre representations, implemented via both a lightweight autoencoder and a latent diffusion model.

- **Chapter 7** concludes the thesis and outlines directions for future research, including extending compositionality to other modalities, incorporating temporal hierarchies, and developing hybrid learning strategies that combine supervised and unsupervised techniques.

Chapter 2

Background

2.1 Disentangled Representation Learning

Representation learning seeks to map high-dimensional observations, such as images and audio waveforms, to compact latent representations that capture salient structure in the data. Such representations support downstream tasks including classification and generation, where the goal is either to predict variables of interest or to synthesise realistic data by manipulating underlying factors of variation.

Disentangled representation learning (DRL) (Bengio et al., 2013a; X. Wang et al., 2024) models data through latent variables, each representing a distinct and semantically meaningful generative factor. The objective is to learn a modular latent space where each dimension corresponds to an interpretable factor.

For example, an image of a physical object may be mapped from high-dimensional pixel space to a low-dimensional representation encoding attributes such as size and shape along separate axes. More generally, the latent space may be partitioned into multi-dimensional subspaces, each capturing a salient property—for instance, pitch and timbre in music (Luo et al., 2019; Cífka et al., 2021; Demerlé et al., 2024), or speaker identity and linguistic content in speech (Hsu et al., 2017; Lian et al., 2022b; Lu et al., 2022).

In contrast, standard non-disentangled representation learning often yields *entangled* representations, in which multiple generative factors are encoded jointly in the

same latent dimensions (Bengio et al., 2013a; Higgins et al., 2017a; Locatello et al., 2019a). As a result, changes in one factor may inadvertently affect others, making the representation difficult to interpret or control. For example, in a speech representation, speaker identity and linguistic content may be mixed together, so that modifying one attribute in the latent space also alters the other. This limits both controllability for generation and robustness for downstream prediction.

A common approach to learning disentangled representations is via generative models, especially variational autoencoders (VAEs) (Higgins et al., 2017a; T. Q. Chen et al., 2018; H. Kim et al., 2018; Kumar et al., 2018). VAEs enable both unconditional generation through latent sampling and conditional generation through latent manipulation. Such representations support generalisation, interpretability, and controllability, which are crucial for building flexible and robust systems.

2.1.1 Generalisability

A key benefit of DRL is generalisability to unseen combinations of factors (Higgins et al., 2017b; Burgess et al., 2019; Locatello et al., 2019a; Creager et al., 2019; Steenkiste et al., 2019). For example, in zero-shot voice conversion (VC), detailed in Section 2.4, the goal is to convert a given speech utterance to sound as if it were spoken by a speaker not seen during training.

Zero-shot VC models typically use autoencoder (AE) architectures to disentangle speaker identity and linguistic content (Hsu et al., 2017; Qian et al., 2019; Nguyen et al., 2022; Chou et al., 2019; D.-Y. Wu et al., 2020; Tang et al., 2022; D. Wang et al., 2021; Tjandra et al., 2021; Lian et al., 2022b; Lu et al., 2022; Lian et al., 2022a; Wei et al., 2023; Lu et al., 2024). An encoder extracts both factors, and a decoder reconstructs speech conditioned on them. VC is performed by swapping speaker identity while preserving content. Whereas conventional VC methods learn speaker embeddings that are restricted to speakers only available from training data, the aforementioned models explicitly learn representations that disentangle speaker from content, thereby enabling applicability to unseen speakers.

DRL also benefits discriminative tasks such as automatic speech recognition (ASR). Similar to the aforementioned VC approaches, Hsu et al. (2018a) propose a generative

process in which speech is modelled by combining a sequence-level and a segment-level latent variable. The sequence-level latent is constrained to remain constant within an utterance and is thus expected to capture relatively stable factors such as speaker identity, room acoustics, and background noise. In contrast, the segment-level latent captures dynamic information, such as phonetic content, required for faithful reconstruction of the original speech.

In the context of ASR, the segment-level latent serves as the informative representation, while the sequence-level latent encodes nuisance factors of variation. Using the segment-level latent to train ASR models improves their robustness to non-linguistic factors such as background conditions or speaker-specific attributes.

This thesis focuses on generative applications of DRL. Literature on disentanglement of pitch and timbre is discussed in Section 2.3, and that of speaker identity and content in Section 2.4.

2.1.2 Variational Autoencoders

VAEs (Kingma et al., 2014; Rezende et al., 2014) are a central backbone of DRL (Higgins et al., 2017a; T. Q. Chen et al., 2018; H. Kim et al., 2018; Kumar et al., 2018), combining reconstruction and regularisation losses to learn structured latent spaces (Tschannen et al., 2018).

However, Locatello et al. (2019b) empirically demonstrate that disentanglement cannot be reliably achieved in the unsupervised setting without inductive biases. The theory follows that, in conventional VAEs, the latent variables are sampled from a standard Normal prior distribution. Due to the rotational invariance of this prior, multiple latent representations are equally probable under the distribution. Consequently, it is impossible to recover a unique, ground-truth set of disentangled latent factors solely by optimising the standard VAE objective.

This theoretical limitation has led to the design of models that incorporate structural constraints. For sequential data, a natural extension is to adapt the VAE framework so that it captures both temporal dependencies and factors that evolve at different rates over time. The disentangled sequential autoencoder (DSAE) (Y. Li et al., 2018) introduces such an inductive bias through an architecture that separates

static and dynamic latent variables. This structure encourages the model to represent invariant and time-varying factors in distinct subspaces. The next section reviews DSAE and its variants, which underpin the methods developed in the remainder of this thesis.

2.2 Disentangled Sequential Autoencoders

A standard approach to modelling time-series data with latent-variable models is the sequential VAE (Girin et al., 2021), in which observations are generated from latent variables while temporal dependencies are captured through autoregressive structure in the prior, encoder, decoder, or a combination of these components. Such models are well suited to sequential generation, but they do not by themselves distinguish between factors that remain relatively stable across a sequence and those that vary from one time step to the next. As a result, latent representations learnt by standard sequential VAEs tend to entangle factors that evolve at different temporal scales. Without explicit structural constraints, such models do not enforce a separation between factors operating at different temporal scales, which motivates the introduction of architectures such as the disentangled sequential autoencoder (DSAE) (Y. Li et al., 2018).

Y. Li et al. (2018) extend VAEs (Kingma et al., 2014; Rezende et al., 2014) in two ways:

1. It learns from time-series data by considering the temporal dependencies across observations at different time steps.

2. It disentangles time-varying (local) factors from time-invariant (global) factors of variation by extracting distinct latent variables to represent each type of factor.

This model retains the core advantages of the VAE framework, serving as a principled probabilistic generative model optimised via the stochastic gradient variational Bayes (SGVB) estimator (Kingma et al., 2014), and as a foundational generative framework for the unsupervised disentanglement of sequential data (Han et al., 2021; Bai et al.,

2021; Naiman et al., 2023; Berman et al., 2024; Simon et al., 2024).

While the dependencies among random variables vary across different DSAE models, depending on specific modelling choices, the original formulation (Y. Li et al., 2018) defines a joint distribution over a sequence of L observations $\mathbf{x}_{1:L}$, the local latent variables $\mathbf{z}_{1:L}$, and the global, static latent variable $\mathbf{s}$ as follows:

$$p_\theta(\mathbf{x}_{1:L}, \mathbf{z}_{1:L}, \mathbf{s}) = p(\mathbf{s}) \prod_{t=1}^{L} p_{\theta_x}(\mathbf{x}_t|\mathbf{z}_t, \mathbf{s}) p_{\theta_z}(\mathbf{z}_t|\mathbf{z}_{<t}), \tag{2.1}$$

where θ_x and θ_z denote the neural network parameters of the decoder p_{θ_x} and the local prior p_{θ_z}, respectively.

The global prior $p(\mathbf{s})$ is chosen as a standard Gaussian distribution, while the local prior is parameterised by a neural network designed to capture the temporal dependencies among the dynamic latent variables $\mathbf{z}_{1:L}$. This separation of prior structures—static for the global latent and dynamic for the local latent—encourages the disentanglement of global and local generative factors. For simplicity, when distinguishing between these parameters is unnecessary, they are collectively denoted as θ throughout the remainder of this thesis.

Importantly, the DSAE model facilitates the disentanglement of $\mathbf{z}_{1:L}$ and $\mathbf{s}$ primarily by distinguishing their temporal resolutions: there is one latent vector $\mathbf{z}_t$ for each observation $\mathbf{x}_t$ at time step t, while a single latent vector $\mathbf{s}$ is shared across the entire sequence $\mathbf{x}_{1:L}$. This distinction in temporal resolution is particularly well motivated in domains such as speech and music, where perceptually meaningful factors naturally evolve at different time scales, e.g., speaker identity or instrument timbre versus linguistic content or melody. As indicated in (2.1), for each $t \in [1, L]$, $\mathbf{x}_t$ can be reconstructed—or sampled—given the static latent $\mathbf{s}$ and the dynamic latent $\mathbf{z}_t$.

Factorised Inference

Following the VAE framework (Kingma et al., 2014; Rezende et al., 2014), the DSAE introduces separate encoders to parameterise approximate posterior distributions over

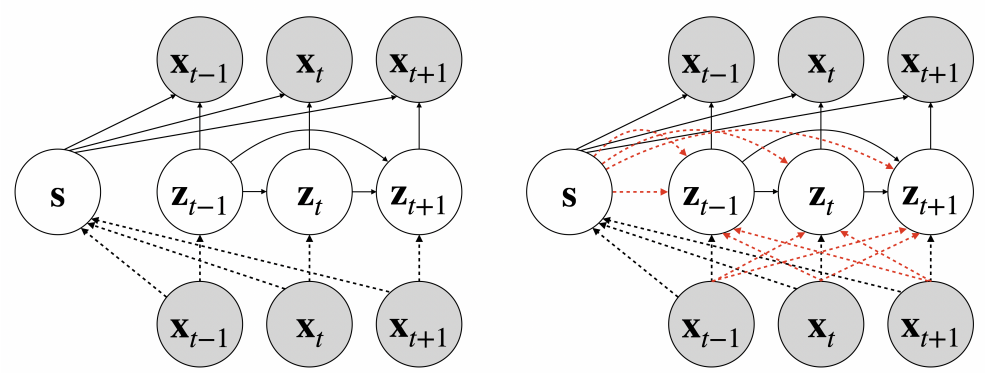

Figure 2.1: The graphical models of DSAE illustrated with three time steps. Solid and dashed lines denote generation and inference procedures, respectively. *Left*: The DSAE configured with the "factorised q" encoder. *Right*: The DSAE configured with the "full q" encoder, with additional dependency highlighted by red arrows.

the two latent variables as:

$$q_\phi(\mathbf{z}_{1:L}, \mathbf{s}|\mathbf{x}_{1:L}) = q_{\phi_s}(\mathbf{s}|\mathbf{x}_{1:L}) \prod_{t=1}^{L} q_{\phi_z}(\mathbf{z}_t|\mathbf{x}_t), \tag{2.2}$$

where ϕ_s and ϕ_z denote the parameters of the global encoder q_{ϕ_s} and the local encoder q_{ϕ_z}, respectively. Y. Li et al. (2018) refer to this factorised approximate posterior as the "factorised q" encoder.

Structured Inference

Additionally, Y. Li et al. (2018) consider a structured posterior which is referred to as the "full q" encoder:

$$q_\phi(\mathbf{z}_{1:L}, \mathbf{s}|\mathbf{x}_{1:L}) = q_{\phi_s}(\mathbf{s}|\mathbf{x}_{1:L}) \prod_{t=1}^{L} q_{\phi_z}(\mathbf{z}_t|\mathbf{x}_{1:L}, \mathbf{s}), \tag{2.3}$$

where $\mathbf{s}$ is first sampled and then used to condition the inference of $\mathbf{z}_{1:L}$. The dependence of dynamic factors on static factors is motivated by, e.g., the fact that the shape of objects could be informative of their motion. At each time step, z_t also depends on the entire sequence and benefits from the context.

Fig. 2.1 depicts the graphical representations of DSAE with either "factorised q" or "full q". As with (2.1), ϕ_s and ϕ_z will be collectively denoted as ϕ throughout the

remainder of the thesis when a distinction is unnecessary.

To enforce the difference in temporal resolution between $\mathbf{z}_{1:L}$ and $\mathbf{s}$, the DSAE imposes a hard constraint on the encoders: the global encoder q_{ϕ_s} applies temporal pooling layers to produce a single vector, while the local encoder q_{ϕ_z} preserves the full temporal resolution of the input sequence. While this architectural bias provides a principled way to separate static and dynamic factors, it does not by itself guarantee that the global latent will be used effectively during training.

Applications

Learning disentangled representations of time-varying and time-invariant factors from time-series data enables a range of applications, including controllable generation and attribute swapping of sequences.

Voice conversion (VC), for instance, aims to modify speaker characteristics while preserving spoken content. This task typically relies on a DSAE model that encodes speaker identity and linguistic information into global and local latent spaces, respectively (Hsu et al., 2017; Y. Li et al., 2018; Zhu et al., 2020; Han et al., 2021; Bai et al., 2021; Naiman et al., 2023; Berman et al., 2024; Lu et al., 2022; Lian et al., 2022b). Given a speech utterance, the encoders extract the two latent variables, and conversion is achieved by swapping the global latent variable with one extracted from a different utterance.

DSAEs have also been applied to pitch-timbre disentanglement (PTD) in musical instrument recordings, where the instrument identity corresponds to the global factor and the melody to the local factor (Tanaka et al., 2022; Tanaka et al., 2025). Analogous to VC, attribute translation can be achieved by replacing either latent variable with one extracted from another recording, thereby modifying timbre or melody while keeping the other attribute unchanged. A more comprehensive review of DSAE-based approaches for PTD and VC is provided in Sections 2.3 and 2.4, respectively.

2.2.1 Training Objective

Similar to the VAE framework (Kingma et al., 2014; Rezende et al., 2014), the DSAE (Y. Li et al., 2018) maximises:

$$\begin{aligned}\mathcal{L}_{\text{DSAE}} &= \mathbb{E}_{q_\phi(\mathbf{z}_{1:L},\mathbf{s}|\mathbf{x}_{1:L})}\big[\log p_\theta(\mathbf{x}_{1:L},\mathbf{z}_{1:L},\mathbf{s}) - \log q_\phi(\mathbf{z}_{1:L},\mathbf{s}|\mathbf{x}_{1:L})\big] \\ &= \sum_{t=1}^{L} \mathbb{E}_{q_\phi(\mathbf{z}_t|\mathbf{x}_{1:L})q_\phi(\mathbf{s}|\mathbf{x}_{1:L})}\big[\log p_\theta(\mathbf{x}_t|\mathbf{z}_t,\mathbf{s})\big] \\ &\quad - \beta_z \sum_{t=1}^{L} \mathbb{E}_{q_\phi(\mathbf{z}_{<t}|\mathbf{x}_{1:L})}\big[\mathcal{D}_{\text{KL}}\big(q_\phi(\mathbf{z}_t|\mathbf{x}_{1:L})\|p_\theta(\mathbf{z}_t|\mathbf{z}_{<t})\big)\big] \\ &\quad - \beta_s \mathcal{D}_{\text{KL}}\big(q_\phi(\mathbf{s}|\mathbf{x}_{1:L})\|p_\theta(\mathbf{s})\big).\end{aligned} \tag{2.4}$$

Here, the weights $\beta_z, \beta_s \in \mathbb{R}_+$ balance the information bottleneck between the two latent variables (Han et al., 2021; B. Vowels et al., 2021; Bai et al., 2021; Naiman et al., 2023; Berman et al., 2024; Simon et al., 2024), drawing inspiration from the β-VAE framework (Higgins et al., 2017a). Y. Li et al. (2018) set both weights to one, reducing (2.4) to a valid lower bound on the log-likelihood $\log p_\theta(\mathbf{x}_{1:L})$, commonly referred to as the evidence lower bound (ELBO). The first term corresponds to the reconstruction of $\mathbf{x}_t$ given $\mathbf{z}_{1:L}$ and $\mathbf{s}$. The last two terms regularise the two latent variables by their corresponding priors $p_\theta(\mathbf{z}_t|\mathbf{z}_{<t})$ and $p_\theta(\mathbf{s})$ through $\mathcal{D}_{\text{KL}}\big(\cdot\,\|\,\cdot\big)$, the Kullback–Leibler divergence (KLD) (Cover et al., 1991).

Parameterisation

Y. Li et al. (2018) define the global prior as a standard Gaussian distribution, introducing no additional parameters: $p(\mathbf{s}) = \mathcal{N}\big(\mathbf{0}, \mathbf{I}_{D_s}\big)$, where $\mathbf{I}_{D_s}$ denotes the identity matrix of size D_s, the dimension of the global latent space. In contrast, the local prior $p_{\theta_z}(\mathbf{z}_t|\mathbf{z}_{<t})$ captures temporal dependencies across the local latent variables and is modelled as a diagonal Gaussian distribution, whose mean $\mu_{\theta_z}(\mathbf{z}_{<t}) \in \mathbb{R}^{D_z}$ and variance $\sigma^2_{\theta_z}(\mathbf{z}_{<t}) \in \mathbb{R}_+^{D_z}$ are produced by autoregressive models, where D_z is the dimension of $\mathbf{z}_t$.

The global encoder and the local encoder each parameterise an approximate pos-

terior as a diagonal Gaussian distribution:

$$q_{\phi_s}(\mathbf{s}|\mathbf{x}_{1:L}) = \mathcal{N}\big(\mu_{\phi_s}(\mathbf{x}_{1:L}), \sigma^2_{\phi_s}(\mathbf{x}_{1:L})\mathbf{I}_{D_s}\big); \tag{2.5}$$

$$q_{\phi_z}(\mathbf{z}_t|\mathbf{x}_{1:L}) = \mathcal{N}\big(\mu_{\phi_z}(\mathbf{x}_{1:L}), \sigma^2_{\phi_z}(\mathbf{x}_{1:L})\mathbf{I}_{D_z}\big), \tag{2.6}$$

where the parameters of the distributions are output by neural networks with parameters ϕ_s and ϕ_z, respectively.

Finally, the decoder that parameterises the likelihood $p_{\theta_x}(\mathbf{x}_t|\mathbf{z}_t, \mathbf{s})$ is defined as a diagonal Gaussian with a fixed variance $\sigma_x^2 \in \mathbb{R}_+^{D_x}$ and a mean $\mu_{\theta_x}(\mathbf{z}_t, \mathbf{s}) \in \mathbb{R}^{D_x}$, where D_x denotes the dimension of the observed data.

With the parameterisation specified above, (2.4) becomes the following equation up to a constant:

$$\begin{aligned}
\mathcal{L}_{\text{DSAE}} = &\sum_{t=1}^{L} \mathbb{E}_{q_\phi(\mathbf{z}_t|\mathbf{x}_{1:L})q_\phi(\mathbf{s}|\mathbf{x}_{1:L})}\big[-\frac{1}{2\sigma_x^2}\sum_{d=1}^{D_x}(x_{t,d} - \mu_{\theta_x,d}(\mathbf{z}_t, \mathbf{s}))^2\big] \\
&+\frac{\beta_z}{2}\sum_{t=1}^{L}\mathbb{E}_{q_\phi(\mathbf{z}_{<t}|\mathbf{x}_{1:L})}\big[\sum_{d=1}^{D_z}(\log\frac{\sigma^2_{\phi_z,d}(\mathbf{x}_{1:L})}{\sigma^2_{\theta_z,d}(\mathbf{z}_{<t})} - \frac{\sigma^2_{\phi_z,d}(\mathbf{x}_{1:L})}{\sigma^2_{\theta_z,d}(\mathbf{z}_{<t})} \\
&\qquad - \frac{(\mu_{\phi_z,d}(\mathbf{x}_{1:L}) - \mu_{\theta_z,d}(\mathbf{z}_{<t}))^2}{\sigma^2_{\theta_z,d}(\mathbf{z}_{<t})}\big] \\
&+\frac{\beta_s}{2}\sum_{d=1}^{D_s}(\log\sigma^2_{\phi_s,d}(\mathbf{x}_{1:L}) - \sigma^2_{\phi_s,d}(\mathbf{x}_{1:L}) - \mu^2_{\phi_s,d}(\mathbf{x}_{1:L})),
\end{aligned} \tag{2.7}$$

where the subscript d denotes the d-th entry of a vector. Evaluating the two KLD terms benefits from their closed-form solutions, owing to the Gaussian parameterisation.

Since the expectations with respect to the latent variables in both the first and second terms are analytically intractable, they are approximated using a Monte Carlo estimate with a single sample drawn from the posterior distributions parameterised by the encoders. The reparameterisation trick (Kingma et al., 2014; Rezende et al., 2014) is employed to enable backpropagation through the sampling process.

2.2.2 Posterior Collapse of the Global Latent Variable

The DSAE framework achieves unsupervised disentanglement of time-invariant (global) and time-varying (local) factors in sequential data through a simple yet effective design. By applying temporal pooling to the global latent variable, the model distinguishes between the temporal resolutions of the latent variables, thereby facilitating the separation of global and local factors.

However, temporal pooling imposes a restrictive constraint on the global latent variable $\mathbf{s}$, resulting in lower information capacity compared to the local latent variables $\mathbf{z}_{1:L}$, which retain higher temporal resolution. This imbalance discourages the model from utilising the global latent during optimisation: the encoders tend to condense most information into the local latent space, and the decoder becomes insensitive to the global latent. This degeneracy is referred to as posterior collapse of the global latent.

In a standard VAE configured with a single latent variable $\mathbf{z}$, posterior collapse refers to the phenomenon where the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ converges to the prior $p(\mathbf{z})$ during optimisation (Y. Wang et al., 2021). This occurs due to what is often described as a preference of information: when the decoder $p_\theta(\mathbf{x}|\mathbf{z})$ is sufficiently expressive to model the data distribution independently of the latent variable $\mathbf{z}$, the optimisation favours minimising the KLD term $\mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})\big)$ by encouraging the approximate posterior to match the prior (X. Chen et al., 2017).

A common case arises when using a recurrent neural network decoder, which defines an autoregressive decoding distribution, i.e., $p_\theta(\mathbf{x}|\mathbf{z}) = \prod_l p_\theta(\mathbf{x}_l|\mathbf{z}, \mathbf{x}_{<l})$, capable of modelling the data distribution by conditioning only on past observations $\mathbf{x}_{<l}$, even without access to $\mathbf{z}$ (Bowman et al., 2016; Serban et al., 2017; Fraccaro et al., 2016). In such scenarios, although the model may generate high-quality samples $\tilde{\mathbf{x}} \sim p_\theta(\mathbf{x}|\tilde{\mathbf{z}})$, these samples are often unfaithful to the input data $\mathbf{x}'$ that was used to infer the latent variable $\tilde{\mathbf{z}} \sim q_\phi(\mathbf{z}|\mathbf{x}')$ (Alemi et al., 2018). As a result, the latent space becomes uninformative, defeating the purpose of representation learning.

Unlike standard VAEs, DSAEs are prone to posterior collapse of the global latent variable $\mathbf{s}$ even without a powerful decoder. As previously mentioned, temporal pooling imposes a restrictive constraint on $\mathbf{s}$, limiting its information capacity compared

to the local latents $\mathbf{z}_{1:L}$, which preserve higher temporal resolution. This imbalance discourages the model from utilising $\mathbf{s}$: the encoder tends to compress most information into the local latent space, and the decoder becomes increasingly insensitive to the global latent during optimisation.

In this case, the collapse of the global latent implies that

$$q_\phi(\mathbf{s}|\mathbf{x}_{1:L}) \approx p_\theta(\mathbf{s}) \quad \text{and} \quad p_\theta(\mathbf{x}|\mathbf{s}, \mathbf{z}_{1:L}) \approx p_\theta(\mathbf{x}|\mathbf{z}_{1:L}), \tag{2.8}$$

indicating that the posterior over $\mathbf{s}$ carries little information about the input sequence and is effectively ignored by the decoder. Posterior collapse hinders the application of DSAEs: replacing global latent variables with those extracted from independent sequences has little effect on the decoder's output. For tasks such as voice conversion and pitch-timbre disentanglement, the vanilla DSAE (Y. Li et al., 2018) is particularly susceptible to this issue in the absence of carefully tuned hyperparameters, as shown in Chapters 3 and 5.

Y. Li et al. (2018) typically address this issue by specifying a global latent space with a larger dimensionality than the local latent space, in order to compensate for the global latent's lower information capacity along the temporal axis. Chapter 3 proposes a training strategy that improves robustness with respect to the dimensionality of the latent spaces.

Inspired by the β-VAE (Higgins et al., 2017a), it is also common practice in the DSAE literature to balance the information bottleneck between the global and local latents by tuning the ratio between β_s and β_z in (2.4) (Han et al., 2021; B. Vowels et al., 2021; Bai et al., 2021; Naiman et al., 2023; Berman et al., 2024; Simon et al., 2024). For example, Lian et al. (2022b) and Lu et al. (2022) investigate the effect of the weight ratio on the successful application of DSAEs to voice conversion. Chapter 5 introduces a model-agnostic and parameter-free approach to mitigate the model's sensitivity to these weights.

The remainder of this section reviews countermeasures proposed in the DSAE literature to address this issue.

2.2.3 Regularising the Aggregated Posterior

The DSAE model, as specified in (2.1), generates a novel sequence by first sampling a global latent variable via $\tilde{\mathbf{s}} \sim p_\theta(\mathbf{s})$, followed by iteratively sampling local latents via $\tilde{\mathbf{z}}_t \sim p_\theta(\mathbf{z}_t|\tilde{\mathbf{z}}_{<t})$. Each time frame is then sampled as $\tilde{\mathbf{x}}_t \sim p_\theta(\mathbf{x}_t|\tilde{\mathbf{z}}_t, \tilde{\mathbf{s}})$ through the decoder. The global latent $\tilde{\mathbf{s}}$ is sampled only once and is shared across all time steps.

For the decoder to successfully generate meaningful outputs from arbitrary samples of the global latent $\tilde{\mathbf{s}}$, training must satisfy two key conditions:

1. The approximate posterior $q_\phi(\mathbf{s}|\mathbf{x}_{1:L})$ must not collapse to the prior $p_\theta(\mathbf{s})$; otherwise, no useful information is encoded in the global latent space, and nothing meaningful is communicated through $\tilde{\mathbf{s}}$.

2. The marginal posterior $q_\phi(\mathbf{s})$, defined below, must closely match the prior $p_\theta(\mathbf{s})$, ensuring that the decoder receives samples that effectively cover the support of the prior distribution, avoiding "holes." In other words, the decoder should be trained on samples that resemble the prior as closely as possible.

The marginal posterior, also referred to as the aggregated posterior (Makhzani et al., 2016; Tolstikhin et al., 2018), is defined as follows:

$$q_\phi(\mathbf{s}) = \mathbb{E}_{p_D(\mathbf{x}_{1:L})}\big[q_\phi(\mathbf{s}|\mathbf{x}_{1:L})\big] = \frac{1}{N}\sum_{i=1}^{N} q_\phi(\mathbf{s}|\mathbf{x}^i_{1:L}), \tag{2.9}$$

where $p_D(\mathbf{x}_{1:L}) = \frac{1}{N}$ is the empirical data distribution assigning a uniform probability mass over N sequences from the training data, and each sequence is indexed by i.

Rewriting the KLD

The KLD corresponding to the global latent from (2.4) can be written as follows according to Hoffman et al. (2016):

$$\mathbb{E}_{p_D(\mathbf{x}_{1:L})}\big[\mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{s}|\mathbf{x}_{1:L})\|p_\theta(\mathbf{s})\big)\big] = \mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{s})\|p_\theta(\mathbf{s})\big) + I_q(\mathbf{s};\mathbf{x}_{1:L}), \tag{2.10}$$

where the mutual information (MI) between $\mathbf{s}$ and $\mathbf{x}_{1:L}$ with respect to $q_\phi(\mathbf{s}|\mathbf{x}_{1:L})$ is defined as:

$$\begin{aligned} I_q(\mathbf{s}; \mathbf{x}_{1:L}) &= \mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{s}, \mathbf{x}_{1:L}) \| q_\phi(\mathbf{s}) p_D(\mathbf{x}_{1:L})\big) \\ &= \mathbb{E}_{p_D(\mathbf{x}_{1:L})} \mathbb{E}_{q_\phi(\mathbf{s}|\mathbf{x}_{1:L})}\big[\log \frac{q_\phi(\mathbf{s}|\mathbf{x}_{1:L}) p_D(\mathbf{x}_{1:L})}{q_\phi(\mathbf{s}) p_D(\mathbf{x}_{1:L})}\big]. \end{aligned} \tag{2.11}$$

(2.10) suggests that by maximising the ELBO averaged over the dataset, $\mathbb{E}_{p_D(\mathbf{x}_{1:L})}\big[\mathcal{L}_{\mathrm{DSAE}}\big]$, where $\mathcal{L}_{\mathrm{DSAE}}$ is defined by (2.4), both the mismatch between the aggregated posterior and the prior—measured by $\mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{s}) \| p_\theta(\mathbf{s})\big)$—and the MI are minimised. The former helps reduce "holes" in the latent space by encouraging the aggregated posterior to cover the support of the prior, while the latter removes information from the global latent, thereby encouraging posterior collapse.

To prevent posterior collapse and satisfy the two key conditions necessary for the decoder to render arbitrary global latent samples meaningfully, several works (Han et al., 2021; Bai et al., 2021; Naiman et al., 2023) propose modifications to the original DSAE by explicitly recovering the MI terms and optimising:

$$\mathbb{E}_{p_D(\mathbf{x}_{1:L})}\big[\hat{\mathcal{L}}_{\mathrm{DSAE}}\big] = \mathbb{E}_{p_D(\mathbf{x}_{1:L})}\big[\mathcal{L}_{\mathrm{DSAE}}\big] + I_q(\mathbf{s}; \mathbf{x}_{1:L}) + I_q(\mathbf{z}_{1:L}; \mathbf{x}_{1:L}). \tag{2.12}$$

This effectively replaces the dataset-averaged KLD term corresponding to the global latent—i.e., the left-hand side of (2.10)—with the KLD involving the aggregated posterior, represented by the first term on the right-hand side of (2.10). An analogous replacement is applied to the KLD term associated with the local latent.

(2.10) highlights the benefit of regularising the aggregated posterior $q_\phi(\mathbf{s})$ rather than the conditional posterior $q_\phi(\mathbf{s}|\mathbf{x}_{1:L})$. Specifically, it avoids minimising the MI, thereby helping to prevent posterior collapse of the global latent variable. Intuitively, the original KLD term encourages each conditional posterior to match the prior, which pushes the posterior for every datapoint toward the same distribution. In contrast, regularising the aggregated posterior does not impose this restriction, allowing the model to retain more information in the global latent space.

However, evaluating the aggregated posterior is intractable due to the need to

marginalise over the entire dataset. This makes it challenging to directly minimise $\mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{s})\|p_\theta(\mathbf{s})\big)$ or to estimate $I_q(\mathbf{s};\mathbf{x}_{1:L})$ for computing (2.12). The same applies to the local latent variables.

Minimising Divergence between the Aggregated Posterior and Prior

Since it is non-trivial to evaluate the KLD involving the aggregated posterior, Han et al. (2021) adapt the Wasserstein autoencoder (WAE) (Tolstikhin et al., 2018) to the DSAE framework and directly approximate the minimisation of a divergence measure $\mathcal{D}\big(q_\phi(\mathbf{s})\|p_\theta(\mathbf{s})\big)$ using likelihood-free approaches that rely only on samples from both distributions. This is done using either the kernel-based maximum mean discrepancy (MMD) (Gretton et al., 2007) or Jensen-Shannon divergence (JSD), the latter of which is closely related to the objective function optimised by generative adversarial networks (GANs) (Goodfellow et al., 2014). However, each approach presents distinct challenges that can impact model performance and computational efficiency.

While MMD provides a non-adversarial framework for distribution matching, its application in generative modelling is not without difficulties. A primary concern is the sensitivity to the choice of kernel function. The performance of MMD-based models can vary significantly depending on the kernel selected, necessitating careful tuning to achieve optimal results (Gretton et al., 2012; F. Liu et al., 2020). Additionally, MMD-based methods often require large batch sizes to accurately estimate the discrepancy between distributions, leading to high computational demands and memory usage (Y. Li et al., 2015).

On the other hand, GAN-based approaches, which utilise adversarial training to match distributions, face challenges related to training stability and sensitivity to hyperparameters (Kodali et al., 2018; S. Liu et al., 2017; Mescheder et al., 2018). Moreover, GANs are highly sensitive to architectural choices, learning rates, and initialization strategies (Salimans et al., 2016; Roth et al., 2017), making them prone to instability and requiring meticulous tuning to ensure effective training.

Estimating Mutual Information Based on Contrastive Learning

Instead of directly minimising $\mathcal{D}\big(q_\phi(\mathbf{s})\|p_\theta(\mathbf{s})\big)$ (e.g., using KLD), Bai et al. (2021) propose C-DSVAE, a contrastively disentangled sequential variational autoencoder, which maximises the following objective:

$$\begin{aligned}\mathbb{E}_{p_D(\mathbf{x}_{1:L})}\big[\mathcal{L}_{\text{C-DSVAE}}\big] &= \mathbb{E}_{p_D(\mathbf{x}_{1:L})}\big[\hat{\mathcal{L}}_{\text{DSAE}}\big] - I_q(\mathbf{s};\mathbf{z}_{1:L}) \\ &= \mathbb{E}_{p_D(\mathbf{x}_{1:L})}\big[\mathcal{L}_{\text{DSAE}}\big] + I_q(\mathbf{s};\mathbf{x}_{1:L}) + I_q(\mathbf{z}_{1:L};\mathbf{x}_{1:L}) - I_q(\mathbf{s};\mathbf{z}_{1:L}).\end{aligned} \tag{2.13}$$

Recall that $\mathbb{E}_{p_D(\mathbf{x}_{1:L})}\big[\hat{\mathcal{L}}_{\text{DSAE}}\big]$, defined in (2.12), replaces the dataset-averaged KLD term corresponding to both latent variables with the KLD terms that regularise the aggregated posteriors, as formulated in (2.10). In other words, C-DSVAE (Bai et al., 2021) regularises the aggregated posteriors by explicitly estimating the MI terms, thereby alleviating the optimisation and hyperparameter tuning challenges posed by MMD- or GAN-based distribution matching techniques (Han et al., 2021).

Bai et al. (2021) further introduce a penalty on the MI between the global and local latent variables, $I_q(\mathbf{s};\mathbf{z}_{1:L})$, to encourage disentanglement. They also show that the objective in (2.13) constitutes a valid ELBO on the marginal log-likelihood $\log p_\theta(\mathbf{x}_{1:L})$.

Bai et al. (2021) approximate the MI terms by InfoNCE (Oord et al., 2018; T. Chen et al., 2020), based on noise-contrastive estimation (NCE) (Gutmann et al., 2010):

$$\mathcal{L}_{\text{InfoNCE}}(\mathbf{u}) = -\log \frac{\exp(\text{sim}(\mathbf{u},\mathbf{v}^+)/\tau)}{\exp(\text{sim}(\mathbf{u},\mathbf{v}^+)/\tau) + \sum_{j=1}^{M}\exp(\text{sim}(\mathbf{u},\mathbf{v}^j)/\tau)}, \tag{2.14}$$

where $\text{sim}(\cdot,\cdot)$ is a similarity measure, usually defined as cosine similarity, and τ denotes a temperature parameter. $\mathbf{v}^j$ and $\mathbf{v}^+$ are the negative and positive "views" of $\mathbf{u} \in \{\mathbf{s},\mathbf{z}_{1:L}\}$, respectively. For each sample of $\mathbf{u}$, a single positive view and M negative views are used to compute the loss.

Using an image as an example, a positive view is an alternative view that shares the same semantic content—such as object identity—with the original image. Positive

views are typically obtained through data augmentation techniques such as random cropping and colour distortion (T. Chen et al., 2020), which preserve object identity. In contrast, a negative view is defined as every other sample in a mini-batch during training, under the assumption that these samples likely depict different objects.

The contrastive predictive coding (CPC) framework (Oord et al., 2018) minimises the InfoNCE loss, $\mathcal{L}_{\text{InfoNCE}}$, to learn latent representations from sequential data, leveraging temporal smoothness as its primary inductive bias. Given a sequence, the positive view is defined as a sample from a few time steps into the future. Importantly, they show that:

$$I(\mathbf{u}; \mathbf{x}_{1:L}) \geq \log(M+1) - \mathcal{L}_{\text{InfoNCE}}(\mathbf{u}), \tag{2.15}$$

indicating that minimising $\mathcal{L}_{\text{InfoNCE}}(\mathbf{u})$ maximises a lower bound on mutual information, and that a larger number of negative samples M tightens the bound.

In the context of the DSAE framework, Bai et al. (2021) propose a strategy for constructing positive and negative views with respect to the two latent variables. To construct positive views for the global latent $\mathbf{s}$, each sequence in a mini-batch is paired with a temporally shuffled version of itself, which is used to infer $\mathbf{v}^+$ in (2.14). Temporal shuffling preserves the static factors that are shared across time, making it a suitable positive counterpart. All other sequences in the mini-batch are treated as negative views and used to infer $\mathbf{v}^j$. Discriminating positive samples from negatives encourages the model to form a discriminative global latent space and mitigates the risk of posterior collapse.

For local latents $\mathbf{z}_{1:L}$, positive views are constructed to preserve dynamic factors while modifying static ones. For video data, Bai et al. (2021) use a combination of frame cropping and colour distortion, following standard image augmentation techniques (T. Chen et al., 2020). For speech data, they generate positive alternatives by applying classic voice conversion algorithms (Verhelst et al., 1993) to modify speaker characteristics while preserving linguistic content.

Bai et al. (2021) also benchmark the estimation of mutual information (MI) terms using a simpler approach based on mini-batch weighted sampling (MWS) (T. Q. Chen et al., 2018; Zhu et al., 2020), and report inferior performance compared to their

proposed contrastive method. They attribute this to the effectiveness of domain-specific data augmentations, which enforce invariance in certain attributes and serve as strong inductive biases for learning disentangled representations.

In an ablation study, Bai et al. (2021) optimise a simplified version of the objective in (2.12), using contrastive estimation but without explicitly minimising $I_q(\mathbf{s}; \mathbf{z}_{1:L})$. Interestingly, they observe that the MWS estimate of this MI term naturally decreases, suggesting that the contrastive signal implicitly encourages disentanglement between global and local latent variables.

However, C-DSVAE (Bai et al., 2021) relies on domain knowledge to apply data augmentations that produce meaningful positive views, as the choice of transformation can significantly impact performance (Tian et al., 2020; Zhang et al., 2022).

In addition, treating each randomly selected sample from a mini-batch as a negative view introduces the risk of false negatives, which can bias learning (Chuang et al., 2020). T. Chen et al. (2020) recommend increasing the batch size to mitigate this issue, consistent with the observation that larger values of M improve the MI lower bound, as described in (2.15). However, this strategy comes at the cost of increased memory consumption. Indeed, Hsu et al. (2017) explored a similar contrastive strategy for distinguishing global latents across different training sequences and found it non-trivial to scale the number of negative samples effectively (Hsu et al., 2018b).

Summary

In summary, Han et al. (2021) and Bai et al. (2021) both address posterior collapse in DSAEs by regularising the aggregated posterior rather than the conditional posterior, thereby avoiding the minimisation of MI that can suppress latent informativeness. Han et al. (2021) adopt likelihood-free distribution matching using either MMD or JSD, enabling flexible regularisation without the need for explicit likelihoods. In contrast, Bai et al. (2021) propose C-DSVAE, which estimates MI through contrastive learning by constructing domain-specific positive and negative views, and explicitly penalises the MI between global and local latents to encourage disentanglement. While the former is conceptually simple and broadly applicable, it requires careful tuning of kernel or discriminator settings. The latter achieves stronger disen-

tanglement but depends heavily on augmentation design and may suffer from false negatives and increased memory requirements due to the need for large batch sizes.

2.2.4 Leveraging Architectural Biases

The original DSAE (Y. Li et al., 2018) relies primarily on its architectural bias based on the assumption that time-invariant and time-varying factors are represented by a single vector and a sequence of vectors, respectively, to achieve unsupervised disentanglement. This section describes approaches that leverage this bias and proposes alternative training strategies (Naiman et al., 2023; Berman et al., 2024).

Modality-Free Contrastive Learning

Inspired by the contrastive learning-based methods in Section 2.2.3, Naiman et al. (2023) exploit the architectural bias and utilise the posterior distributions to sample both positive and negative views, without resorting to domain-specific data augmentation methods, unlike C-DSVAE (Bai et al., 2021).

The key intuition is that, after a warm-up phase of C training epochs, the encoder $q_{\phi^C}(\mathbf{u}|\mathbf{x}_{1:L})$, where $\mathbf{u} \in \{\mathbf{s}, \mathbf{z}_{1:L}\}$, will produce similar posterior distributions for sequences that share the same underlying latent factors. Given a mini-batch of N training sequences and their encoded posteriors, they compute pairwise distributional distances using:

$$\mathcal{D}^C_{i,j} = \mathcal{D}_{\mathrm{KL}}\big(q_{\phi^C}(\mathbf{u}|\mathbf{x}^i_{1:L}) \| q_{\phi^C}(\mathbf{u}|\mathbf{x}^j_{1:L})\big), \quad \text{for } i, j \in \{1, \ldots, N\}. \tag{2.16}$$

Contrastive pairs are then defined heuristically: The positive set of distributions for the i-th sequence, the anchor, is the first third of N distributions of the sorted i-th row of $D^C_{i,j}$ in ascending order, while the last third of the distributions constructs the negative set.

Although this approach is modality-agnostic and does not require handcrafted augmentations, it assumes that the posterior has not collapsed to the prior by the C-th epoch. If posterior collapse occurs, then $\mathcal{D}^C_{i,j} \to 0$ for all i, j, and no meaningful contrastive signal can be extracted. To mitigate this risk, Naiman et al. (2023) follow

common practice by setting the global latent dimensionality larger than the local one and carefully tuning β_s and β_z in (2.4) to balance their information capacities.

In summary, Naiman et al. (2023) exploit the DSAE's natural tendency to cluster similar latent factors in the posterior space for contrastive learning, enabling augmentation-free training. However, their method still requires extensive hyperparameter tuning and does not directly address posterior collapse.

Alternative Training Strategies

To address the issue of posterior collapse of the global latent in DSAE, as described in Section 2.2.2, common strategies are discussed, including:

1. Balancing the information capacity between the global and local latent variables, by adjusting their respective dimensionalities and the weights of the two KLD terms in the ELBO (2.4); and

2. Regularising the aggregated posterior rather than the per-datapoint posterior, as illustrated in (2.10), which augments the ELBO with mutual information terms (2.12).

While effective, these approaches often come with practical drawbacks: they introduce optimisation difficulties (Han et al., 2021), rely on domain-specific augmentations (Bai et al., 2021), or require extensive hyperparameter tuning (Han et al., 2021; Naiman et al., 2023). Moreover, the resulting hyperparameters are often dataset-specific, and it is unclear whether tuning is conducted based on the classification accuracy of downstream supervised tasks (Bai et al., 2021), or on validation losses derived from the model's objective functions. In the former case, the use of supervised signals during model selection undermines the claim of unsupervised training, as it introduces supervision through the evaluation protocol.

Inspired by the elegant design of the original model (Y. Li et al., 2018), this raises a natural research question: Is it possible to leverage the intrinsic architectural bias of the DSAE framework while remaining modality-independent and avoiding excessive complexity in the training objective?

Berman et al. (2024) propose sampling the global latent variable exclusively from

a single time frame: $\tilde{\mathbf{s}}_k \sim q_\phi(\mathbf{s}|\mathbf{x}_k)$, where $k \in \{1, \ldots, L\}$, and empirically show that setting $k = 1$ yields strong performance. This strategy builds on the assumption that time-invariant factors are shared across all time steps, and that a single frame is inherently incapable of capturing time-varying information. As a result, sampling from a single frame helps prevent dynamic information from leaking into the global latent space.

Conversely, to avoid static information from contaminating the dynamic latent space, Berman et al. (2024) subtract the intermediate representations of the static factors from those of the dynamic factors, thereby encouraging disentanglement between the two.

The model is referred to as the Single-Frame DSAE (SF-DSAE), and its objective is expressed as:

$$\begin{aligned}\mathcal{L}_{\text{SF-DSAE}} = & \sum_{t=2}^{L} \mathbb{E}_{q_{\phi_z}(\mathbf{z}_t|\mathbf{x}_{\leq t},\mathbf{z}_{<t})q_{\phi_s}(\mathbf{s}|\mathbf{x}_1)}\big[\log p_\theta(\mathbf{x}_t|\mathbf{z}_t,\mathbf{s})\big] \\ & + \alpha \mathbb{E}_{q_{\phi_z}(\mathbf{z}_1|\mathbf{x}_1)q_{\phi_s}(\mathbf{s}|\mathbf{x}_1)}\big[\log p_\theta(\mathbf{x}_1|\mathbf{z}_1,\mathbf{s})\big] \\ & - \beta_z \sum_{t=1}^{L} \mathbb{E}_{q_{\phi_z}(\mathbf{z}_t|\mathbf{x}_{\leq t},\mathbf{z}_{<t})}\big[\mathcal{D}_{\text{KL}}\big(q_{\phi_z}(\mathbf{z}_t|\mathbf{x}_{\leq t},\mathbf{z}_{<t})\|p_\theta(\mathbf{z}_t|\mathbf{z}_{<t})\big)\big] \\ & - \beta_s \mathcal{D}_{\text{KL}}\big(q_{\phi_s}(\mathbf{s}|\mathbf{x}_1)\|p_\theta(\mathbf{s})\big).\end{aligned} \tag{2.17}$$

This objective extends $\mathcal{L}_{\text{DSAE}}$ in (2.4) by introducing an additional reconstruction term for the first frame, weighted by $\alpha \in \mathbb{R}^+$. To reduce the need for hyperparameter tuning, Berman et al. (2024) set a common value for both β_z and β_s, and use equal dimensionality for the global and local latent spaces. In an ablation study, they show that both design choices—sampling the global latent from a single frame and applying latent-space subtraction—are critical for achieving meaningful disentanglement.

Nevertheless, the model exhibits limitations, particularly when applied to less structured datasets. First, the time frame at step k used to infer the global latent does not always contain sufficient static information. For example, audio recordings of musical instruments often include interleaved silent and active frames; sampling a silent frame can result in a degenerate or collapsed global representation. Second, the subtraction operation assumes additivity between the static and dynamic latent

variables. However, there is limited theoretical justification for this assumption, and its validity remains supported only by empirical results on specific datasets reported by Berman et al. (2024).

2.2.5 Summary of Limitations

This section reviews key works in the DSAE literature for unsupervised disentanglement of time-invariant and time-varying factors in sequential data. To summarise, Y. Li et al. (2018) introduce the DSAE framework to separate global and local factors, but it is prone to posterior collapse. Han et al. (2021) address this by regularising the aggregated posterior using likelihood-free metrics, though their approach requires careful tuning. Bai et al. (2021) achieve strong disentanglement via contrastive MI estimation with domain-specific augmentations, at the cost of increased memory usage and sensitivity to augmentation design. Naiman et al. (2023) leverage latent clustering for augmentation-free contrastive learning, but still rely on extensive hyperparameter tuning and do not explicitly prevent collapse. Berman et al. (2024) enforce disentanglement through single-frame sampling and latent subtraction, but assume structured inputs and underperform on sparse or silent data.

This thesis aims to alleviate posterior collapse and improve the robustness of DSAEs to hyperparameters and architectural choices. To this end, alternative training strategies are proposed that avoid introducing additional trainable parameters, domain-specific components, or complex estimators into the training objective.

2.3 Pitch and Timbre Disentanglement

2.3.1 Overview

Pitch and timbre are two significant descriptors of musical instrument sounds. While pitch is the perceived frequency of a sound, timbre is the quality or colour that distinguishes different sounds sharing the same pitch and loudness (Acoustical Society of America, 1994; McAdams et al., 2017).

To establish a well-defined evaluation protocol for determining whether a trained neural network has successfully learnt disentangled representations, the literature

on pitch-timbre disentanglement (PTD) often defines timbre by instrument identity (Engel et al., 2017; Bitton et al., 2018; Luo et al., 2019; Y. Wu et al., 2023; Esling et al., 2018; Luo et al., 2020; Tanaka et al., 2021; Cífka et al., 2021; Tanaka et al., 2022; Tanaka et al., 2025; Demerlé et al., 2024). The approach circumvents the elusive and multifaceted nature of timbre by associating it with discrete, identifiable categories and conveniently evaluates PTD models with pitch and instrument classifiers.

This overview first outlines the general framework underlying key works in the PTD literature and then describes a common evaluation protocol.

An Autoencoding Framework

A main application of PTD is to generate novel instrument sounds $\hat{\mathbf{x}}$ by conditioning a decoder $\mathrm{Dec}(\cdot;\theta)$, parameterised by θ, on two separate feature representations: $\mathbf{z}_P$, encoding pitch information such as note height or melody contour, and $\mathbf{z}_I$, capturing timbral characteristics of a musical instrument. This can be expressed as:

$$\hat{\mathbf{x}} = \mathrm{Dec}(\mathbf{z}_P, \mathbf{z}_I; \theta). \tag{2.18}$$

Ideally, manipulating or replacing either input should not interfere with the rendering of the other attribute.

Approaches to PTD can be characterised primarily by how the feature representations $\{\mathbf{z}_P, \mathbf{z}_I\}$ are obtained. Methods that rely on ground-truth annotations of pitch and instrument often derive the features as functions of their respective labels $\{y_P, y_I\}$, either through a trainable lookup table or intermediate feature maps of networks (Engel et al., 2017; Bitton et al., 2018; Luo et al., 2019; Y. Wu et al., 2023), enabling explicit control over the annotated attributes. Alternatively, the features are typically extracted from audio inputs $\mathbf{x}$, which may facilitate generalisation beyond the pitch and instrument categories seen during training (Tanaka et al., 2021; Cífka et al., 2021; Tanaka et al., 2022).

The features are obtained via $\mathrm{Enc}(\cdot;\phi_U)$, generally referred to as a function with trainable weights ϕ_U:

$$\mathbf{z}_U = \mathrm{Enc}(\mathbf{a}; \phi_U), \tag{2.19}$$

where $U \in \{P, I\}$, and $\mathbf{a}$ denotes either audio inputs $\mathbf{x}$ or attribute annotations y_U.

An autoencoding PTD framework typically minimises the following objective function:

$$\mathcal{L}_{\text{PTD}} = \mathcal{D}(\mathbf{x}, \hat{\mathbf{x}}) + \gamma \mathcal{R}(\mathbf{z}_U), \tag{2.20}$$

which minimises a reconstruction loss measured by a distance function $\mathcal{D}(\cdot, \cdot)$, in addition to regularisation terms $\mathcal{R}(\cdot)$ weighted by γ.

The regularisation terms are generally necessary to promote the disentanglement of $\mathbf{z}_P$ and $\mathbf{z}_I$, including cycle consistency (Bitton et al., 2018; Y. Wu et al., 2023), maximum likelihood (Luo et al., 2019), metric learning (Esling et al., 2018; Luo et al., 2020; Tanaka et al., 2021; Tanaka et al., 2022), adversarial loss (Demerlé et al., 2024), and within- and between-sample variance (Y. Wu et al., 2025).

Beyond auxiliary regularisers, architectural priors and information bottlenecks — such as discretised representations (Luo et al., 2020; Cífka et al., 2021) and temporal hierarchies (Cífka et al., 2021; Tanaka et al., 2022; Demerlé et al., 2024) — are critical for facilitating disentanglement in the absence of attribute annotations.

Evaluation

Defining timbre by instrument identity facilitates a well-defined evaluation protocol. A typical approach examines whether the decoder can faithfully render different pitches or instruments given manipulated input representations.

Within this pipeline, the encoder must produce disentangled representations to which the decoder is sensitive. To implement this pipeline, pitch and instrument classifiers are trained using a dataset of size N, denoted as $\{(y_U^i, \mathbf{x}^i)\}_{i \in [N]}$, where each datapoint comprises a ground-truth annotation y_U^i and audio $\mathbf{x}^i$. Synthetic sounds are then generated from a pre-trained PTD model:

$$\begin{aligned} &\mathbf{x}^{m,n} = \text{Dec}(\mathbf{z}_P^m, \mathbf{z}_I^n; \theta), \quad \text{where} \\ &\mathbf{z}_P^m = \text{Enc}(\mathbf{a}^m; \phi_P) \quad \text{and} \quad \mathbf{z}_I^n = \text{Enc}(\mathbf{a}^n; \phi_I). \end{aligned} \tag{2.21}$$

Here, $\mathbf{x}^{m,n}$ is generated by combining the pitch and timbre representations of $\mathbf{x}^m$ and $\mathbf{x}^n$, respectively.

A successful PTD model should extract and render the pitch and timbre representations such that the pitch and instrument classifiers can accurately recognise pitch y_P^m and instrument y_I^n given $\mathbf{x}^{m,n}$. As the classifiers are trained using real data, this evaluation pipeline also requires the PTD model to generate realistic synthetic sounds; otherwise, the distributional gap between the real and the synthesised data would reduce the classification accuracy, even if the two representations are well disentangled.

For PTD approaches that explicitly extract latent representations as a function of input audio, i.e., $\mathbf{a} \coloneqq \mathbf{x}$, another evaluation method focusing on the learnt representations is applicable. It involves training supervised pitch and instrument classifiers on a dataset $\{(y_U^i, \mathbf{z}_U^i)\}_{i\in[N]}$, where each datapoint comprises a ground-truth annotation y_U^i and the corresponding learnt representation $\mathbf{z}_U^i$ extracted using the pre-trained encoders.

The performance of disentanglement is then assessed based on the classification accuracy of pitch and instrument. Successfully disentangled representations of pitch (respectively, timbre) are expected to yield high classification accuracy for pitch (respectively, instrument) and low accuracy for instrument (respectively, pitch).

In the following sections, we categorise the principal contributions to the PTD literature according to the primary inductive biases they adopt: label conditioning, metric learning, and temporal resolution modelling.

2.3.2 Label Conditioning Methods

The majority of works relying on ground-truth annotations of pitch or instrument do not explicitly learn separate representations of pitch and timbre — the labels act as surrogate conditioning information on the decoder for sound synthesis (Engel et al., 2017; Bitton et al., 2018; Y. Wu et al., 2023) and $\mathbf{a} \coloneqq \mathbf{y}_U$ in (2.19).

Engel et al. (2017) present an early exploration into PTD using a WaveNet-style autoencoder framework (Oord et al., 2016). Their model comprises an encoder that processes raw audio waveforms into temporal embeddings and a WaveNet decoder that reconstructs audio conditioned on these embeddings. While the primary model learns embeddings capturing timbral characteristics, it does not explicitly disentangle

pitch and timbre representations.

A preliminary attempt at PTD involves conditioning the decoder with pitch annotations of the input isolated notes. However, the effect is limited, as the decoder is not sensitive to the pitch condition, and the learnt embeddings remain largely entangled with pitch information. They report that pitch information can be separated from the learnt embeddings to some extent by providing the decoder with the ground-truth pitch condition when the dimensionality of the embeddings is drastically reduced. Nevertheless, this imposes a trade-off between disentanglement and audio reconstruction quality.

Since the main focus of Engel et al. (2017) is raw audio generation, they employ a plain autoencoder framework without auxiliary regularisation to enforce disentanglement.

Bitton et al. (2018) propose the modulated variational autoencoder (MoVE) for many-to-many musical timbre transfer. MoVE adopts a VAE framework, comprising an encoder that maps input audio to an embedding, and a decoder that reconstructs audio conditioned on the embedding alongside external labels.

MoVE conditions the decoder on both pitch (octave and pitch class) and domain (instrument) labels. The pitch labels facilitate explicit control over synthesised sounds. The domain condition serves as a switch between different instruments, enabling many-to-many timbre transfer without necessitating separate decoders for each instrument, as required in prior works such as Mor et al. (2019). With information of both the pitch and instrument captured through the conditioning labels, the learnt embedding encodes the data variance explained by descriptors used to compare the qualities of different sounds (Peeters et al., 2011). MoVE does not explicitly learn separate latent spaces for pitch and timbre; instead, it relies on the combination of the embedding and the conditioning labels to achieve controllable synthesis.

Y. Wu et al. (2023) introduce TransPlayer. Like MoVE, TransPlayer adopts an autoencoder framework; however, it diverges in its approach to disentangling pitch and timbre. While MoVE conditions the decoder on both pitch and instrument labels, relying on the combination of these labels and a learnt bottleneck for synthesis, TransPlayer focuses on isolating pitch information within the latent representation.

Similar to MoVE, the conditioning instrument labels allow for many-to-many timbre transfer without necessitating separate decoders for each. Inspired by AutoVC (Qian et al., 2019), the model employs a narrow bottleneck to prevent instrument information from leaking into the latent representation, serving as a key inductive bias for disentanglement.

To further enforce the separation of pitch and timbre, TransPlayer incorporates a consistency loss as an auxiliary regularisation technique. This loss ensures that decoding with either the original or a different instrument condition, followed by re-encoding, yields the same pitch representation in the bottleneck. This strategy promotes the stability of the pitch representation across varying timbral conditions, enhancing the model's ability to disentangle and manipulate pitch and timbre independently.

Unlike the aforementioned works, the model proposed by Luo et al. (2019) explicitly learns disentangled representations for both pitch and timbre. They employ a Gaussian mixture VAE (Jiang et al., 2017; Dilokthanakul et al., 2016), utilising two separate encoders to learn distinct latent spaces for timbre and pitch, each modelled as a Gaussian mixture corresponding to instrument identity and pitch classes, respectively. During reconstruction, latent variables are sampled from these mixtures and concatenated as input to the decoder.

The auxiliary regularisation is constructed using supervision from instrument and pitch labels to encourage disentanglement — $\mathbf{z}_U = \text{Enc}(\mathbf{x}; \phi_U)$ is regularised to maximise the likelihood under a Gaussian component, $p_{\phi_U}(\mathbf{z}_U | y_U)$, parameterised by a mean and a variance through a trainable lookup table indexed by the ground-truth y_U.

Compared to the other label conditioning methods, Luo et al. (2019) more closely follow the general PTD framework outlined by (2.20), reconstructing audio from separate latent representations extracted by their respective encoders as functions of the input audio. The following discussion of the PTD literature similarly builds upon this framework, organised according to the different inductive biases employed for disentanglement.

2.3.3 Metric-Learning-Based Methods

Metric-learning-based approaches regularise latent spaces by enforcing similarity measures or contrastive strategies grounded in domain knowledge. These methods encourage PTD by shaping the latent representations such that pitch (resp. timbre) latents derived from data sharing the same pitch (resp. timbre) are encouraged to be close in the latent space.

Esling et al. (2018) propose a method that integrates perceptual ratings from timbre studies into the training of a VAE. By applying multidimensional scaling to subjective listening test data, they construct a perceptual timbre space (McAdams et al., 2005). This space serves as a target for regularising the VAE's latent space, ensuring that the model's internal representations align with human auditory perception of timbral similarity. The primary focus of this work is to bridge the gap between perceptual studies and generative modelling, rather than explicitly disentangling pitch and timbre.

Luo et al. (2020) introduce an unsupervised framework for PTD that employs a VAE architecture with distinct latent variables: a categorical variable for pitch and a continuous variable for timbre. The model leverages pitch-shifting data augmentation to create pairs of inputs that share timbre but differ in pitch. Contrastive loss functions (T. Chen et al., 2020) are applied to these pairs to encourage the model to learn pitch-invariant timbre representations. Additionally, a cycle-consistency loss ensures that decoding and re-encoding processes preserve the timbre representation. This methodology enables the disentanglement of pitch and timbre without relying on explicit pitch or instrument labels.

Tanaka et al. (2021) present a VAE-based approach that avoids categorical constructions for pitch and timbre to enhance generalisation to unseen attributes. Instead, they employ weak supervision by grouping inputs that share the same attributes and measure Euclidean distances to regularise their latent representations of pitch and timbre. These losses are designed to pull together latent representations of samples with shared attributes (either pitch or timbre) and push apart those without. This strategy allows the model to learn disentangled representations based on relational information, rather than explicit labels, facilitating the handling of novel pitches and

instruments.

Collectively, these metric learning-based methods more explicitly learn disentangled pitch and timbre representations by introducing auxiliary regularisation based on similarity measures that rely on domain knowledge, alleviating the requirement of ground-truth annotations.

2.3.4 Temporal Resolution Modelling

Temporal resolution modelling approaches design models with hierarchical latent structures to disentangle factors operating at different temporal scales. These methods explicitly encode both pitch and timbre representations as functions of input audio, rather than relying on ground-truth annotations. A core inductive bias employed is that timbre representations remain relatively stable (low variance) within a sample, while pitch representations are dynamic (high variance). Most methods impose a hard constraint based on this bias: the timbre representation is a vector associated with the entire input sequence, whereas the pitch representation preserves the time resolution of the input sequence.

Methods with Data Augmentation

Cífka et al. (2021) propose a self-supervised VQ-VAE (Oord et al., 2017) for one-shot music style transfer. The model imposes a hard architectural constraint where the timbre latent is represented as a single vector, and the pitch latent is a sequence of vectors. To limit the information capacity and prevent timbre leakage into the pitch latent, the pitch representation is further discretised. Data augmentation techniques, such as pitch shifting and timbre transformation, are applied. Instead of deriving pitch and timbre latents from the same data, they are extracted from augmented versions of the reconstruction target, encouraging disentanglement through pitch- and timbre-invariant transformations.

Tanaka et al. (2022) introduce methods for unsupervised disentanglement of timbre, pitch, and variation features from musical instrument sounds. To accommodate timbre variation over time, they introduce an additional sequence of temporal timbre representations alongside the global timbre and local pitch representations. The

models rely on domain-dependent data augmentation and impose timbre- and pitch-invariant objective functions. In Tanaka et al. (2025), a re-entry training framework is presented, applying the network for three-factor disentanglement twice in series with shared weights, refining the characteristics extracted by the encoders and effectively performing implicit data augmentation.

Methods without Data Augmentation

Demerlé et al. (2024) propose a method that combines audio control and style transfer using latent diffusion (Ho et al., 2020; Song et al., 2021; Rombach et al., 2022). The model imposes a hard architectural constraint with a global timbre representation and a sequential pitch representation. A two-stage training strategy is employed to promote an informative global timbre representation. Additionally, an adversarial loss serves as the main auxiliary regulariser to further enhance disentanglement.

Y. Wu et al. (2025) introduce an unsupervised method that learns disentangled content and style representations from sequences of observations by leveraging domain-general statistical differences between content and style, without relying on domain-specific labels. Unlike the aforementioned methods, the method does not impose a hard constraint on the temporal resolution of the latent representations.

Collectively, these temporal resolution modelling methods contribute to the PTD literature by introducing architectural constraints and training strategies that leverage temporal characteristics of pitch and timbre.

2.3.5 Summary of Limitations

This section has reviewed key methods for PTD, grouped by their primary inductive biases: label conditioning, metric learning, and temporal resolution modelling. All methods follow a common autoencoding framework that synthesises musical audio from pitch and timbre representations, with disentanglement typically encouraged through architectural constraints or auxiliary regularisation.

Label conditioning methods rely on ground-truth pitch and instrument labels to control synthesis, often without extracting representations from audio (Engel et al., 2017; Bitton et al., 2018; Y. Wu et al., 2023). Consequently, these methods lack the

ability to infer attributes from reference audio and are unsuitable for downstream tasks such as analysis or attribute transfer. Although these models offer explicit control via labels, they are constrained by their reliance on annotated data and limited generalisability.

Metric learning-based approaches shape the latent space using domain knowledge through perceptual similarity (Esling et al., 2018), contrastive learning (Luo et al., 2020), or weak supervision (Tanaka et al., 2021). These methods reduce dependency on labels and support generalisation to unseen categories. However, they often require carefully designed training protocols and handcrafted objectives that may be difficult to scale.

Temporal resolution modelling methods leverage the inherent timescale differences between pitch and timbre to impose architectural priors on latent representations. These models enable disentanglement without labels, but many rely on domain-specific data augmentation (Cífka et al., 2021; Tanaka et al., 2022; Tanaka et al., 2025), inheriting the same scalability challenges as metric learning methods. In addition, the use of adversarial losses (Demerlé et al., 2024) may complicate training, while the strict requirement for structured data fragments (Y. Wu et al., 2025) limits applicability to real-world music.

Chapter 3 builds upon the framework of disentangled sequential autoencoder (DSAE) reviewed in Section 2.2, leveraging the temporal resolution modelling assumption to achieve robust, unsupervised PTD without reliance on domain-dependent augmentation. Furthermore, Chapter 6 proposes a supervised framework tailored to mixtures of musical sources. By binding each source to a unique disentangled representation, and conditioning a decoder on a set of such representations, the framework enables the composition of arbitrary pitch-timbre combinations.

2.4 Zero-Shot Voice Conversion

2.4.1 Overview

Voice conversion (VC) is the task of transforming a source speaker's voice to resemble that of a target speaker while preserving the underlying linguistic content. Zero-shot

VC extends this task to scenarios in which the target speaker is not encountered during training. A key approach underpinning recent progress in zero-shot VC is the learning of disentangled representations that separate speaker identity from linguistic content in speech utterances.

The disentangled sequential autoencoder (DSAE) (Y. Li et al., 2018), introduced in Section 2.2, assumes that time-series data can be generated by a time-invariant (global) latent variable and time-variant (local) latent variables that preserve temporal information (Han et al., 2021; Bai et al., 2021; Naiman et al., 2023; Berman et al., 2024). This assumption is widely shared by contemporary approaches to zero-shot VC (Hsu et al., 2017; Qian et al., 2019; Nguyen et al., 2022; Chou et al., 2019; D.-Y. Wu et al., 2020; Tang et al., 2022; D. Wang et al., 2021; Tjandra et al., 2021; Lian et al., 2022b; Lu et al., 2022; Lian et al., 2022a; Wei et al., 2023; Lu et al., 2024). These methods facilitate disentanglement by assuming that speaker identity and linguistic content can be represented by global and local latent variables, respectively.

Similar to the task of pitch–timbre disentanglement (PTD) discussed in Section 2.3, the goal is to learn an autoencoder composed of an encoder and a decoder:

$$\hat{\mathbf{x}} = \mathrm{Dec}(\mathbf{z}_S, \mathbf{z}_C; \theta); \quad \mathbf{z}_U = \mathrm{Enc}(\mathbf{x}; \phi_U), \quad \text{where} \quad U \in \{S, C\}. \tag{2.22}$$

According to the key assumption that distinguishes the temporal resolutions of the two latent variables, $\mathbf{z}_S \in \mathbb{R}^{L_S}$ and $\mathbf{z}_C \in \mathbb{R}^{L_C \times T_C}$ are defined to denote the speaker and content representations, respectively, and $\mathbf{x} \in \mathbb{R}^{L_X \times T_X}$. In practice, $T_X \geq T_C > 1$. Voice conversion is then achieved by replacing the speaker latent $\mathbf{z}_S$ while preserving the content latents $\mathbf{z}_C$ during decoding.

As discussed in Section 2.2.2, the DSAE framework faces an optimisation challenge arising from the mismatch in information capacity between the global and local variables: the model often collapses the global latent by routing all useful information through the local latent variables. In zero-shot VC, this challenge manifests in the content latents $\mathbf{z}_C$ capturing most of the information and rendering the speaker latent $\mathbf{z}_S$ uninformative (Lian et al., 2022b; Lu et al., 2022). In what follows, existing approaches to zero-shot VC are outlined according to how they maintain an informative

speaker representation and prevent speaker information from leaking into the content representation.

2.4.2 Removing the Bottleneck of Speaker Representations

In DSAEs, (2.10) suggests that regularising the global encoder with a prior minimises the mutual information (MI) between the input sequence and the global latent, promoting posterior collapse of the global latent. To avoid the aggressive regularisation and learn informative speaker representations, some works simplify their models by discarding the prior distribution (Qian et al., 2019; Chou et al., 2019; D.-Y. Wu et al., 2020; Tang et al., 2022; D. Wang et al., 2021).

However, the lack of a prior distribution prohibits their models from unconditionally sampling novel speaker characteristics and smoothly interpolating between speakers (Hsu et al., 2019; Stanton et al., 2021).

2.4.3 Constraining Content Representations

Architectural Constraints

To enhance the disentanglement between speaker and content information, Qian et al. (2019) balance the information bottleneck between the global (speaker) and the local (content) latents by carefully tuning their dimensionality. Furthermore, they leverage speaker embeddings pre-trained with supervised discriminative tasks (Heigold et al., 2016; Wan et al., 2018) to sidestep learning the speaker encoder from scratch, stabilising the training process.

Chou et al. (2019) apply an adaptive instance normalisation layer (X. Huang et al., 2017) to the content representation. Their evaluation demonstrates that the additional layer helps remove static information such as speaker characteristics from the content latent and yields improved disentanglement without the aid of pre-trained speaker embeddings.

Quantisation

A common drawback of heuristic architectural constraints is the requirement for extensive hyperparameter tuning. D.-Y. Wu et al. (2020) show further improvement on disentanglement by quantising the content latent space, exploiting the discrete nature of the acoustic units of speech (Oord et al., 2017; Chorowski et al., 2019; A. H. Liu et al., 2020). Combined with the instance normalisation layer, their approach outperforms continuous counterparts. Both D. Wang et al. (2021) and Tang et al. (2022) employ the quantisation approach and propose further improvement in terms of disentanglement using auxiliary loss terms based on MI (Cheng et al., 2020) and metric learning (Hoffer et al., 2015).

Introducing Supervisory Signals

Lian et al. (2022a) and Lu et al. (2024) incorporate text data to construct a conditional prior and regularise the content latents. The rationale is to prevent the content latents from encoding acoustic information, thereby discouraging the leakage of speaker information. Wei et al. (2023) construct inputs to be pairs of utterances that share the same speaker and modify the prior distributions to be conditioned on inputs, thereby improving disentanglement and avoiding over-regularisation.

2.4.4 Summary of Limitations

Despite the variety of strategies developed to improve disentanglement, existing methods present several limitations. Models without a prior over the speaker latent cannot support unconditional generation or interpolation. Architectural constraints require extensive hyperparameter tuning. Quantisation improves disentanglement but complicates gradient propagation, posing challenges for optimisation and potentially degrading reconstruction quality (Bengio et al., 2013b; Jang et al., 2017; Maddison et al., 2017). Supervisory signals improve regularisation but require additional domain-specific resources.

Chapter 5 presents a simple and effective approach based on Gaussian dropout (Alemi et al., 2017; Rey et al., 2021) to regularise the content latents $\mathbf{z}_C$, where the

noise of the dropout is parameterised by the posterior distribution of the speaker latent $\mathbf{z}_S$. The approach does not introduce additional parameters or loss terms. Applying the dropout to existing models for zero-shot VC consistently improves disentanglement robustness with respect to hyperparameter choices.

Chapter 3

Improving Robustness of Unsupervised Disentanglement: A Two-Stage Training Strategy

As introduced in Section 2.2, the disentangled sequential autoencoder (DSAE) provides an architectural bias for separating global and local factors in sequential data. This chapter investigates the failure mode of posterior collapse of the global latent in the context of pitch-timbre disentanglement for musical audio.

To address this issue, TS-DSAE is proposed as a two-stage training framework that first encourages an informative global posterior and then uses this posterior as a sequence-specific prior during full training, together with auxiliary regularisation for disentanglement. The method remains fully unsupervised and avoids both adversarial training and domain-specific data augmentation. The proposed approach is evaluated on both synthetic and real-world monophonic music datasets using quantitative and qualitative analyses. [1]

[1] This chapter is based on work that was published in Luo et al. (2022). The implementation and audio samples are accessible at `https://github.com/yjlolo/dSEQ-VAE`.

3.1 Introduction

Building on the DSAE formulation reviewed in Section 2.2, this section analyses posterior collapse of the global latent variable and its impact on disentanglement. As discussed in Section 2.2.2, this collapse is especially likely when the local latent space has sufficient capacity to explain the observations without relying on the global latent. In that case, disentanglement between global and local factors becomes highly sensitive to architectural choices and hyperparameter settings.

Existing extensions to DSAE address this problem using additional objectives such as adversarial distribution matching, self-supervised learning based on domain-specific augmentations, or more complex generative parameterisations. While effective in some settings, these approaches either increase optimisation difficulty, reduce modality-agnostic applicability, or introduce additional design complexity.

To improve robustness while preserving the basic simplicity of the DSAE framework, this chapter proposes TS-DSAE, a two-stage training strategy supplemented by auxiliary regularisation terms that encourage factor invariance and manifestation. The first stage promotes an informative global posterior under a constrained bottleneck, and the second stage uses this posterior as a sequence-specific prior during full training. The results show that, unlike the vanilla DSAE, TS-DSAE consistently mitigates collapse of the global latent and supports more reliable disentanglement under varying model configurations.

3.2 Related Work

3.2.1 Structured Priors for Sequential Data

This section situates TS-DSAE relative to prior approaches that address the utilisation of global latent variables in sequential representation learning. Hsu et al. (2017) introduced the factorised hierarchical variational autoencoder (FHVAE), which constructs a hierarchical prior where each input is governed by a sequence-level prior atop a segment-level prior. The two-stage training framework shares this philosophy; however, it employs a strong bottleneck during the constrained training phase

to naturally promote a globally informative posterior. This posterior can then be directly employed as the sequence-level prior in the complete model training stage. In contrast, FHVAE initialises and learns the prior from scratch, lacking a strong inductive bias, and benefits from a discriminative objective function. Furthermore, the model amortises the learning of sequence-level priors via a global encoder, ensuring that memory consumption does not scale with the size of the training data, unlike FHVAE.

Despite being an elegant disentanglement framework, the vanilla DSAE (Y. Li et al., 2018) is prone to collapse of the global latent space, as discussed in Section 2.2.2. Han et al. (2021) introduced the recurrent Wasserstein autoencoder (R-WAE), which minimises the Wasserstein distance between the aggregated posterior and the prior. This approach employs either maximum mean discrepancy or generative adversarial networks, both of which present challenges in parameter tuning and optimisation. Other models, such as S3-VAE (Zhu et al., 2020) and C-DSVAE (Bai et al., 2021), exploit self-supervised learning and incorporate domain-specific ad hoc loss functions or data augmentation. In contrast, the proposed TS-DSAE operates without any form of supervision, adversarial training, or domain-dependent data augmentation.

3.2.2 Progressive Bottlenecks for Disentanglement

The variational deep state-space model (VDSM) employs a pretraining phase combined with KL-annealing to encourage the utilisation of the global latent space (B. Vowels et al., 2021). While this is similar in spirit to the constrained training stage used in this work, the proposed method differs by training only the global variable during pretraining and avoiding KL-annealing altogether, thereby reducing the reliance on extensive hyperparameter tuning. Additionally, VDSM uses n decoders, each dedicated to modelling a specific object identity in video data, with n manually specified based on the dataset. This design limits model generalisability, increases computational overhead, and may complicate optimisation.

To prevent a model from bypassing the target latent factors of interest for data reconstruction, Lezama (2019) propose a progressive autoencoder-based framework to address the "shortcut problem" in static data. The model is first trained with a low-

capacity latent space to focus on learning the factors of interest, after which the latent capacity is increased to improve reconstruction quality. The final stage incorporates supervision from human annotations to guide the learning of desired factors. While the two-stage training in TS-DSAE follows a similar philosophy, it differs in operating entirely without supervision and is specifically designed for sequential data.

3.2.3 Multi-View Representation Learning

The constrained training stage also draws parallels with multi-view representation learning. For instance, variational canonical correlation analysis (VCCA) (W. Wang et al., 2016) formulates a model in which different views of a common object are sampled from distributions conditioned on a shared latent variable. NestedVAE (M. Vowels et al., 2020) learns common factors through staged information bottlenecks, where a low-level VAE is trained conditioned on the latent space obtained from a high-level VAE. Similarly, in the proposed model, multiple time frames within a sequence are treated as distinct "views" of a shared underlying factor—specifically, the global latent variable.

3.2.4 Applications to Music Audio

The exploration of unsupervised disentangled representation learning for music audio remains limited. Both Luo et al. (2020) and Cífka et al. (2021) adopt self-supervised learning strategies to decorrelate instrument pitch and timbre. The latter, who also focus on monophonic melodies, employ pitch-shifting—a domain-specific augmentation—and constrain local capacity by learning discrete latent variables, potentially introducing optimisation challenges. In contrast, the proposed approach retains the simplicity of the DSAE framework while improving robustness in a straightforward and modality-agnostic manner.

3.3 Method

This section describes the proposed TS-DSAE framework. As illustrated in Fig. 3.1, the method combines a two-stage training strategy with auxiliary regularisation terms

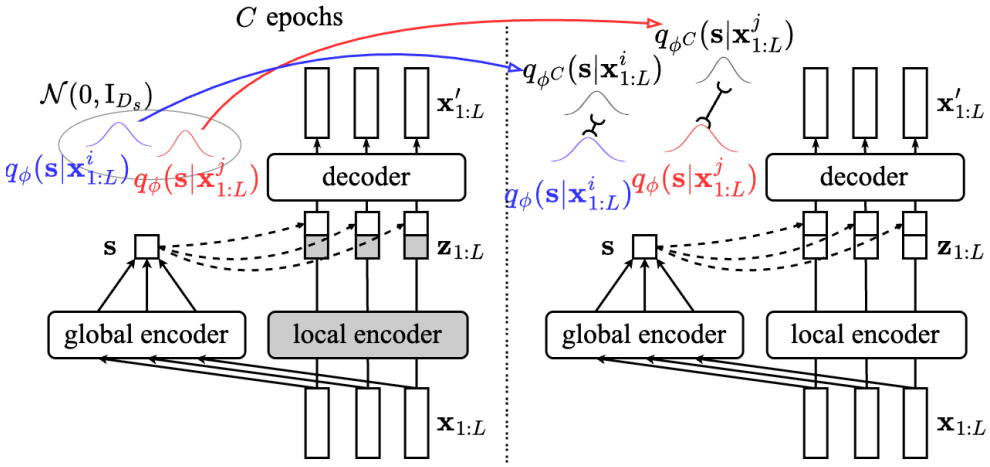

Figure 3.1: System diagrams of Two-Stage DSAE. *Left*: The constrained training stage where the local modules are frozen and the global latent follows a standard Gaussian prior. *Right*: The stage of informed-prior training where the global latent is regularised by a sequence-specific prior defined as the associated posterior learnt from the first stage. The dashed arrows denote broadcast along the time-axis.

that encourage the decoder to preserve and render the intended separation between global and local factors.

3.3.1 Two-Stage Training Framework

The collapse of the global latent, discussed in Section 2.2.2, can be ascribed to the simplicity of the uni-modal prior $p(\mathbf{s})$ which is not expressive enough to capture the multi-modal global factors. As a result, the approximate posterior $q_\phi(\mathbf{s}|\mathbf{x}_{1:L})$ becomes over-regularised. The issue is further exaggerated by the relatively capacity-rich local latents $\mathbf{z}_{1:L}$ which are allowed to carry information at the same frame resolution as $\mathbf{x}_{1:L}$. To mitigate the problem, the training is divided into two stages, *constrained training* and *informed-prior training*.

Stage One: Constrained Training

During constrained training, some parameters of the local module—specifically, the local encoder and the transition network—are frozen after initialisation. In this way, the local latents $\mathbf{z}_t$ resemble random projections of the input and are not optimised to capture the most informative aspects of the data. This strongly encourages the

decoder to focus on the global latent $\mathbf{s}$ for reconstruction.

As a result, $q_\phi(\mathbf{s}|\mathbf{x}_{1:L})$ is biased to capture the global factors that are shared across the entire sequence. From an optimisation perspective, this is equivalent to eliminating the second term (the KLD terms for $\mathbf{z}_t$) from (2.4).

This first training stage is reminiscent of multi-view representation learning (W. Wang et al., 2016; M. Vowels et al., 2020) where the goal is to learn a single representation shared by multiple views of the underlying image. The proposed method treats the sequence of frames as different "views" of the underlying static factors and does not additionally require any form of supervision to group data.

Stage Two: Informed-Prior Training

The training proceeds to the second stage after C epochs of constrained training. During this stage, all the model parameters are unfrozen and trained regularly using the full objective (2.4) with a modification.

In particular, the standard Normal prior $\mathcal{N}\big(\mathbf{0}, \mathbf{I}_{D_s}\big)$, previously configured for the global latent $\mathbf{s} \in \mathbb{R}^{D_s}$ during the first stage, is replaced by:

$$p_i(\mathbf{s}) = q_{\phi^C}(\mathbf{s}|\mathbf{x}^i_{1:L}), \tag{3.1}$$

where ϕ^C denotes the parameters of the global encoder at the C-th epoch. That is, each input sequence admits a sequence-specific prior that has been learnt from the first stage, whereby the KLD term for $\mathbf{s}$ in (2.4) is written as:

$$\mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{s}|\mathbf{x}^i_{1:L})\|p_i(\mathbf{s})\big) = \mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{s}|\mathbf{x}^i_{1:L})\|q_{\phi^C}(\mathbf{s}|\mathbf{x}^i_{1:L})\big). \tag{3.2}$$

Note that q_ϕ is differentiated from q_{ϕ^C} to emphasise that a "snapshot" of the global encoder, denoted as q_{ϕ^C}, is taken at the C-th epoch. This frozen network is used to parameterise the sequence-specific prior, while training continues with q_ϕ, initialised from ϕ^C.

In other words, the posterior continues to be trained while the distribution of each sequence is "anchored" to its associated prior—namely, the posterior obtained from constrained training—which is expected to capture sequence-level global factors. This

way, although the local module is introduced over the training, the global latent variables of sequences no longer commonly share the uni-modal prior, thereby mitigating the effect of over-regularisation.

3.3.2 Factor Invariance and Manifestation

Although the two-stage strategy aids the learning of the global latent representation, it does not explicitly enforce the decoder to utilise the encoded global latent. In practice, the decoder may still convey all necessary information through the local latents, effectively disregarding the global latent, even though the encoder produces global posteriors that align with the informative prior acquired during the first stage.

To address this limitation, auxiliary constraints are introduced to strengthen the influence of the global latent by encouraging the decoder to render the global latent in its output.

Consider the following scheme of inference, replacement, decoding, and inference: given the *inferred* variables $\mathbf{z}^i_{1:L} \sim q_\phi(\mathbf{z}_{1:L}|\mathbf{x}^i_{1:L})$ and $\mathbf{s}^i \sim q_\phi(\mathbf{s}|\mathbf{x}^i_{1:L})$, one can *replace* $\mathbf{s}^i$ with $\mathbf{s}^j$ inferred from another sequence j, and *decode* $\mathbf{x}^{\mathbf{s}^i \to \mathbf{s}^j}_{1:L} \sim p_\theta(\mathbf{x}_{1:L}|\mathbf{z}^i_{1:L}, \mathbf{s}^j)$. This generated sequence can then be passed through the encoders to *infer* $\mathbf{z}^{\mathbf{s}^i \to \mathbf{s}^j}_{1:L} \sim q_\phi(\mathbf{z}_{1:L}|\mathbf{x}^{\mathbf{s}^i \to \mathbf{s}^j}_{1:L})$ and $\mathbf{s}^{\mathbf{s}^i \to \mathbf{s}^j} \sim q_\phi(\mathbf{s}|\mathbf{x}^{\mathbf{s}^i \to \mathbf{s}^j}_{1:L})$. The replacement step can be applied to either of the two latent variables.

Invariance

If $\mathbf{z}_{1:L}$ and $\mathbf{s}$ have been successfully disentangled, the difference between $\mathbf{z}^{\mathbf{s}^i \to \mathbf{s}^j}_{1:L}$ and $\mathbf{z}^i_{1:L}$ would be minimal because replacing the global factor should not affect the subsequently inferred local factor. Similarly, replacing the local factor should not affect the subsequently inferred global factor—the difference between $\mathbf{s}^{\mathbf{z}^i_{1:L} \to \mathbf{z}^j_{1:L}}$ and $\mathbf{s}^i$ is therefore expected to be small.

One can impose the desired invariance by introducing the following regularisation

terms to (2.4):

$$-\mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{z}_{1:L}|\mathbf{x}_{1:L}^{\mathbf{s}^i\to\mathbf{s}^j})\|q_\phi(\mathbf{z}_{1:L}|\mathbf{x}_{1:L}^i)\big) \quad \text{and} \tag{3.3}$$

$$-\mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{s}|\mathbf{x}_{1:L}^{\mathbf{z}_{1:L}^i\to\mathbf{z}_{1:L}^j})\|q_\phi(\mathbf{s}|\mathbf{x}_{1:L}^i)\big). \tag{3.4}$$

Manifestation

On the other hand, replacing the global factor is supposed to render a change to the subsequent samples—$\mathbf{s}^{\mathbf{s}^i\to\mathbf{s}^j}$ should be close to $\mathbf{s}^j$ in order to faithfully manifest the swapping. Similarly, $\mathbf{z}_{1:L}^{\mathbf{z}_{1:L}^i\to\mathbf{z}_{1:L}^j}$ should be close to $\mathbf{z}_{1:L}^j$.

The following constraints are introduced to (2.4) to mirror the desired effect:

$$-\mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{s}|\mathbf{x}_{1:L}^{\mathbf{s}^i\to\mathbf{s}^j})\|q_\phi(\mathbf{s}|\mathbf{x}_{1:L}^j)\big) \quad \text{and} \tag{3.5}$$

$$-\mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{z}_{1:L}|\mathbf{x}_{1:L}^{\mathbf{z}_{1:L}^i\to\mathbf{z}_{1:L}^j})\|q_\phi(\mathbf{z}_{1:L}|\mathbf{x}_{1:L}^j)\big). \tag{3.6}$$

To summarise, invariance of the local and global latent variables is encouraged by maximising (3.3) and (3.4), respectively. Simultaneously, posteriors of the replaced factors are regularised to follow the target posteriors through maximising (3.5) and (3.6). In practice, each input sequence i in a mini-batch is paired with a randomly sampled input sequence j from the same mini-batch, and the described scheme of inference, replacement, decoding, and inference is applied. Notably, this procedure does not rely on any form of supervision or data augmentation.

While the above KLD terms encourage meaningful behaviour, they can still be minimised with a trivial global latent space, which is undesired. Thus, the two-stage training plays a crucial role in obtaining robust disentanglement. Further, note that the individual KLD terms above vary in terms of magnitude and thus importance to the gradient and so could benefit from balancing. However, scaling was found to be unnecessary for successful disentanglement in practice, and a detailed investigation is left for future work.

To summarise, TS-DSAE constitutes a two-stage training framework that facilitates the exploitation of additional divergences to achieve robust unsupervised disentanglement, as empirically validated in Section 3.5.

3.4 Experimental Setup

3.4.1 Datasets

Both an artificial and a real-world music audio dataset are considered. The former facilitates control over the underlying factors of variation, while the latter demonstrates the applicability of the proposed model to realistic data.

dMelodies

The artificial dataset is compiled by synthesising audio from monophonic symbolic music gathered from the dMelodies dataset (Pati et al., 2020). Each melody is a two-bar sequence with 16 eighth notes, subject to several global factors, i.e., tonic, scale, and octave, and local factors, i.e., direction of arpeggiation, and rhythm.

In order to facilitate analysis, the global factors are normalised by considering only the melodies of C Major in the fourth octave. Melodies starting or ending with a rest note are also discarded to avoid spurious amplitude values and boundary effects during audio synthesis with `FluidSynth`.[2] From the remaining melodies, 3k samples are randomly selected, split into 80% training and 20% validation sets, and synthesised at a sampling rate of 16kHz using violin and trumpet sound fonts from `MuseScore_General.sf3`.[3]

The amplitude of each audio sample is normalised with respect to its maximum value. The number of samples rendered with the two instruments is uniformly distributed.

URMP

For the real-world audio recordings, the violin and trumpet tracks from the URMP dataset (B. Li et al., 2019) are selected. The preprocessing procedure of Hayes et al. (2021) is followed, where the amplitude of each audio recording, resampled to 16kHz, is normalised in a corpus-wide fashion for each instrument subset. The audio samples are then divided into four-second segments, and segments with mean pitch confidence lower than 0.85 are discarded, as assessed by the full CREPE model (J. W. Kim et

[2] `https://www.fluidsynth.org/`
[3] `https://musescore.org/en/handbook`

al., 2018), a state-of-the-art pitch extractor. The process results in 1,545 violin and 534 trumpet samples in the training set, and 193 violin and 67 trumpet samples for validation.

Note that for both datasets, the underlying local and global factors are expected to correspond to melody and instrument identity, respectively. The audio samples are transformed and represented as log-amplitude mel-spectrograms with 80 mel frequency banks, derived from a short-time Fourier transform with a 128ms Hann window and a 16ms hop, yielding $\mathbf{x}_{1:T} \in \mathbb{R}^{80\times 251}$.

Separate models are trained and evaluated on each dataset independently. The artificial dataset serves as a setting for analysing a controlled set of factors of variation, including the analysis of an additional global factor introduced in Section 3.5.5, while the real-world dataset demonstrates applicability to more realistic audio.

We note that no separate held-out test set is used in these experiments, and evaluation is conducted on validation splits. As a result, the reported results do not constitute a strict assessment of generalisation performance. However, since the primary objective of the experiments is to compare the relative behaviour of different models under controlled settings, using a consistent validation protocol across all methods provides a meaningful basis for comparative analysis.

3.4.2 Implementation

Architecture

To verify the robustness of TS-DSAE with respect to unsupervised global–local disentanglement, various combinations of model architectures and hyperparameters are explored, including the dimensionality of the latent variables and the design choices of the encoder and decoder.

The notation `net-[layers]` is used to denote module architectures, where `net` indicates the type of network and `[layers]` specifies the number of neurons in each layer. `Tanh` is used as the non-linear activation between layers of fully connected networks (FCNs), and the long short-term memory (LSTM) is used for recurrent neural networks (RNNs). Let ψ_u denote the parameters of a Gaussian projection

layer `Gau-D` that maps hidden representations to the mean $\mu_{\psi_u}(\cdot)$ and log-variance $\log \sigma^2_{\psi_u}(\cdot)$ of a diagonal Gaussian distribution of dimensionality D_u. Specifically, ψ_u is defined for $u \in \{x, z, s\}$, corresponding to the Gaussian layers used in the decoder (ψ_x), the local encoder (ψ_z), and the global encoder (ψ_s), respectively.

Following standard VAE conventions, the decoder defines a diagonal Gaussian distribution $\mathcal{N}\big(\mu_{\psi_x}(\cdot), \mathbf{I}_{D_x}\big)$. Under this assumption, the log-likelihood term reduces to the squared ℓ_2 distance between the observation $\mathbf{x}_t$ and $\mu_{\psi_x}(\cdot)$, up to an additive constant and scaling factor.

The two inference networks proposed in the original DSAE (Y. Li et al., 2018), "factorised q" and "full q", are examined, as illustrated in Fig. 2.1. The "factorised q" implements the global encoder $q_\phi(\mathbf{s}|\mathbf{x}_{1:T})$ as `FCN-[64,64]-Avg-Gau-16`, where `Avg` denotes average pooling across the time axis, and the size of the global latent is fixed at $D_s = 16$ throughout the main experiments. The local encoder $q_\phi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$ parameterises a factorised posterior as `FCN-[64,64]-Gau-{8,16,32}`, with different sizes of the local latent investigated, $D_z \in \{8, 16, 32\}$.

For the transition network $p_\theta(\mathbf{z}_t|\mathbf{z}_{<t})$, `RNN-[32,32]-Gau-{8,16,32}` is used. The decoder $p_\theta(\mathbf{x}_t|\mathbf{z}_t, \mathbf{s})$ is `FCN-[64,64]-Gau-80` taking as input the concatenation of $\mathbf{z}_{1:T}$ and time-axis broadcast $\mathbf{s}$.

For "full q", $q_\phi(\mathbf{s}|\mathbf{x}_{1:T})$ follows that of "factorised q"; and $q_\phi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T}, \mathbf{s})$ corresponds to `biRNN-[64,64]-Gau-{8,16,32}` which takes as input the concatenation of $\mathbf{x}_{1:L}$ and time-axis broadcast $\mathbf{s}$ inferred from $q_\phi(\mathbf{s}|\mathbf{x}_{1:T})$. `biRNN` denotes a bi-directional LSTM, where the outputs of the forward and backward LSTM are averaged along the time-axis before the Gaussian layer. Both the transition network and decoder follow those of "factorised q". Note that the implementation can be considered as a minimal effort to realise both of the graphical models in Fig. 2.1, which allows for undistracted examination of the proposed method.

Optimisation

The implementation is based on `PyTorch v1.9.0`, and optimisation is performed using ADAM (Kingma et al., 2015) with default parameters learning rate $lr = 0.001$, and $[\beta_1, \beta_2] = [0.9, 0.999]$ without weight decay. A batch size of 128 is used, and models

are trained for up to 4k epochs. Early stopping is applied if (2.4) obtained from the validation set stops improving for 300 epochs. For models adopting the proposed two-stage training frameworks presented in Section 3.3, the number of epochs for the first stage is fixed to $C = 300$ across all configurations. This value was chosen by observing that the training and validation losses flatten beyond this point, which indicates convergence. It is not a tuned optimum, and performance is insensitive to the exact value.

3.5 Experiments and Results

Three baseline methods are considered: (1) *DSAE*; (2) *DSAE-f*, which employs constrained training and freezes the global encoder after C epochs; and (3) *TS-DSAE w/o regs*, which adopts the two-stage training framework but excludes the auxiliary loss terms (3.3)–(3.6).

Models discussed in Section 3.2 (Zhu et al., 2020; Bai et al., 2021; Han et al., 2021) are not included and are left for future work. The main focus is to improve upon DSAE (Y. Li et al., 2018) through minimal modifications, thereby providing a stronger and more general backbone that can be complementary to existing approaches.

3.5.1 Instrument Classification

Disentanglement is first evaluated through the task of instrument classification. In particular, a linear discriminant analysis (LDA) classifier is trained on the global latent variables $\mathbf{s} \sim q_\phi(\mathbf{s}|\mathbf{x}_{1:T})$ sampled from the training set, and evaluated on the validation set using the macro F1-score for instrument identity with a default decision threshold of 0.5. While the F1-score depends on the choice of threshold, the use of macro averaging mitigates the effect of class imbalance in datasets such as URMP, where violin samples are more prevalent than trumpet samples. Furthermore, the same evaluation protocol is applied consistently across all models, making it suitable for comparative analysis.

To assess the effect of latent manipulation, each validation sequence i is paired with another sequence j recorded with the other instrument, and the inference-

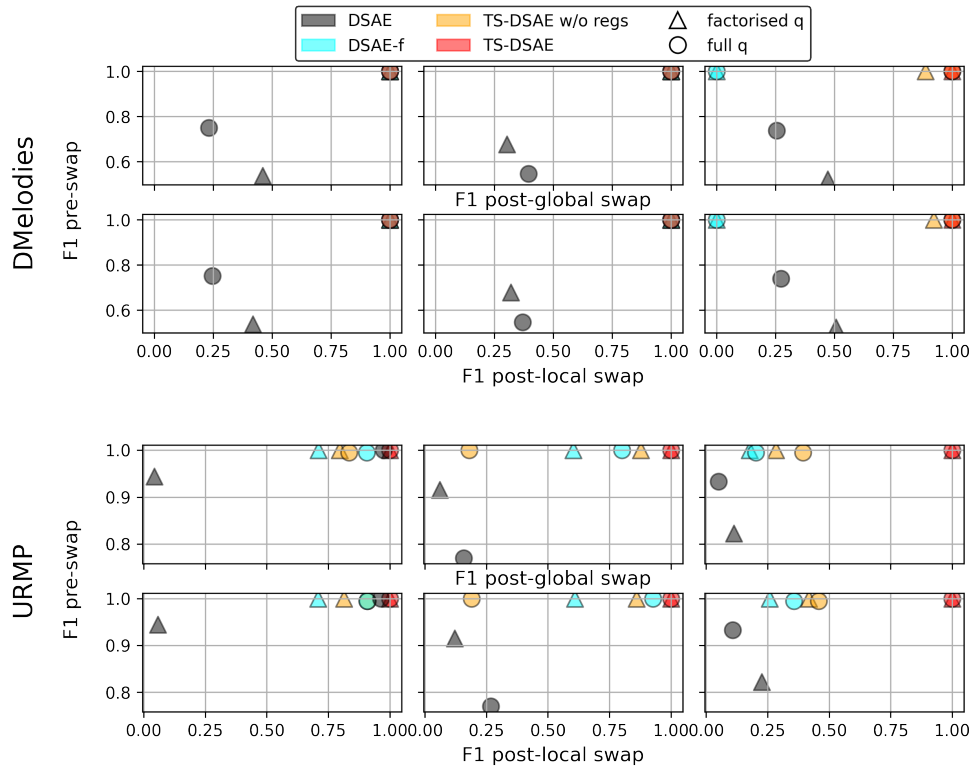

Figure 3.2: Macro F1 score of instrument classification derived from applying LDA to the global latent space. The vertical axis (*pre-swap*) shows scores based on global latents extracted before performing the replacement-decoding-inference scheme discussed in Section 3.3.2. The two horizontal axes, *post-global swap* and *post-local swap*, correspond to the global latents obtained after replacing the global and the local latents, respectively. Size of the local latent space increases from left to right columns, 8, 16, and 32, respectively. Top panel refers to the dMelodies dataset, while bottom panel corresponds to URMP.

replacement-decoding-inference scheme is applied. Following the notation in Section 3.3.2, $\mathbf{s}^{\mathbf{s}^i \to \mathbf{s}^j}$ should be predictive of the instrument of sample j, while $\mathbf{s}^{\mathbf{z}^i_{1:T} \to \mathbf{z}^j_{1:T}}$ should reflect the original instrument of sample i. Three metrics are reported: accuracy using the original global latents (*pre-swap*), after replacing the global latent (*post-global swap*), and after replacing the local latents (*post-local swap*). For LDA training, the mean of the Gaussian posterior $q_\phi(\mathbf{s}|\mathbf{x}_{1:T})$ is used as the feature representation.

The results are summarised in Fig. 3.2. The proposed TS-DSAE (red), with either factorised or full q, is consistently located at the top right corner of the plot, across all the sizes of the local latent space. This indicates its robust disentanglement as well

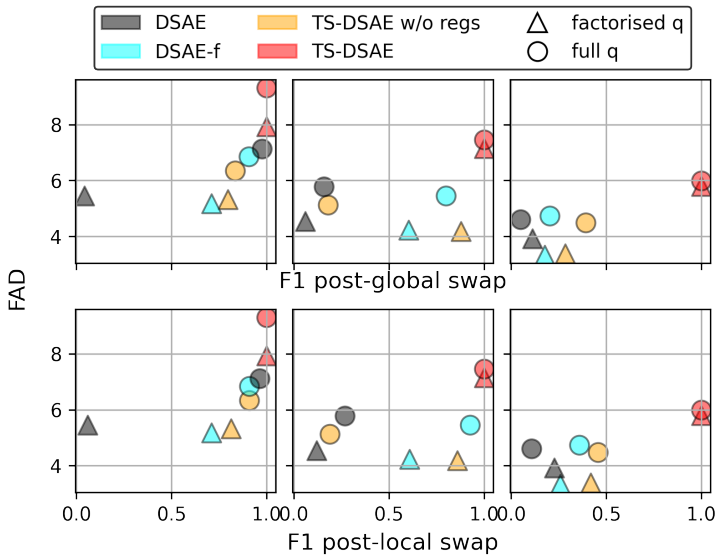

Figure 3.3: FAD (the lower the better) of reconstruction versus macro F1 score for instrument classification, evaluated using URMP. Left to right panels refer to size of the local latent 8, 16, and 32. See Section 3.5.2 for details.

as a linearly separable global latent space. From the left to right column, the competing methods DSAE-f (cyan) and TS-DSAE without the additional regularisations (orange) move from top right to left-hand side of the plot, showing the inclination for a collapsed global latent space with the increased local latent capacity. Being located at the lower left of the plot, DSAE (gray) attains the worst performance in most configurations. This highlights the issue of positing the standard Gaussian prior in the global latent space.

The overall high pre-swap and low post-swap F1 especially towards high-dimensional $\mathbf{z}_t$ implies that the decoder tends to ignore $\mathbf{s}$, even though the mean parameter of $q_\phi(\mathbf{s}|\mathbf{x}_{1:T})$ is discriminative with respect to the instrument identity. The competing models appear to suffer the most from the size of $\mathbf{z}_t$ as a large local latent space can easily capture all the necessary information for reconstruction.

3.5.2 Reconstruction Quality

The trade-off between disentanglement and reconstruction is examined using Fréchet Audio Distance (FAD) (Kilgour et al., 2019), a metric shown to correlate with perceptual audio quality. Specifically, multivariate Gaussians are computed on both a

set of embeddings derived from the generated data (post-global or post-local swap) $\mathcal{N}_g(\mu_g, \Sigma_g)$ and a set of embeddings extracted from the real data $\mathcal{N}_r(\mu_r, \Sigma_r)$. Dowson et al. (1982) show that the Fréchet distance between two Gaussians is given by

$$F(\mathcal{N}_r, \mathcal{N}_g) = \|\mu_r - \mu_g\|^2 + \mathrm{tr}(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r \Sigma_g}),$$

where tr denotes the trace of a matrix. Following Kilgour et al. (2019), a VGGish model is used to extract the embeddings from either the real or generated audio. [4]

Results are reported for the URMP dataset in Fig. 3.3, as both datasets yield comparable trends. As expected, FAD improves with increasing dimensionality of the local latent variable $\mathbf{z}_t$. However, TS-DSAE is the only model that overcomes this trade-off: while competing models exhibit a loss in disentanglement performance as FAD improves (moving from right to left in the plot), TS-DSAE maintains disentanglement alongside enhanced reconstruction quality.

Although TS-DSAE yields the worst FAD, it is the only model that preserves disentanglement as the size of the local latent space increases. The next chapter will build upon the two-stage training strategy and introduce an alternative regularisation to tackle the trade-off between reconstruction and disentanglement.

3.5.3 Raw Pitch Accuracy

This section evaluates the information captured in the local latents $\mathbf{z}_{1:T}$ by applying the full CREPE model (J. W. Kim et al., 2018) to audio re-synthesised from the mel-spectrogram. The re-synthesis is performed using `InverseMelScale` and `GriffinLim`, both available in `torchaudio v0.9.0`.

Following the notation from Section 3.3.2, pitch contours are extracted from reconstructed samples (pre-swap), $\mathbf{x}_{1:T}^{\mathbf{s}^i \to \mathbf{s}^j}$ (post-global swap) which is expected to retain the pitch contour of $\mathbf{x}_{1:T}^i$, and $\mathbf{x}_{1:T}^{\mathbf{z}_{1:T}^i \to \mathbf{z}_{1:T}^j}$ (post-local swap) which should reflect the pitch contour of $\mathbf{x}_{1:T}^j$. Note that for models with collapsed global latents $\mathbf{s}$, the accuracy of post-global swap will remain high as the decoder ignores the global latent and relies solely on $\mathbf{z}_{1:T}$.

[4] The implementation of FAD is available at `https://github.com/google-research/google-research/tree/master/frechet_audio_distance`.

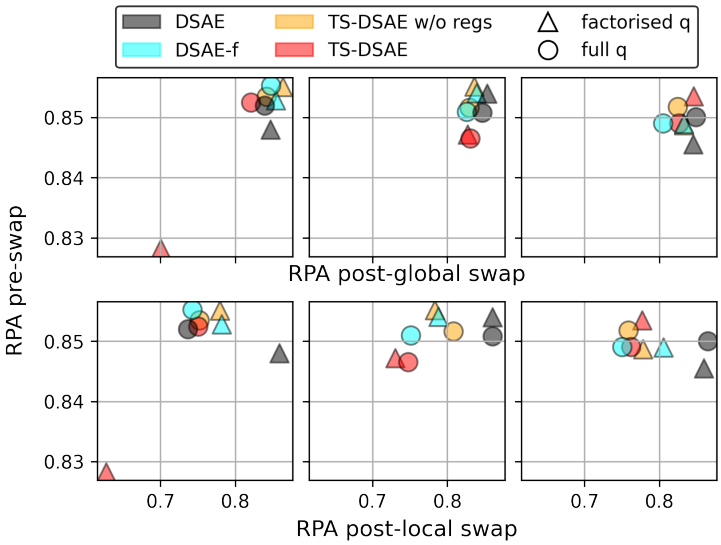

Figure 3.4: RPA evaluated using CREPE on re-synthesised audio from URMP. Left to right panels refer to size of the local latent 8, 16, and 32. See Section 3.5.3 for details.

Reference pitch contours are obtained by applying CREPE to the input audio, and the raw pitch accuracy (RPA) is reported with a 0.5-semitone threshold (Salamon et al., 2014). This introduces a limitation, as the evaluation depends on the accuracy of the pitch extractor rather than annotated ground-truth. However, CREPE has been shown to perform reliably on monophonic instrument sounds (J. W. Kim et al., 2018), which aligns with the characteristics of the datasets considered in this work, and the same extraction procedure is applied consistently across all models and conditions, such that any estimation errors are expected to be systematic and are unlikely to affect the relative comparison of methods. Because the reference comes from CREPE rather than verified annotation, CREPE's own errors cannot be detected or corrected. They affect all models equally, however, so the comparison between methods stays valid.

Results on the URMP dataset are reported in Fig. 3.4. TS-DSAE consistently improves in RPA with increasing $\mathbf{z}_t$ dimensionality. Apart from the post-local swap setting, TS-DSAE performs comparably with the competing models towards the larger $\mathbf{z}_t$, while simultaneously achieving effective disentanglement (Fig. 3.2 and Fig. 3.3).

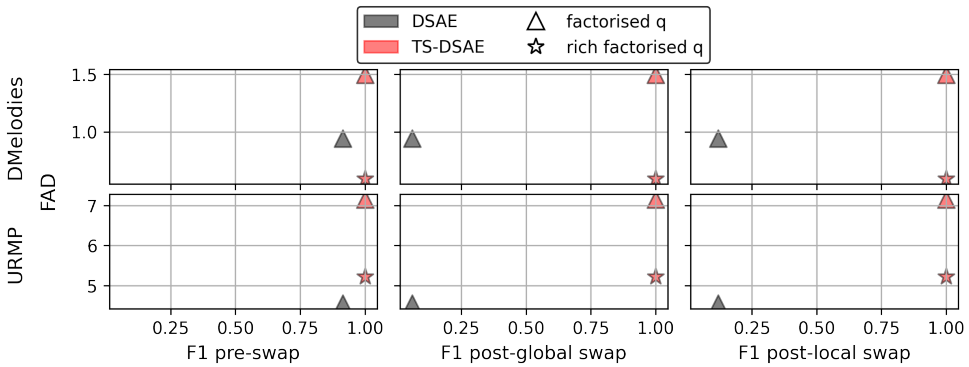

Figure 3.5: FAD (the lower the better) against disentanglement in terms of instrument classification. The triangles correspond to models configured with a factorised decoder, while the stars denote TS-DSAE enriched by an expressive decoder. All models share the same factorised inference network. Left to right panels refer to size of the local latent 8, 16, and 32. See Section 3.5.4 for details.

3.5.4 Richer Decoders

To further mitigate the trade-off between disentanglement and reconstruction, a richer decoder is constructed and evaluated. In this variant, the reconstruction of $\mathbf{x}_t$ is conditioned on the entire sequence of local latents $\mathbf{z}_{1:T}$, rather than on the current time frame $\mathbf{z}_t$. The size of $\mathbf{z}_t$ is fixed to 16, and the inference network adopts the "factorised q" configuration. Comparisons are made across DSAE, TS-DSAE, and TS-DSAE augmented with the enriched decoder.

As shown in Fig. 3.5, the enriched TS-DSAE maintains perfect accuracy for instrument classification for both datasets (indicated by the red triangles and stars), while improving FAD over its factorised-decoder counterpart (the red stars are positioned lower than the red triangles). On dMelodies, the enriched TS-DSAE outperforms DSAE equipped with the factorised decoder in terms of FAD. A more comprehensive evaluation across configurations—including autoregressive decoders, which are known to induce posterior collapse even in standard VAEs—is left for future work.

3.5.5 Multiple Global Factors

The dMelodies dataset is extended to include both the fourth and fifth octaves, introducing octave number as an additional global factor of variation alongside instrument identity. The decoder-enriched TS-DSAE, described in Section 3.5.4, is trained on

this dataset, and the results are presented in Fig. 3.6. In this experiment, the global latent $\mathbf{s}$ inferred from a source sample (bottom-left corner) is replaced with that from one of three target samples (top row). The resulting outputs, shown in the second to last columns of the bottom row, demonstrate the impact of this replacement on the generated audio.

Each sample is annotated using $\{i, o, m\}$ to denote instrument, octave, and melody, respectively. For example, the source sample $\{i_1, o_1, m_1\}$ and the first target $\{i_2, o_1, m_2\}$ share the same octave but differ in instrument identity, which is reflected in the spectral characteristics. Replacing $\mathbf{s}$ accordingly yields $\{i_{1\to2}, o_1, m_1\}$, where the instrument changes to i_2 while the octave and the melody remain fixed.

In the second case, the target $\{i_1, o_2, m_3\}$ differs from the source only in octave, characterised by the level of pitch contour. Swapping $\mathbf{s}$ produces $\{i_1, o_{1\to2}, m_1\}$, indicating a successful transformation of octave while preserving instrument and melody.

Finally, using the third target $\{i_2, o_2, m_4\}$—which shares no attributes with the source—results in $\{i_{1\to2}, o_{1\to2}, m_1\}$, where both instrument and octave are modified. In all three cases, the melody m_1 remains unchanged, demonstrating successful global–local disentanglement.

3.6 Summary

This chapter has presented TS-DSAE, a robust framework for unsupervised disentanglement of sequential data, which consistently performs well across a broad range of settings. The evaluation has primarily focused on the model's ability to achieve robust disentanglement. Further exploration into its capacity for multi-modal data generation via unconditional prior sampling is left for future work. Additionally, investigating the applicability of TS-DSAE to modalities beyond music audio remains an important direction.

The probabilistic graphical model of DSAEs assumes a single global latent variable fixed over time for each input sequence. While effective in structured domains, this assumption may be overly restrictive for more general scenarios where global factors evolve slowly over time. One promising relaxation is to introduce a hierarchy

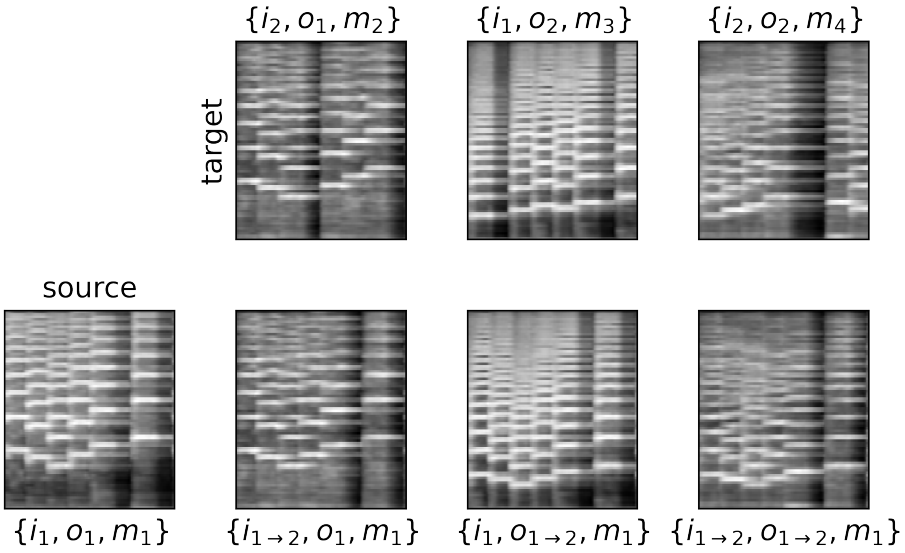

Figure 3.6: Global latent replacement using the top three samples as targets and the bottom-left sample as the source. Each sample is annotated with instrument (i), octave (o), and melody (m). Replacing the global latent $\mathbf{s}$ transforms instrument and/or octave while preserving the source melody, demonstrating global–local disentanglement. See Section 3.5.5 for details.

of latent variables operating at multiple temporal resolutions (Saxena et al., 2021). A natural extension of the two-stage training strategy would be to adopt multiple stages of constrained training with gradually increasing network capacity, enabling such hierarchical representations. This can be viewed as a temporal analogue of the progressive framework proposed by Z. Li et al. (2020).

Despite the substantial improvement in robustness, the challenge of balancing disentanglement and reconstruction persists. Although differential scaling of the regularisation terms (3.3)–(3.6) may help, as discussed in Section 3.3.2, the next chapter proposes a teacher–student framework that simplifies the objective by replacing the four auxiliary terms with a single constraint.

Chapter 4

Addressing the Trade-Off between Reconstruction and Disentanglement: A Teacher-Student Framework

As shown in the previous chapter, the disentangled sequential autoencoder (DSAE) (Y. Li et al., 2018) provides an effective framework for learning global and local latent representations from sequential data. However, it exhibits a tendency toward posterior collapse, where the global latent variable is underutilised, especially in the presence of high-capacity local latents. Moreover, the model is sensitive to architectural hyperparameters and often requires careful tuning to achieve reliable disentanglement.

The two-stage disentangled sequential autoencoder (TS-DSAE), introduced in the previous chapter, addresses these issues by incorporating a staged training strategy and auxiliary constraints that explicitly encourage the global latent to encode invariant information. This design significantly improves the robustness of disentanglement across architectures and hyperparameter settings.

However, TS-DSAE achieves this robustness at the cost of degraded reconstruction

fidelity. Although the auxiliary losses are effective in preventing collapse, they may over-constrain the model, reducing its capacity to accurately reconstruct audio signals. This raises a natural question: can the disentanglement benefits of TS-DSAE be retained while improving reconstruction quality?

This chapter introduces the Jacobian disentangled sequential autoencoder (J-DSAE), which preserves the two-stage optimisation strategy of TS-DSAE but shifts the regularisation from the latent space to the observation space. Specifically, a Jacobian-based constraint is incorporated that encourages changes in the global latent space to induce corresponding variations in the output signal. As a result, the decoder is constrained to acknowledge perturbations applied in the global latent space, thereby preventing collapse and underutilisation of the global latent variable. [1]

4.1 Introduction

For sequential data, the DSAE framework (reviewed in Section 2.2) models observations using global and local latent variables operating at different temporal resolutions, and has been widely applied to both speech and musical audio.

As discussed in Section 2.2.2, a critical failure mode of DSAEs is posterior collapse, whereby the decoder reconstructs the sequence using only the local latents, leaving the global latent unused. In the previous chapter, TS-DSAE was proposed to counteract this behaviour through latent-space regularisation and a two-stage training schedule. While effective, these additional constraints can reduce reconstruction quality by overly limiting the model's flexibility.

To address this trade-off, J-DSAE is proposed, which shifts regularisation into the observation space using a Jacobian-based constraint. Instead of enforcing global latent usage through latent-space penalties, the model is encouraged to reconstruct audio in a manner that reflects perturbations applied to the global latent variable.

Following the teacher–student paradigm, a compact teacher model is trained to capture the variation of the reconstructed signal with respect to the global latent. The teacher is not required to reconstruct the audio signal in detail, but only to

[1] This chapter is based on work published in Luo et al. (2024c).

manifest the influence of the global latent. The student decoder is then trained to reproduce the same patterns of variation while optimising for audio reconstruction. This formulation ensures that the student decoder produces outputs that both respect global latent structure and maintain audio quality.

J-DSAE is evaluated on real-world monophonic musical recordings from the URMP dataset (B. Li et al., 2019), targeting the disentanglement of pitch (local) and timbre (global). Compared to TS-DSAE, J-DSAE achieves improved synthesis quality, as measured by Fréchet Audio Distance (FAD) (Kilgour et al., 2019), and stronger pitch disentanglement, as evaluated by raw pitch accuracy (RPA) (Salamon et al., 2014). These results highlight the method's effectiveness in overcoming the reconstruction–disentanglement trade-off.

4.2 Related Work

4.2.1 Disentangled Sequential Autoencoder

The DSAE models sequential data by learning both a global latent variable $\mathbf{s}$ and a sequence of local latent variables $\mathbf{z}_{1:L}$, optimised via (2.4). Following the parameterisation described in Section 2.2.1, the dynamic prior $p_\theta(\mathbf{z}_{1:L}) = \prod_{t=1}^{L} p_\theta(\mathbf{z}_t|\mathbf{z}_{<t})$ is factorised as a product of diagonal Gaussians and implemented using an autoregressive model. The static prior $p(\mathbf{s})$ is assumed to be a standard Gaussian. The joint posterior $q_\phi(\mathbf{z}_{1:L}, \mathbf{s}|\mathbf{x}_{1:L})$ is factorised into two diagonal Gaussians, each parameterised by a separate neural network encoder to facilitate the disentanglement of $\mathbf{s}$ from $\mathbf{z}_{1:L}$.

Although the difference in temporal resolution introduces a strong inductive bias towards disentanglement, the model often encodes most information into the local latent variables, leaving the global latent underutilised. Empirical evidence of this posterior collapse, obtained using recordings of musical instruments, is presented in Section 3.5, while similar observations for speech data have been reported by Lian et al. (2022b) and Lu et al. (2022).

Section 2.2 surveys existing extensions to the vanilla DSAE, which typically rely on either domain-specific self-supervised signals (Bai et al., 2021; Naiman et al., 2023;

Zhu et al., 2020) or extensive hyperparameter tuning (Han et al., 2021; Berman et al., 2024).

4.2.2 Two-Stage Disentangled Sequential Autoencoder

Chapter 3 introduces TS-DSAE as a solution to the underutilisation of the global latent variable. In the first stage, the local encoder $q_\phi(\mathbf{z}_{1:L}|\mathbf{x}_{1:L})$ is frozen, encouraging the model to rely on $\mathbf{s}$ for reconstructing the sequence. In the second stage, the model resumes training under the full ELBO (2.4), but with the static prior $p(\mathbf{s})$ replaced by a sequence-dependent prior $p_i(\mathbf{s}) = q_{\phi^C}(\mathbf{s}|\mathbf{x}^i_{1:L})$ for each sequence $\mathbf{x}^i_{1:L}$, where ϕ^C denotes the encoder parameters obtained at the end of the first stage after C epochs of training. This approach preserves the information encoded in $\mathbf{s}$ during the initial training phase.

As detailed in Section 3.3.2, disentanglement is further encouraged via auxiliary regularisation terms (3.3)–(3.6) applied to both static and dynamic latent variables. By perturbing one latent factor while keeping the other fixed, these constraints enforce invariance in the unperturbed factor and promote the model to be sensitive to the perturbed one.

As demonstrated in Section 3.5, TS-DSAE improves the robustness of disentanglement across a wide range of model configurations without relying on domain-specific signals or extensive tuning. However, the auxiliary constraints may be overly strong, resulting in reduced sampling fidelity. This trade-off affects both reconstruction quality and timbre similarity in downstream tasks such as attribute translation.

4.2.3 Jacobian Constraint

Lezama (2019) proposes a teacher-student framework that splits autoencoder training into two stages. Conceptually, this shares similarities with the staged training strategy presented in the previous chapter. The key idea is to first isolate the factors of interest using a compact latent space—thereby encouraging the model to fully exploit these dimensions—and then expand the latent space to incorporate nuisance factors and improve reconstruction quality.

More specifically, given images of human faces, a low-capacity deterministic autoencoder (the *teacher*) learns latent representations $\mathbf{s} \in \mathbb{R}^{D_s}$ that are predictive of annotated facial attributes via supervised learning. A student network is then constructed by augmenting the teacher with an additional latent variable $\mathbf{z} \in \mathbb{R}^{D_z}$, where $D_z > D_s$, to enhance reconstruction fidelity. Crucially, the student model is regularised using a Jacobian constraint:

$$\frac{\partial \hat{\mathbf{x}}^{\mathrm{Tea}}}{\partial \mathbf{s}^{\mathrm{Tea}}} = \frac{\partial \hat{\mathbf{x}}^{\mathrm{Stu}}}{\partial \mathbf{s}^{\mathrm{Stu}}}, \tag{4.1}$$

where $\hat{\mathbf{x}}$ and $\mathbf{s}$ denote the reconstructions and latent variables of the teacher (superscript Tea) and student (superscript Stu) networks, respectively.

That is, the Jacobian—representing how the output varies with respect to perturbations in the target factors of interest—is required to match between teacher and student. In Eq. (2)–(7) of Lezama (2019), they show that (4.1) can be approximated by minimising a loss function $\mathcal{L}_{\mathrm{Jac}}$ that regularises the student model:

$$\begin{aligned} \mathcal{L}_{\mathrm{Jac}}(E_u^{\mathrm{Stu}}, D^{\mathrm{Stu}}; \mathbf{x}^i, \mathbf{x}^j) = \|\mathbf{s}^{\mathrm{Stu},\,i} - \mathbf{s}^{\mathrm{Tea},\,i}\|_2^2 \\ + \|(D^{\mathrm{Tea}}(\mathbf{s}^{\mathrm{Tea},\,i}) - D^{\mathrm{Tea}}(\mathbf{s}^{\mathrm{Tea},\,j})) \\ - (D^{\mathrm{Stu}}(\mathbf{s}^{\mathrm{Stu},\,i}, \mathbf{z}^i) - D^{\mathrm{Stu}}(\mathbf{s}^{\mathrm{Stu},\,j}, \mathbf{z}^i))\|_2^2, \end{aligned} \tag{4.2}$$

where i and j denote the indices of a randomly selected pair of data points. E_u^{M} and D^{M} denote the encoder-decoder pairs, where the superscript $\mathrm{M} \in \{\mathrm{Tea}, \mathrm{Stu}\}$ identifies the model type, and the subscript $u \in \{s, z\}$ distinguishes between encoders for the target and nuisance factors. The corresponding latent variables are defined as:

$$\mathbf{s}^{\mathrm{M},\,i} = E_s^{\mathrm{M}}(\mathbf{x}^i) \quad \text{and} \quad \mathbf{z}^i = E_z^{\mathrm{Stu}}(\mathbf{x}^i). \tag{4.3}$$

Note that $\mathbf{s}$ can be computed from either the teacher or student encoder, while $\mathbf{z}$ is exclusive to the student model.

The Jacobian constraint (4.2) is applied during the second stage, where only the student network is optimised. Initially, the teacher encoder-decoder pair $(E_s^{\mathrm{Tea}}, D^{\mathrm{Tea}})$ is trained to extract informative global factors $\mathbf{s}^{\mathrm{Tea}}$, ensuring that D^{Tea} is sensitive to

their perturbations. Subsequently, the student network is trained with the Jacobian constraint: the first term in (4.2) aligns the student's global latent with that of the teacher, while the second term enforces that the variation of the reconstructed output due to $\mathbf{s}$ is consistent across both networks. Lezama (2019) demonstrates that this approach outperforms adversarial baselines in terms of reconstruction quality.

TS-DSAE and the Jacobian method (Lezama, 2019) share a common strategy: prioritising the learning of a disentangled latent representation, which then guides the subsequent training of the full model. However, there are important differences: TS-DSAE is fully unsupervised, models sequential data, and is grounded in a probabilistic framework, whereas Lezama (2019) uses supervised training, deterministic autoencoders, and static image data. The next section highlights these connections and extends TS-DSAE with a Jacobian constraint to address the trade-off between reconstruction quality and unsupervised disentanglement.

4.3 Method

J-DSAE is proposed as a model that combines TS-DSAE, introduced in the previous chapter, with the Jacobian constraint described in Section 4.2.3. The two-stage training strategy employed in TS-DSAE can be reformulated as a teacher-student framework, which naturally accommodates the integration of the Jacobian constraint.

4.3.1 TS-DSAE as a Teacher-Student Framework

In the first stage of TS-DSAE, the global encoder and decoder can be defined as:

$$E_s^{\text{Tea}}(\mathbf{x}_{1:L}) \coloneqq q_{\phi^C}(\mathbf{s}|\mathbf{x}_{1:L}) \quad \text{and} \quad D^{\text{Tea}}(\mathbf{s}) \coloneqq p_{\theta^C}(\mathbf{x}_{1:L}|\mathbf{0}, \mathbf{s}), \tag{4.4}$$

where ϕ^C and θ^C are the parameters obtained at the final epoch C of the first stage. That is, the encoder-decoder pair optimised during the first stage serves as the teacher network. Note that in J-DSAE, the local latents are simply set to $\mathbf{z}_{1:L} = 0$ at this stage, rather than freezing the local encoder as in TS-DSAE. This is found to be equally effective in encouraging the model to encode informative global latents $\mathbf{s}$.

Similarly, the encoders and decoder used in the second stage of TS-DSAE can be interpreted as components of the student network:

$$\begin{aligned} &E_s^{\mathrm{Stu}}(\mathbf{x}_{1:L}) \coloneqq q_\phi(\mathbf{s}|\mathbf{x}_{1:L}), \quad E_z^{\mathrm{Stu}}(\mathbf{x}_{1:L}) \coloneqq q_\phi(\mathbf{z}_{1:L}|\mathbf{x}_{1:L}); \\ &D^{\mathrm{Stu}}(\mathbf{s}, \mathbf{z}_{1:L}) \coloneqq p_\theta(\mathbf{x}_{1:L}|\mathbf{z}_{1:L}, \mathbf{s}). \end{aligned} \tag{4.5}$$

These definitions reveal the connection between TS-DSAE and the teacher-student framework: the first stage yields a teacher VAE, which guides the training of the student DSAE in the second stage through the replacement of the global prior, as given in (3.1):

$$p_i(\mathbf{s}) = q_{\phi^C}(\mathbf{s}|\mathbf{x}^i_{1:L}). \tag{4.6}$$

With this formulation, the Kullback–Leibler divergence (KLD) term defined in (3.2) becomes:

$$\mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{s}|\mathbf{x}^i_{1:L}) \| p_i(\mathbf{s})\big) = \mathcal{D}_{\mathrm{KL}}\big(E_s^{\mathrm{Stu}}(\mathbf{x}^i_{1:L}) \| E_s^{\mathrm{Tea}}(\mathbf{x}^i_{1:L})\big). \tag{4.7}$$

This connection allows us to directly adapt (4.2) to regularise the second stage of TS-DSAE:

$$\begin{aligned} \mathcal{L}_{\mathrm{Jac}}(E_s^{\mathrm{Stu}}, E_z^{\mathrm{Stu}}, D^{\mathrm{Stu}}; \mathbf{x}^i_{1:L}, \mathbf{x}^j_{1:L}) = \mathcal{D}_{\mathrm{KL}}\big(E_s^{\mathrm{Stu}}(\mathbf{x}^i_{1:L}) \| E_s^{\mathrm{Tea}}(\mathbf{x}^i_{1:L})\big) \\ + \| (D^{\mathrm{Tea}}(\mathbf{s}^{\mathrm{Tea},\,i}) - D^{\mathrm{Tea}}(\mathbf{s}^{\mathrm{Tea},\,j})) \\ - (D^{\mathrm{Stu}}(\mathbf{s}^{\mathrm{Stu},\,i}, \mathbf{z}^i_{1:L}) - D^{\mathrm{Stu}}(\mathbf{s}^{\mathrm{Stu},\,j}, \mathbf{z}^i_{1:L}))\|_2^2. \end{aligned} \tag{4.8}$$

The latent variables are sampled from the approximate posteriors defined by the encoders in (4.4) and (4.5):

$$\mathbf{s}^{\mathrm{M},\,i} \sim E_s^{\mathrm{M}}(\mathbf{x}^i_{1:L}) \quad \text{and} \quad \mathbf{z}^i_{1:L} \sim E_z^{\mathrm{Stu}}(\mathbf{x}^i_{1:L}), \tag{4.9}$$

rather than being deterministically extracted, as in (4.3).

The teacher network $(E_s^{\mathrm{Tea}}, D^{\mathrm{Tea}})$ is frozen after C training epochs and is used to regularise the student network $(E_s^{\mathrm{Stu}}, E_z^{\mathrm{Stu}}, D^{\mathrm{Stu}})$. From the second stage onwards, while E_z^{Stu} is trained from scratch, E_s^{Stu} and D^{Stu} are initialised by the parameters of E_s^{Tea} and D_s^{Tea}, respectively. The probabilistic formulation replaces the mean squared

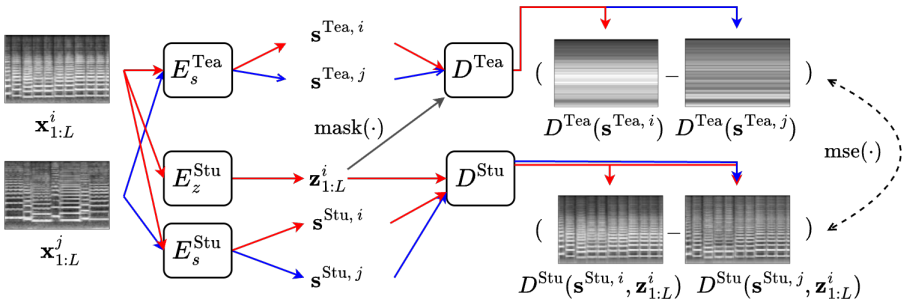

Figure 4.1: Illustration of the Jacobian constraint adapted in this approach (4.8), denoted by $\text{mse}(\mathbf{x}_{1:L})$. In practice, for each sequence $\mathbf{x}^i_{1:L}$ from a minibatch, a paired sequence $\mathbf{x}^j_{1:L}$ is randomly sampled from the same batch and is not required to share or differ in specific attributes. Hence, **the framework remains fully unsupervised**.

error (the first term in (4.2)) with a divergence term between distributions (the first term in (4.8)). The second term in (4.8) is illustrated in Fig. 4.1: the Jacobian is computed from a *random* pair of samples drawn from a mini-batch, and no explicit supervision is required. The variation in the observation domain induced by $\mathbf{s}^{\text{M},\,i}$ and $\mathbf{s}^{\text{M},\,j}$ is constrained to be consistent across the teacher and student networks.

4.3.2 Optimising J-DSAE

To summarise, J-DSAE first optimises the teacher network for C epochs by maximising the following objective:

$$\begin{aligned}\mathcal{L}_{\text{Tea}}(\theta, \phi; \mathbf{x}^i_{1:L}) = \mathbb{E}_{q_\phi(\mathbf{z}_{1:L}, \mathbf{s}|\mathbf{x}^i_{1:L})}\big[\log p_\theta(\mathbf{x}^i_{1:L}|\mathbf{0}, \mathbf{s}^i)\big] \\ - \mathcal{D}_{\text{KL}}\big(q_\phi(\mathbf{z}_{1:L}|\mathbf{x}^i_{1:L})\|p_\theta(\mathbf{z}_{1:L})\big) - \mathcal{D}_{\text{KL}}\big(q_\phi(\mathbf{s}|\mathbf{x}^i_{1:L})\|p(\mathbf{s})\big),\end{aligned} \tag{4.10}$$

and then proceeds to optimise the student network by maximising:

$$\begin{aligned}\mathcal{L}_{\text{Stu}}(\theta, \phi; \mathbf{x}^i_{1:L}, \mathbf{x}^j_{1:L}) = \mathbb{E}_{q_\phi(\mathbf{z}_{1:L}, \mathbf{s}|\mathbf{x}^i_{1:L})}\big[\log p_\theta(\mathbf{x}^i_{1:L}|\mathbf{z}^i_{1:L}, \mathbf{s}^i)\big] \\ - \mathcal{D}_{\text{KL}}\big(q_\phi(\mathbf{z}_{1:L}|\mathbf{x}^i_{1:L})\|p_\theta(\mathbf{z}_{1:L})\big) - \mathcal{L}_{\text{Jac}}(E^{\text{Stu}}_s, E^{\text{Stu}}_z, D^{\text{Stu}}; \mathbf{x}^i_{1:L}, \mathbf{x}^j_{1:L}).\end{aligned} \tag{4.11}$$

Crucially, beyond incorporating $\mathcal{L}_{\text{Jac}}$, J-DSAE differs from TS-DSAE in that it removes the auxiliary loss terms defined in (3.5)–(3.6). As shown in Section 4.6, this simplification leads to improved sampling quality and disentanglement performance.

Unlike the original method proposed by Lezama (2019), described in Section 4.2.3, J-DSAE applies the Jacobian constraint to sequential data rather than static images, exploits the inductive bias and staged training strategy of TS-DSAE, operates in a fully unsupervised setting without explicit labels, and optimises a VAE with prior distributions instead of a deterministic AE.

4.4 Experiment

Following Section 3.4, a subset of the URMP dataset (B. Li et al., 2019) is used for the experiments. It comprises 1,545 (193) violin and 534 (67) trumpet training (validation) samples. Preprocessing of the dataset closely follows the procedure described in the previous chapter. Audio recordings are first resampled to 16 kHz and normalised across the corpus per instrument. They are then divided into four-second segments and converted into mel spectrograms with 80 frequency banks. The transformation is based on a short-time Fourier transform with a Hann window of 128 ms and a hop size of 16 ms, resulting in $\mathbf{x}_{1:L} \in \mathbb{R}^{80\times 251}$. Given monophonic instrument recordings as input, the global latent variable $\mathbf{s}$ is expected to capture timbre-related features, while the local latent sequence $\mathbf{z}_{1:L}$ is expected to encode time-varying pitch information.

4.4.1 Implementation

To effectively verify the impact of the proposed Jacobian constraint, the architecture of the global and local encoders in J-DSAE closely mirrors that of TS-DSAE described in Section 3.4.2. Unlike the previous chapter, which explores various combinations of latent dimensionalities, (D_s, D_z) is fixed to $(16, 32)$ throughout the evaluation. This configuration was previously identified as the most challenging for the original DSAE, as it makes the global latent especially prone to collapse due to the relatively high capacity of the local latents.

In addition to the factorised decoder ($D^{\mathrm{M}}_{\mathrm{fac}}$) used in the previous chapter, an autoregressive (AR) decoder ($D^{\mathrm{M}}_{\mathrm{ar}}$) is implemented, which reconstructs the current observation $\mathbf{x}_t$ conditioned on prior observations, with the aim of further enhancing

reconstruction quality:

$$
\begin{aligned}
D_{\text{fac}}^{\text{M}} &\coloneqq \prod_{t=1}^{L} p_\theta(\mathbf{x}_t|\mathbf{s},\ \mathbb{1}_{\text{M=Stu}} \cdot \mathbf{z}_t) \quad \text{and} \\
D_{\text{ar}}^{\text{M}} &\coloneqq \prod_{t=1}^{L} p_\theta(\mathbf{x}_t|\mathbf{s},\ \mathbb{1}_{\text{M=Stu}} \cdot [\mathbf{x}_{<t};\ \mathbf{z}_{\leq t}]), \\
\text{where } \mathbb{1}_{M=\text{Stu}} &= \begin{cases} 1 & \text{if M} = \text{Stu} \\ 0 & \text{if M} = \text{Tea} \end{cases}
\end{aligned}
\tag{4.12}
$$

Here, D_*^{Stu} continues training from the initial parameters of D_*^{Tea} after the first C epochs, as established in the staged teacher-student framework described in Sections 4.3.1 and 4.3.2. The optimisation procedure follows closely to that of the previous chapter, using the ADAM optimiser with a learning rate $lr = 0.001$ and a batch size of 128. Training is capped at 4k epochs with an early stop triggered if $\mathcal{L}_{\text{Stu}}$ (4.11) evaluated from the validation set fails to improve for 300 epochs. The number of epochs for the teacher training stage is fixed at $C = 300$, consistent with the configuration used in the TS-DSAE experiments.

4.5 Evaluation Protocol

To ensure consistency with the evaluation procedure outlined in Section 3.5, disentanglement in J-DSAE is assessed from three complementary perspectives: instrument classification, pitch accuracy, and reconstruction fidelity. Each perspective follows the paradigm of infer-swap-reconstruct as previously introduced.

4.5.1 Instrument Classification

The presence of timbre-related information in the global latent variable $\mathbf{s}$ is evaluated via an instrument classification task applied to post-swap reconstructions. A classifier based on convolutional neural networks is trained using the original mel spectrograms from the training set. This classifier is then applied to the generated mel spectrograms from the validation set, where latent variables underlying these samples have been selectively replaced. Following the protocol of Section 3.5.1, given a pair of data (i, j),

two types of replacements are considered:

- *z-swap:* The local latent sequence $\mathbf{z}^i_{1:L}$ is replaced with $\mathbf{z}^j_{1:L}$, while the global latent $\mathbf{s}^i$ is kept fixed, resulting in $\mathbf{x}^{\mathbf{z}^i_{1:L} \to \mathbf{z}^j_{1:L}}_{1:L}$.

- *s-swap:* The global latent $\mathbf{s}^i$ is replaced with $\mathbf{s}^j$, while $\mathbf{z}^i_{1:L}$ is retained, producing $\mathbf{x}^{\mathbf{s}^i \to \mathbf{s}^j}_{1:L}$.

The classifier is evaluated on each swapped reconstruction. For z-swap, the target label corresponds to the instrument identity of sample i; for s-swap, the target is the instrument identity of sample j. Disentanglement is reflected by alignment between classifier predictions and these ground truths. Evaluation is conducted using the macro F1-score with a fixed decision threshold of 0.5, which introduces limitations associated with threshold-dependent metrics, as discussed in Section 3.5.1, but is sufficient for the comparative evaluation considered in this work.

4.5.2 Raw Pitch Accuracy

Pitch-related content encoded in $\mathbf{z}_{1:L}$ is evaluated by extracting pitch contours from reconstructed waveforms using the full CREPE model (J. W. Kim et al., 2018). Mel spectrograms are converted back to waveforms using `InverseMelScale` and `GriffinLim`, as implemented in `torchaudio` v0.9.0. Two types of the reconstructed waveform are evaluated in terms of the raw pitch accuracy (RPA) computed using a 50-cent tolerance threshold (Salamon et al., 2014):

- *s-swap:* Pitch is extracted from $\mathbf{x}^{\mathbf{s}^i \to \mathbf{s}^j}_{1:L}$ and compared to the pitch of sample i.

- *z-swap:* Pitch is extracted from $\mathbf{x}^{\mathbf{z}^i_{1:L} \to \mathbf{z}^j_{1:L}}_{1:L}$ and compared to the pitch of sample j.

Using CREPE to extract pitch contours from both model outputs and original input audio introduces limitations associated with pitch extraction accuracy, as discussed in Section 3.5.3. However, it remains appropriate for the comparative analysis considered in this work.

4.5.3 Reconstruction Quality

Following the approach described in Section 3.5.2, reconstruction and sampling quality are assessed using Fréchet Audio Distance (FAD) (Kilgour et al., 2019), a metric shown to correlate with perceptual similarity. The following four conditions are considered:

- *Recon:* The validation set is first grouped by instrument, and FAD is measured between embeddings extracted from reconstructed mel spectrograms and those from the original inputs $\mathbf{x}_{1:L}$ within each group. This serves as a direct measure of reconstruction fidelity.

- *s-swap:* For each randomly paired sample (i, j), a reconstruction $\mathbf{x}_{1:L}^{\mathbf{s}^i \to \mathbf{s}^j}$ is generated using the global latent from sample j and the local latent from sample i. Embeddings derived from these reconstructions are grouped according to the instrument identity of sample j, and FAD is computed between these and the original recordings from the same group. This measures the model's ability to translate timbral identity.

- *z-swap:* In this setting, $\mathbf{z}_{1:L}^i$ is replaced by $\mathbf{z}_{1:L}^j$ while retaining $\mathbf{s}^i$. FAD is then calculated between the resulting samples $\mathbf{x}_{1:L}^{\mathbf{z}_{1:L}^i \to \mathbf{z}_{1:L}^j}$ and original recordings corresponding to the instrument used in sample i, thus evaluating the preservation of timbre.

- *Rand:* A sample $\tilde{\mathbf{x}}_{1:L}$ is generated by sampling a random global latent $\tilde{\mathbf{s}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{D_s})$ and combining it with the inferred local latent $\mathbf{z}_{1:L}$ via $D^{\text{Stu}}(\tilde{\mathbf{s}}, \mathbf{z}_{1:L})$. The pre-trained instrument classifier is used to predict the instrument label $\tilde{y}$ of the generated sample. FAD is then computed between $\tilde{\mathbf{x}}_{1:L}$ and original recordings annotated with instrument label $\tilde{y}$. This evaluates how closely the generated timbre aligns with plausible examples of the predicted class.

This evaluation protocol provides a comprehensive view of the model's ability to control, preserve, and manipulate independent factors of variation, while maintaining high perceptual quality across both reconstruction and generation.

	Global (F1) ↑		Local (RPA) ↑		$\text{FAD}_{\text{Violin}}$ ↓				$\text{FAD}_{\text{Trumpet}}$ ↓			
	s-swap	z-swap	s-swap	z-swap	Recon	s-swap	z-swap	Rand	Recon	s-swap	z-swap	Rand
TS-DSAE	1.00	1.00	0.87	0.86	3.49	4.26 (0.15)	4.53 (0.09)	4.85 (0.09)	2.08	4.00 (0.48)	3.46 (0.36)	2.88 (0.02)
J-DSAE	1.00	1.00	**0.93**	**0.92**	**2.17**	**3.22** (0.06)	**3.37** (0.07)	**3.37** (0.02)	**1.20**	**2.73** (0.49)	**2.25** (0.27)	**2.12** (0.14)
TS-DSAE (AR)	1.00	1.00	0.82	0.83	2.18	3.19 (0.11)	3.47 (0.07)	3.56 (0.03)	1.41	3.61 (0.48)	2.92 (0.38)	2.59 (0.24)
J-DSAE (AR)	1.00	1.00	**0.90**	**0.90**	**1.49**	**2.35** (0.05)	**2.57** (0.03)	**2.56** (0.07)	**0.89**	**1.05** (0.12)	**0.99** (0.04)	**1.88** (0.09)

Table 4.1: Disentanglement in terms of macro F1 score and RPA, alongside audio quality measured in FAD. *Recon* column reports the FAD measured between embeddings derived from the reconstructed and original data, while *Rand* compares embeddings extracted from the real recordings and those from generated data conditioned on randomly sampled global latents. See Section 4.5 for the complete definitions of the columns. The numbers in parentheses are standard deviation of three random paired sequences (for s- and z-swap) or samples of $\tilde{\mathbf{s}}$ (for Rand).

4.6 Results

4.6.1 Quantitative Metrics

The results of the quantitative evaluation are summarised in Table 4.1. Across all metrics for disentanglement and reconstruction, J-DSAE consistently outperforms TS-DSAE.

In particular, improvements in RPA indicate stronger disentanglement, with the local latent $\mathbf{z}_{1:L}$ capturing pitch-related information more effectively. Simultaneously, J-DSAE achieves lower FAD in all evaluation settings described in Section 4.5, reflecting enhanced sample quality in timbre translation (via s-swap), preservation (via z-swap), and unconditional synthesis (via Rand).

Remarkably, even when using a factorised decoder, J-DSAE surpasses TS-DSAE that employs a more expressive autoregressive decoder. This outcome suggests that the Jacobian constraint provides a more effective inductive bias than the latent-space auxiliary losses (3.5)–(3.6) used in TS-DSAE.

It is hypothesised that the auxiliary objectives in TS-DSAE can be satisfied through latent representations that exploit superficial correlations, encoding information that does not correspond to natural variation in the observation space. By contrast, the Jacobian constraint operates in the output domain and explicitly encourages changes in the generated signal to align with variations in the global latent. This is made possible by grounding the student model on the behaviour learnt by the teacher network $(E_s^{\text{Tea}}, D^{\text{Tea}})$.

4.6.2 Sampling from Timbre Space

To further investigate the structure of the learnt timbre space, a reduced global latent dimensionality ($D_s = 2$) is adopted, while maintaining the local latent dimensionality at $D_z = 32$. The left panel of Fig. 4.2 visualises the outputs of the teacher decoder D^{Tea} for grid-sampled values of $\mathbf{s}$, i.e., $\mathbf{s}^{\text{grid}} = (m, n)$ where $m, n \in \{-3, 0, 3\}$.

The generated mel spectrograms demonstrate smooth and coherent variation across the grid, while exhibiting static spectral patterns over time. This illustrates that $\mathbf{s}$ effectively captures global, timbral attributes that are disentangled from temporal dynamics. Perceptible transitions in spectral shape, such as shifts in brightness or transitions between instrument classes (e.g., trumpet to violin), confirm the interpretability of the learnt space.

Let $\mathbf{s}^{\text{diag},\,m} = (m, -m)$ denote global latents along the diagonal, with $m \in \{-3, 0, 3\}$, and let $\mathbf{x}_{1:L}^{\text{seed},\,i}$ represent seed examples from the first column of the right panel for $i \in \{1, 2\}$. For each seed, local latents are obtained via $\mathbf{z}_{1:L}^{\text{seed},\,i} \sim E_z^{\text{Stu}}(\mathbf{x}_{1:L}^{\text{seed},\,i})$. By combining these with different $\mathbf{s}^{\text{diag},\,m}$ and decoding through D^{Stu}, new samples $\mathbf{x}_{1:L}^{\text{new},\,(m,i)}$ are generated. These share pitch content with the seed but vary in timbre, validating the modular and disentangled nature of the representation.

4.6.3 Attribute Translation

Attribute translation results are shown in Fig. 4.3. In the z-swap condition (blue block), the pitch information is altered while timbre remains unchanged, producing

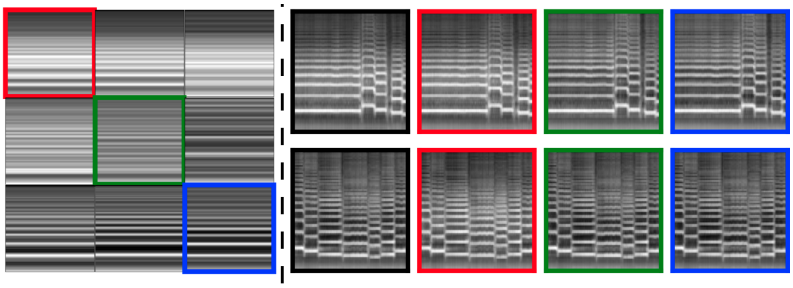

Figure 4.2: Visualisation of the learnt timbre space. Left: outputs of D^{Tea} from a 2D grid in global latent space. Right: outputs of D^{Stu} generated by combining fixed local latents (pitch) from the leftmost column with varying global latents (timbre) from the diagonal of the left panel, with correspondences indicated by the colour of the frame. See Section 4.6.2 for details.

outputs that preserve the original instrument identity but adopt the pitch contour of the target. In the s-swap condition (red block), the global latent is replaced, resulting in samples that translate the timbre while maintaining the original pitch. These outcomes verify the independent controllability of the latent subspaces and their alignment with semantically meaningful attributes. [2]

4.7 Summary

Disentangled representation learning with autoencoder-based models frequently relies on strong regularisation to guide factor separation, often at the expense of reconstruction fidelity. This chapter introduced J-DSAE, which integrates a Jacobian-based constraint into the TS-DSAE framework to address this trade-off.

Unlike conventional approaches that constrain the latent space directly, J-DSAE employs a teacher-student framework to regularise the observation space, encouraging semantically meaningful variation in the generated output. This leads to improved

[2] Audio samples are available at `https://www.jaco-dsae.xyz/`.

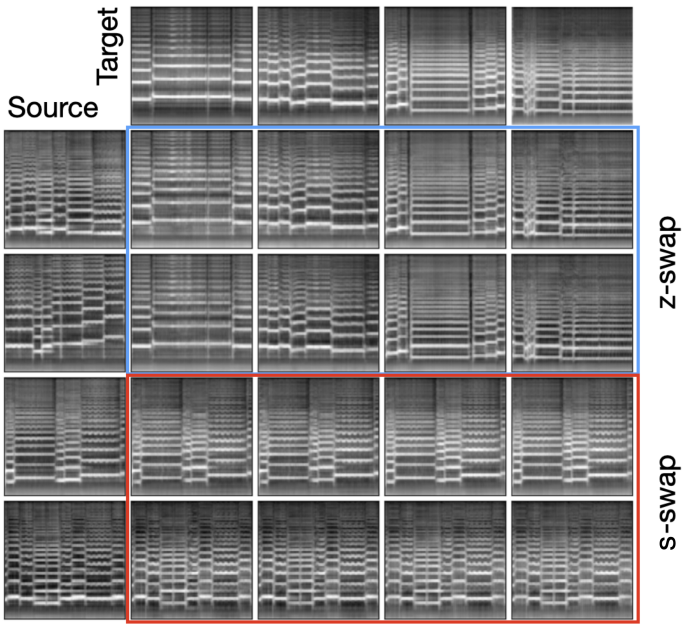

Figure 4.3: Attribute translation via latent swapping. Blue: z-swap alters pitch while preserving timbre. Red: s-swap changes timbre while maintaining pitch. See Section 4.6.3.

disentanglement of pitch and timbre without compromising synthesis quality, as evidenced by improvements in both RPA and FAD.

The findings suggest that the observation-level Jacobian constraint can better guide disentanglement by ensuring that changes in global latent variables correspond to meaningful variation in the generated signal. While the current method depends on the successful training of a teacher network to capture global attributes, further research is needed to understand the conditions under which teacher effectiveness is maximised and generalised.

Chapter 5

Balancing the Information Bottleneck: Adaptive Gaussian Dropout

Disentangled sequential autoencoders (DSAEs) offer a principled framework for separating global and local generative factors in sequential data, as reviewed in Section 2.2. Section 2.4 elaborates on the framework's application to speech, where this factorisation aligns naturally with the time-invariant nature of speaker identity and the dynamic nature of linguistic content. Such a structure has shown promise in zero-shot voice conversion (VC), where the goal is to modify the speaker characteristics of an utterance while preserving its linguistic content, even for speakers unseen during training.

A recurring challenge in training DSAEs is the collapse of the global latent representation. Owing to the lower temporal resolution and the regularising effect of the prior, the global representation is often overshadowed by the frame-level local latents. Chapter 3 and Chapter 4 propose countermeasures to this issue and validate their effectiveness using monophonic music audio. This chapter focuses instead on speech disentanglement, where the collapse of the global representation hampers the model's ability to isolate speaker identity from local linguistic content. Although evaluated in

the context of zero-shot VC, the proposed method is modality-agnostic and orthogonal to the approaches proposed in the previous two chapters on music. Exploring how these strategies may complement one another across both the music and speech domains is left as an avenue for future work.

Specifically, a simple yet effective regularisation technique—posterior variance-parameterised Gaussian dropout (pvpGD)—is introduced to address the challenge. By applying a multiplicative noise channel to the local latent space, with the noise magnitude scaled according to the posterior variance of the global latent, the method dynamically adjusts the information capacity of the local representation in accordance with global latent usage. Intuitively, when the model bypasses the global latent for data reconstruction, the posterior variance of the global latent increases to match closely to the standard Gaussian prior. This in turn amplifies the noise applied to the local latents, thereby strengthening the information bottleneck and hindering reconstruction quality. In other words, the model is penalised for communicating all information solely through the local latents, which mitigates the collapse of the global latent. The approach is fully unsupervised, does not require labelled speaker data or architectural modifications, and integrates seamlessly into existing DSAE-based VC models. Empirical evaluations on zero-shot VC demonstrate substantial improvements in speaker similarity while preserving or enhancing intelligibility, thereby validating the efficacy of pvpGD in rebalancing latent utilisation within the DSAE framework.[1]

5.1 Introduction

Building on Section 2.2, this chapter focuses on the collapse of the global latent variable in the context of speech disentanglement and zero-shot VC.

DSAEs provide a probabilistic formulation of this factorisation, but in practice the high temporal resolution and expressive capacity of local latents often dominate the training dynamics, leading to underutilisation or collapse of the global latent space (Lu et al., 2022; Lian et al., 2022b).

[1] This chapter is based on work published in Luo et al. (2024b).

Several studies have attempted to resolve this issue by introducing supervised speaker encoders (Jia et al., 2018; Variani et al., 2014), contrastive learning objectives (Yuan et al., 2021; Luong et al., 2021), or architectural bottlenecks such as dimensionality reduction (Qian et al., 2019), instance normalisation (Chou et al., 2019), and vector quantisation (Tjandra et al., 2021). While these approaches offer partial solutions, they often depend on labelled data, introduce additional training complexity, or impose rigid constraints on model design. Moreover, many of them remove the prior distribution over the global latent variable, resulting in deterministic speaker representations. Although this design choice mitigates posterior collapse by not introducing noise and uncertainty to speaker representations, it limits the model's flexibility for tasks such as unconditionally sampling novel speakers and smoothly interpolating between speaker identities (Hsu et al., 2019; Stanton et al., 2021).

This chapter introduces pvpGD, a stochastic regularisation method that adaptively injects multiplicative noise into the local latent representation. The noise level is determined by the posterior variance of the global latent, such that higher uncertainty in the global variable induces stronger noise in the local latent, discouraging it from absorbing global information. Conversely, when the global posterior is confident, less noise is applied, preserving the fidelity of local representations. In contrast to conventional Gaussian dropout, which requires manual tuning of the dropout strength (S. Wang et al., 2013; Rey et al., 2021; Srivastava et al., 2014), pvpGD derives its dropout strength from the posterior variance of the global latent variable. Even though one might argue for the potential benefit of introducing an additional scaling factor on the posterior variance to more explicitly control the noise level, a similar effect can already be achieved through the existing weight β_s on the Kullback–Leibler divergence (KLD) term associated with the global latent space in (2.4). Therefore, pvpGD is kept simple and free from additional hyperparameters. The method remains applicable as long as the posterior over the global latent is Gaussian.

Applied to existing zero-shot VC models, pvpGD enhances the separation of speaker and linguistic representations without requiring supervision or architectural alterations, while preserving an explicit prior distribution for the global latent space. Evaluations demonstrate consistent gains in speaker similarity and competitive char-

acter error rates across a range of DSAE variants and regularisation regimes. These results highlight pvpGD as a robust and general-purpose method for mitigating global latent collapse in speech disentanglement tasks.

5.2 Related Work

5.2.1 Speaker-Content Disentanglement

Section 2.2 elaborates on the DSAE (Y. Li et al., 2018) and its optimisation challenges, notably the tendency for posterior collapse. Moreover, contemporary models for zero-shot voice conversion (VC) often adopt the architectural inductive biases embedded in the DSAE framework, as reviewed in Section 2.4.

To briefly summarise, the DSAE describes the generative process of a sequence $\mathbf{x}_{1:L}$ using the factorisation:

$$\prod_{t=1}^{L} p_\theta(\mathbf{x}_t|\mathbf{s},\mathbf{z}_t)\, p_\theta(\mathbf{z}_t|\mathbf{z}_{<t})\, p_\theta(\mathbf{s}),$$

where the global latent variable $\mathbf{s}$ captures sequence-level attributes such as speaker identity, and the frame-level latent variables $\mathbf{z}_{1:L}$ account for dynamic content, such as phonetics. Following the variational autoencoder (VAE) framework (Kingma et al., 2014; Rezende et al., 2014), the model introduces approximate posteriors $q_\phi(\mathbf{z}_{1:L}|\mathbf{x}_{1:L})$ and $q_\phi(\mathbf{s}|\mathbf{x}_{1:L})$ to optimise a modified evidence lower bound (ELBO), incorporating weighting coefficients β_s and β_z to modulate the information bottlenecks, in a manner analogous to the β-VAE (Higgins et al., 2017a). Concretely, the objective function, identical to (2.4) but restated here for convenience, is written as:

$$\begin{aligned}
\mathcal{L}_{\text{DSAE}} &= \mathbb{E}_{q_\phi(\mathbf{z}_{1:L},\mathbf{s}|\mathbf{x}_{1:L})}\big[\log p_\theta(\mathbf{x}_{1:L},\mathbf{z}_{1:L},\mathbf{s}) - \log q_\phi(\mathbf{z}_{1:L},\mathbf{s}|\mathbf{x}_{1:L})\big] \\
&= \sum_{t=1}^{L} \mathbb{E}_{q_\phi(\mathbf{z}_t|\mathbf{x}_{1:L})q_\phi(\mathbf{s}|\mathbf{x}_{1:L})}\big[\log p_\theta(\mathbf{x}_t|\mathbf{z}_t,\mathbf{s})\big] \\
&\quad - \beta_z \sum_{t=1}^{L} \mathbb{E}_{q_\phi(\mathbf{z}_{<t}|\mathbf{x}_{1:L})}\big[\mathcal{D}_{\text{KL}}\big(q_\phi(\mathbf{z}_t|\mathbf{x}_{1:L})\|p_\theta(\mathbf{z}_t|\mathbf{z}_{<t})\big)\big] \\
&\quad - \beta_s \mathcal{D}_{\text{KL}}\big(q_\phi(\mathbf{s}|\mathbf{x}_{1:L})\|p_\theta(\mathbf{s})\big).
\end{aligned} \tag{5.1}$$

In zero-shot VC, $q_\phi(\mathbf{s}|\mathbf{x}_{1:L})$ and $q_\phi(\mathbf{z}_{1:L}|\mathbf{x}_{1:L})$ serve to encode speaker identity and linguistic content, respectively. Conversion is performed by retaining $\mathbf{z}_{1:L}$ from a source utterance while substituting $\mathbf{s}$ from a target utterance. However, a critical challenge lies in the imbalance of information capacity: the static $\mathbf{s}$ is significantly more constrained than the sequence $\mathbf{z}_{1:L}$, leading to speaker information leakage into $\mathbf{z}_{1:L}$ and degraded conversion quality. Reducing β_s can alleviate this leakage by allowing $\mathbf{s}$ to encode richer speaker characteristics, yet may result in an under-regularised latent space. Conversely, increasing β_s tightens the posterior distribution, potentially causing the KLD term in (2.4) to vanish, i.e.,

$$\mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{s}|\mathbf{x}_{1:L})\|p_\theta(\mathbf{s})\big) \to 0,$$

thereby rendering $\mathbf{s}$ uninformative.

5.2.2 Counteracting the Collapse of Speaker Representations

Numerous models based on the DSAE framework introduce mechanisms to preserve speaker identity within the global latent variable, as discussed in Section 2.4.

AutoVC (Qian et al., 2019) bypasses the KL terms in the ELBO, instead limiting the representational capacity of $\mathbf{z}_{1:L}$ by restricting its spatial and temporal dimensions. The speaker embedding $\mathbf{s}$ is obtained from a speaker classifier trained independently. Other methods incorporate speaker supervision using auxiliary or adversarial losses (Nguyen et al., 2022; Luong et al., 2021).

AdaIN-VC (Chou et al., 2019) applies instance normalisation (X. Huang et al., 2017) to $\mathbf{z}_{1:L}$ to remove global statistics and sets $\beta_s = 0$ to eliminate the bottleneck for $\mathbf{s}$. Vector quantisation (VQ) (Oord et al., 2017) has also been used to discretise $\mathbf{z}_{1:L}$, thereby reducing its capacity and discouraging the encoding of speaker-specific information (D.-Y. Wu et al., 2020; Tjandra et al., 2021).

Despite differences in supervision and architectural design, a unifying strategy across these models is to constrain the expressiveness of $\mathbf{z}_{1:L}$ in order to promote reliance on $\mathbf{s}$. In contrast, Lian et al. (2022b) and Lu et al. (2022) preserve the original DSAE structure without external constraints, achieving fully unsupervised

disentanglement. Retaining the prior $p(\mathbf{s})$ additionally allows for unconditional sampling of novel speaker identities and smooth interpolation across speakers (Stanton et al., 2021; Hsu et al., 2019). Nevertheless, effective disentanglement in such settings hinges on careful tuning of β_s and β_z (Lian et al., 2022b; Lu et al., 2022).

5.2.3 Gaussian Dropout as an Information Bottleneck

An ideal representation should be sufficiently expressive for downstream tasks while remaining compressive with respect to the input (Bengio et al., 2013a; Tishby et al., 1999). One means of enforcing this constraint is through the injection of noise, such as Gaussian perturbations, which act as an information bottleneck (Alemi et al., 2017; Rey et al., 2021).

For example, information bottleneck layers can be applied to systems that combine an autoregressive (AR) model and a VAE to regularise the information pathway. The main idea in such systems is to use the AR model to capture local details and nuisance factors, while reserving the VAE for global factors that are more semantically meaningful. However, AR models are often capable of modelling the data distribution alone without relying on information encoded by the VAE, similar to posterior collapse discussed in Chapter 2.2.2. To counteract this issue, X. Chen et al. (2017) and Gulrajani et al. (2017) propose architectural restrictions on AR models, whereas Rey et al. (2021) introduce Gaussian dropout (GD) as channel-wise multiplicative noise to limit the information capacity of the AR model, thereby encouraging the system to make use of the VAE outputs.

More specifically, GD is applied to a representation $\mathbf{r}$ as:

$$\mathbf{r}' = \mathbf{r} \times \epsilon, \quad \text{where } \epsilon \sim \mathcal{N}\big(\mathbf{1}, \sigma^2\mathbf{I}\big). \tag{5.2}$$

The noise level is controlled by a hyperparameter σ. When compared against additive Gaussian noise (Alemi et al., 2017), Rey et al. (2021) show that multiplicative GD achieves a better trade-off between data reconstruction quality and expressivity of the learnt representation. This chapter investigates the use of this mechanism to regularise the local latents $\mathbf{z}_{1:L}$ of DSAEs, thereby limiting their capacity and incentivising the

model to encode global factors, such as speaker identity, in the global latent $\mathbf{s}$.

5.3 Method

This section introduces a novel parameterisation of the variance used in multiplicative Gaussian noise, which acts as a form of dropout or information bottleneck without requiring additional hyperparameters.

5.3.1 Objective Function

To focus on verifying the effectiveness of the proposed method in mitigating posterior collapse of the global latent, both the prior and approximate posterior distributions are configured in factorised forms. Specifically, the prior distributions are simplified as $p_\theta(\mathbf{z}_{1:T}) = \prod_{t=1}^{T} p(\mathbf{z}_t)$ and $p_\theta(\mathbf{s}) = p(\mathbf{s})$, where both $p(\mathbf{z}_t)$ and $p(\mathbf{s})$ are set to $\mathcal{N}\big(\mathbf{0}, \mathbf{I}\big)$. The approximate posterior distribution $q_\phi(\mathbf{z}_{1:L}, \mathbf{s}|\mathbf{x}_{1:L})$ is factorised as $q_\phi(\mathbf{s}|\mathbf{x}_{1:L}) \prod_{t=1}^{T} q_\phi(\mathbf{z}_t|\mathbf{x}_{1:L})$. Under these assumptions, the objective function reduces to:

$$
\begin{aligned}
\mathcal{L}'_{\mathrm{DSAE}} &= \mathbb{E}_{q_\phi(\mathbf{z}_{1:L}, \mathbf{s}|\mathbf{x}_{1:L})}\big[\log p_\theta(\mathbf{x}_{1:L}, \mathbf{z}_{1:L}, \mathbf{s}) - \log q_\phi(\mathbf{z}_{1:L}, \mathbf{s}|\mathbf{x}_{1:L})\big] \\
&= \sum_{t=1}^{L} \mathbb{E}_{q_\phi(\mathbf{z}_t|\mathbf{x}_{1:L}) q_\phi(\mathbf{s}|\mathbf{x}_{1:L})}\big[\log p_\theta(\mathbf{x}_t|\mathbf{z}_t, \mathbf{s})\big] \\
&\quad - \beta_z \sum_{t=1}^{L} \mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{z}_t|\mathbf{x}_{1:L}) \| p(\mathbf{z}_t)\big) \\
&\quad - \beta_s \mathcal{D}_{\mathrm{KL}}\big(q_\phi(\mathbf{s}|\mathbf{x}_{1:L}) \| p(\mathbf{s})\big).
\end{aligned}
\tag{5.3}
$$

The main idea of the proposed method is to apply multiplicative Gaussian noise to the local latent $\mathbf{z}_t \sim q_\phi(\mathbf{z}_t|\mathbf{x}_{1:L})$, and to use the noise-perturbed local latent together with the global latent $\mathbf{s}$ to reconstruct the data $\mathbf{x}_t$. The additional noise introduces an extra information bottleneck on the local latent, which limits its ability to encode all information and helps to prevent the global latent from collapsing. The next section specifies the parameterisation of this noise.

5.3.2 Noise Parameterisation

Let the standard deviation of the diagonal Gaussian posterior $q_\phi(\mathbf{s}|\mathbf{x}_{1:L})$ be denoted as $\sigma_s \coloneqq \sigma_\phi(\mathbf{s}) = (\sigma_s^{(1)}, \ldots, \sigma_s^{(D_s)}) \in \mathbb{R}^{D_s}$, where D_s is the dimensionality of the global latent space. The proposed parameterisation is defined as:

$$\log \sigma_{\text{pvpGD}} \coloneqq \frac{1}{D_s} \sum_{d=1}^{D_s} \log \sigma_s^{(d)} = \log \left[\prod_{d=1}^{D_s} \sigma_s^{(d)} \right]^{\frac{1}{D_s}}. \tag{5.4}$$

That is, the standard deviation used in pvpGD is given by the geometric mean of the standard deviations in σ_s.

The pvpGD noise can be applied to any latent variable; in the present setting, it is applied to the local latent $\mathbf{z}_t \sim q_\phi(\mathbf{z}_t|\mathbf{x}_{1:L})$, resulting in a perturbed version:

$$\mathbf{z}'_t = \mathbf{z}_t \times \epsilon, \quad \text{where } \epsilon \sim \mathcal{N}\big(\mathbf{1}, \sigma^2_{\text{pvpGD}}\mathbf{I}\big). \tag{5.5}$$

Unlike conventional GD (Rey et al., 2021; S. Wang et al., 2013), pvpGD introduces no additional hyperparameters, such as a fixed dropout probability. As shown in (5.4), σ^2_{pvpGD} can be computed simply by exponentiating the arithmetic mean of $\log \sigma_s^2$, a quantity commonly produced by a fully connected layer in standard VAE implementations.

5.3.3 Interpretation

The proposed pvpGD mitigates the posterior collapse of the global latent variable $\mathbf{s}$ by coupling the bottleneck applied to the local latent $\mathbf{z}_t$ with the posterior variance of $\mathbf{s}$. Over-regularisation leads to $\mathcal{D}_{\text{KL}}\big(q_\phi(\mathbf{s}|\mathbf{x}_{1:L})\|p(\mathbf{s})\big) \to 0$, resulting in the collapse of $\mathbf{s}$. According to (5.4) and (5.5), this increases the value of σ_{pvpGD}, thereby imposing a stronger bottleneck on $\mathbf{z}_t$. Since a collapsed $\mathbf{s}$ also impairs the capacity of $\mathbf{z}_t$, the model is incentivised to limit the expansion of σ_s, thus mitigating collapse.

On the other hand, making σ_{pvpGD} arbitrarily small to increase the capacity of $\mathbf{z}_t$ effectively removes the stochasticity of $q_\phi(\mathbf{s}|\mathbf{x}_{1:L})$, which is penalised by the KL divergence term $\mathcal{D}_{\text{KL}}\big(q_\phi(\mathbf{s}|\mathbf{x}_{1:L})\|p(\mathbf{s})\big)$ in the ELBO, given the standard normal prior $p(\mathbf{s})$. In other words, the proposed pvpGD counteracts the collapse of $\mathbf{s}$ by introducing

a bottleneck to $\mathbf{z}_t$, with its strength determined by the capacity of $\mathbf{s}$.

By coupling the capacities of the two latent variables, the model is encouraged to operate in a regime where neither latent dominates the reconstruction process. In this context, a reasonable equilibrium refers to a training regime in which $\mathbf{s}$ remains sufficiently informative to avoid posterior collapse, while $\mathbf{z}_t$ retains enough effective capacity to model time-varying details without becoming an unrestricted information channel. As a result, $\mathbf{s}$ is regularised, yet retains a suitable amount of information, which is empirically verified by Fig. 5.3.

5.4 Experiments

The proposed pvpGD is applied to AdaIN-VC (Chou et al., 2019) and DSAE (Lian et al., 2022b; Lu et al., 2022), as well as to their speaker-supervised and VQ counterparts. In the context of zero-shot VC, empirical results are presented to demonstrate that pvpGD effectively mitigates the collapse of speaker representations across a wide range of bottleneck strength. Specifically, the experiment sets $\beta_s = \{0, 0.5, 1, 4, 32\}$ and $\beta_z = 1$ in (5.3).

5.4.1 Datasets

The experiments are conducted using the VCTK dataset (Veaux et al., 2013), with 99 English speakers allocated to the training set and 10 to the test set. A validation set is created by withholding 10% of the training data. The test set comprises an equal number of male and female speakers. In addition to reserving specific speakers for testing, a total of 47 out of 500 unique utterances are randomly excluded from the training set.

In total, the training and test sets contain 32,822 and 4,011 utterances, respectively. All test utterances are from speakers unseen during training, with 378 of them also containing spoken content not observed in the training set. Furthermore, the 25 utterances used for intra-lingual conversion in VCC2020 (Zhao et al., 2020) are included for evaluation purposes only.

All recordings are downsampled to 22,050 Hz. Logarithmic mel spectrograms with

80 frequency bands are extracted from a short-time Fourier transform using a window size of 1024 and a hop size of 256.

5.4.2 Evaluation

A total of 4,011 random source–target pairs are constructed from the VCTK test set, where all speakers are unseen. Among these, 378 pairs involve conversions between utterances with unseen spoken content. The VCC2020 evaluation set provides an additional 400 source–target pairs for inter-dataset evaluation.

In zero-shot VC, the speaker acceptance rate (SAR) is reported to assess the success of speaker identity transfer, and the character error rate (CER) is used to evaluate the preservation of linguistic content (W.-C. Huang et al., 2022; T.-H. Huang et al., 2021). SAR is computed as the proportion of converted utterances that are accepted by a speaker verification (SV) system. Separate SV systems are built for the VCTK and VCC2020 test sets, using speaker embeddings obtained from a publicly available pre-trained speaker encoder.[2] CER is measured using a pre-trained wav2vec 2.0 model.[3]

5.4.3 Implementation of Common Modules

The global encoder $q_\phi(\mathbf{s}|\mathbf{x}_{1:L})$ follows the speaker encoder design from AdaIN-VC (Chou et al., 2019). The input mel spectrogram is first passed through a convolution bank that concatenates the outputs of eight one-dimensional convolution (Conv) modules. The result is then processed by six blocks of Conv and subsampling layers, followed by temporal average pooling to obtain an utterance-level embedding. This embedding is projected onto a latent space of size $D_s = 256$ and subsequently passed through two fully connected (FC) layers that form the Gaussian reparameterisation layer, yielding the mean μ_s and standard deviation σ_s for parameterising $q_\phi(\mathbf{s}|\mathbf{x}_{1:L})$.

The local encoder $q_\phi(\mathbf{z}_t|\mathbf{x}_{1:L})$ is adapted from a VQ model for acoustic unit discovery (Niekerk et al., 2020a). It begins with a Conv layer with stride two, halving the temporal resolution. This is followed by five blocks comprising layer normalisation,

[2] `https://github.com/resemble-ai/Resemblyzer`
[3] `https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self`

ReLU activation, and an FC layer. The final FC layer maps the intermediate embedding $\mathbf{e}_t$ to a local latent space of size $D_z = 64$, which is then passed to a Gaussian reparameterisation layer to produce μ_z and σ_z at each time step.

The decoder $p_\theta(\mathbf{x}_t|\mathbf{z}_t, \mathbf{s})$ follows the architecture of AutoVC (Qian et al., 2019). It first upsamples $\mathbf{z}_t$ (or its perturbed variant $\mathbf{z}'_t$) by a factor of two via linear interpolation and repeats the global latent $\mathbf{s}$ across T time steps. A 512-dimensional LSTM consumes the concatenated $\mathbf{z}_t$ and $\mathbf{s}$ as input, followed by three blocks of Conv and batch normalisation (BN). This is followed by two 1024-dimensional LSTM layers and an FC layer that maps the embedding back to the mel spectrogram size $D_x = 80$. A Gaussian reparameterisation layer then produces the output mean μ_x, with σ_x held fixed in practice.

A PostNet, composed of five Conv and BN blocks, refines the reconstruction μ_x by predicting its residual from the input mel spectrogram. The final audio waveform is synthesised using the pre-trained MelGAN vocoder.[4]

5.4.4 Implementation of individual models

AdaIN-VC

AdaIN-VC extracts the global latent $\mathbf{s}$ using the speaker encoder described in Section 5.4.3. The content encoder includes an instance normalisation (IN) layer before the Gaussian reparameterisation layer. The IN layer normalises the content branch using channel-wise statistics, which is intuitively expected to remove speaker information (Chou et al., 2019). Formally, given an intermediate representation $\mathbf{h}$:

$$\mathrm{IN}(\mathbf{h}) = \frac{\mathbf{h} - \mu(\mathbf{h})}{\sigma(\mathbf{h})}, \tag{5.6}$$

where $\mu(\mathbf{h})$ and $\sigma(\mathbf{h})$ denote the channel-wise mean and standard deviation. The local latent at each time step is then sampled as follows:

$$\mathbf{z}_t \sim \mathcal{N}\big(\mu_{\phi_z}(\mathrm{IN}(\mathbf{h}_c)), \sigma^2_{\phi_z}(\mathrm{IN}(\mathbf{h}_c))\mathbf{I}\big), \tag{5.7}$$

[4] `https://github.com/descriptinc/melgan-neurips/tree/master`

where $\mathbf{h}_c$ represents the hidden states of the local encoder that precede the IN and Gaussian parameterisation layer. The application of pvpGD follows (5.5) to regularise the local latents.

Rather than concatenating $\mathbf{z}$ and $\mathbf{s}$ as input to the decoder, AdaIN-VC retains speaker information by replacing the BN layers in the three convolutional blocks with adaptive IN layers. Each adaptive IN layer first applies IN and then performs an affine transformation with a learnable mean and standard deviation conditioned on the global latent $\mathbf{s}$ (Chou et al., 2019). Formally, given intermediate representations of the decoder $\mathbf{h}_d$:

$$\text{AdaIN}(\mathbf{h}_d, \mathbf{s}) = \gamma(\mathbf{s}) \cdot \text{IN}(\mathbf{h}_d) + \beta(\mathbf{s}), \tag{5.8}$$

where $\gamma(\mathbf{s})$ and $\beta(\mathbf{s})$ are scale and shift parameters predicted from the global latent $\mathbf{s}$ through simple linear projections. The input $\mathbf{h}_d$ to the first adaptive IN layer of the decoder is either the local latents $\mathbf{z}_{1:T}$ or their noised counterpart $\mathbf{z}'_{1:T}$ if pvpGD is applied. This design explicitly conditions the affine transformation on $\mathbf{s}$, enabling speaker information to modulate the normalised content representation. When optimised by (5.3), positive β_s and β_z provide a principled probabilistic extension of the original formulation (Chou et al., 2019), and the original model is recovered when both β_s and β_z are set to zero.

DSAE

The DSAE model is optimised using (5.3) and incorporates the speaker encoder, content encoder, and decoder described in Section 5.4.3. When pvpGD is applied, it follows the formulation given in (5.4).

DSAE-S

A speaker-supervised variant, referred to as DSAE-S, is also implemented for comparison. The speaker encoder is replaced with the same pre-trained model used for SV in Section 5.4.2, following the design of AutoVC (Qian et al., 2019). Specifically, an alternative version is considered in which the normalisation on the output of the pre-trained encoder is removed. A Gaussian reparameterisation layer is appended to

allow fine-tuning of the speaker embedding. This design maintains the probabilistic structure of the model and allows optimisation via (5.3), which also permits the use of pvpGD as a regularisation method.

DSAE-VQ

DSAE-VQ is a VQ variant in which the Gaussian reparameterisation layer is removed from the local encoder. Instead, a discrete latent $\mathbf{z}_t$ is learnt from the continuous intermediate representation $\mathbf{e}_t$ using a codebook of size 512.

Concretely, let $\mathcal{E} = \{\mathbf{e}^{(k)}\}_{k=1}^{512}$ denote the trainable codebook. At each time step,

$$k_t = \underset{1 \le k \le 512}{\operatorname{argmin}} \left\| \mathbf{e}_t - \mathbf{e}^{(k)} \right\|_2^2, \qquad \mathbf{z}_t = \mathbf{e}^{(k_t)}, \tag{5.9}$$

and the straight-through estimator (Bengio et al., 2013b) is used to backpropagate through the non-differentiable nearest-neighbour assignment. In this case, the second KL term in (5.3) is replaced with a VQ loss: $\alpha \sum_{t=1}^{T} \|\mathbf{e}_t - \mathrm{sg}(\mathbf{z}_t)\|_2^2$, where $\alpha = 0.25$ and $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation, following existing methods (Niekerk et al., 2020a; D. Wang et al., 2021). When pvpGD is enabled, it is applied to $\mathbf{e}_t$ rather than $\mathbf{z}_t$ to avoid additional instability during training caused by the quantisation process.

5.4.5 Optimisation

The models are optimised using the Adam optimiser (Kingma et al., 2015) with a batch size of 256. Each training sample consists of a mel spectrogram segment of length $T = 128$, corresponding to approximately 1.5 seconds of audio.

The learning rate is set to 5×10^{-4}, with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and gradient clipping is applied with a threshold of three.[5] Training is stopped early if the value of $\mathcal{L}'_{\mathrm{DSAE}}$ (5.3), evaluated in the validation set, does not improve for a patience period of 13,000 steps.

[5]Settings adapted from `https://github.com/KimythAnly/AGAIN-VC` (Y.-H. Chen et al., 2021).

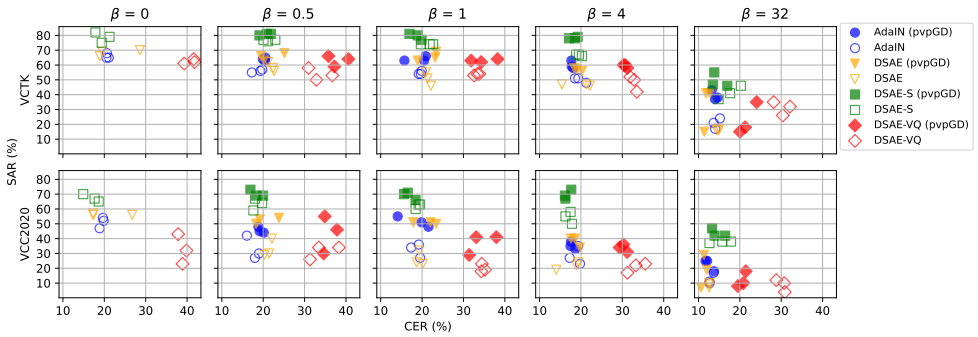

Figure 5.1: Evaluation of zero-shot VC on VCTK and VCC2020 using three random seeds per model. Speaker conversion in terms of speaker acceptance rate (SAR, higher is better) versus preservation of linguistic content in terms of character error rate (CER, lower is better) across β_s (denoted as β for simplicity in the figure).

5.5 Results

5.5.1 Applying pvpGD to Different Models and Values of β_s

The main results are presented in Fig. 5.1, where each model is evaluated using three random seeds. A model with an unconstrained speaker embedding ($\beta_s = 0$) is also included for reference. In this case, pvpGD is not applicable since the added Gaussian noise must be parameterised by the variance of a stochastic speaker latent variable.

The results are not intended as a rigorous comparison across existing models or colour groups, given the lack of extensive hyperparameter tuning per model. In particular, models such as DSAE-VQ demand greater tuning effort due to the additional complexity introduced by vector quantisation. Instead, the aim is to demonstrate the effectiveness of the proposed pvpGD across varying values of β_s.

When $\beta_s > 0$, models augmented with pvpGD consistently outperform their vanilla counterparts in speaker conversion, as indicated by SAR. This is evidenced by the filled markers generally appearing above the corresponding unfilled markers within each shape group. Exceptions include one run of DSAE and DSAE-VQ at $\beta_s = 32$. In terms of CER, there is no systematic difference between models with and without pvpGD, though notable improvements are observed in certain cases, such as DSAE at $\beta_s = 32$, DSAE-S at $\beta_s = 1$, and DSAE-VQ at $\beta_s \in \{4, 32\}$.

Overall, the speaker-supervised model DSAE-S serves as a reference point and

tends to outperform unsupervised variants; pvpGD narrows this gap effectively. DSAE-VQ typically yields lower intelligibility, as shown by its markers being shifted rightward, due to the VQ module. Additionally, the benefits of pvpGD appear most consistent within a reasonable range of β_s, and begin to diminish at higher values (e.g., $\beta_s = 32$), aligning with prior findings (Lian et al., 2022b; Lu et al., 2022). Importantly, pvpGD improves SAR without compromising CER.

Although models with $\beta_s = 0$ often achieve competitive SAR through unconstrained speaker representations, constrained models equipped with pvpGD (e.g., AdaIN-VC at $\beta_s = 1$) may outperform in CER while maintaining similar SAR. Similar observations hold for DSAE and DSAE-VQ. Beyond improved intelligibility, models with $\beta_s > 0$ also support applications such as sampling from the prior and interpolating between speaker identities (Section 5.1), and promote disentangled representations of speaker attributes (Hsu et al., 2019; Higgins et al., 2017a).

5.5.2 Comparing pvpGD with Vanilla Gaussian Dropout

Fig. 5.2 compares pvpGD to conventional multiplicative Gaussian dropout (Rey et al., 2021), using DSAE with $\beta_s = 1$. Here, GD is applied with dropout strengths $p = \{0.1, 0.3, 0.5\}$, replacing the variance term σ^2_{pvpGD} in (5.5) with $\sigma^2_{\text{GD}} = \frac{p}{1-p}$.

The results indicate that pvpGD offers a favourable trade-off, improving SAR by approximately 10% (up to 20% on VCC2020) at the cost of a modest 5% loss in CER. In one run, pvpGD does not incur any degradation in intelligibility. Among the GD variants, $p = 0.3$ yields the most balanced trade-off. However, pvpGD achieves a comparable effect without introducing an additional hyperparameter, thereby simplifying optimisation.

5.5.3 Limiting Expansion of Posterior Variance

The mechanism of pvpGD is further validated in Fig. 5.3, using DSAE to analyse its influence on posterior variance. The left panel reports the average KLD, $\mathbb{E}_{\mathbf{x}}\big[\mathcal{D}_{\text{KL}}\big(q_\phi(\mathbf{s}|\mathbf{x})\|p(\mathbf{s})\big)\big]$, and the active unit (AU) metric (Burda et al., 2016). while the right panel shows the arithmetic mean of σ_s and σ_{pvpGD}. The AU metric quantifies the number of active dimensions in the speaker latent space of size $D_s = 256$,

with a dimension considered active if its variance across samples exceeds 0.01.

As shown in the left panel, AU trends similarly to KLD, and pvpGD maintains the activity level until $\beta_s = 32$, where the gap narrows. Nonetheless, pvpGD retains two active units compared to one in the vanilla case, effectively doubling the information capacity.

The arithmetic means of the standard deviation of the global latent σ_s from the models with or without pvpGD, shown by the yellow and blue bars, respectively, complement the left panel. Specifically, without pvpGD, σ_s is close to 1.0 across all values of β_s, collapsing to the standard Gaussian prior, while the application of pvpGD significantly reduces σ_s, preventing the collapse.

In conclusion, pvpGD delivers consistent gains when β_s is within a reasonable range. When β_s becomes excessively large, the added penalty on the KLD—due to shrinking σ_s—may outweigh improvements in reconstruction. Future work may explore strategies to better exploit the enhanced information capacity observed in Fig. 5.3, potentially leading to further gains in reconstruction quality.

5.6 Summary

This chapter has introduced an alternative approach—orthogonal to the staged training frameworks presented in Chapter 3 and Chapter 4—for enhancing the robustness of global–local disentanglement in (DSAEs). Whereas previous chapters concentrated on two-stage training strategies to improve stability with respect to the dimensionality

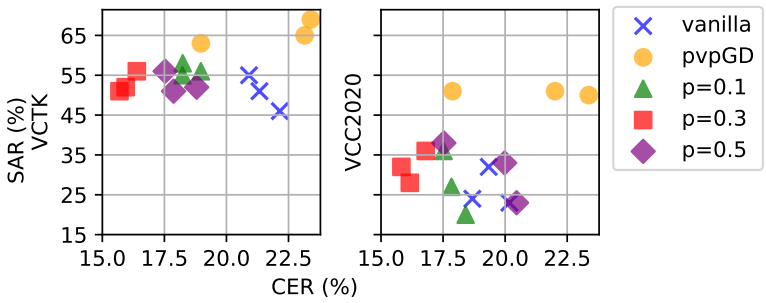

Figure 5.2: Comparing pvpGD against conventional Gaussian dropout using DSAE with $\beta_s = 1$.

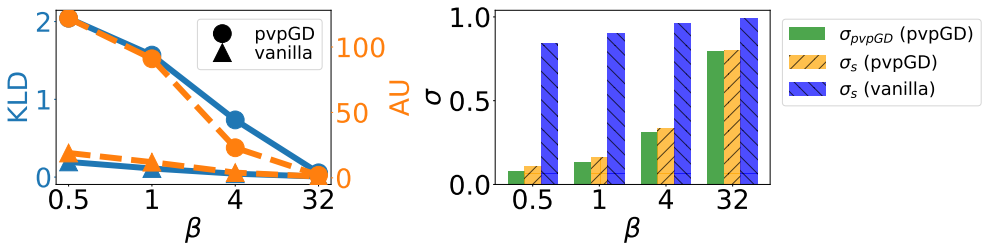

Figure 5.3: Left: KLD (solid) and AU (dashed). Right: σ_s and σ_{pvpGD}. β_s is denoted as β for simplicity in the figure.

mismatch between global and local latent variables, the present chapter has focused on mitigating sensitivity to the hyperparameter β_s, which regulates the strength of the information bottleneck in the global latent space.

The proposed method, pvpGD, is a dropout-based regularisation technique that requires no additional parameters. By parameterising the noise level according to the posterior variance, pvpGD allows the model to adaptively inject Gaussian noise conditioned on the input, thereby facilitating a more effective balance between global and local latent information. Its effectiveness has been demonstrated in the context of zero-shot VC, where it improves the disentanglement between speaker identity and linguistic content across various existing VC models and a wide range of β_s values.

Taken together with the methods proposed in Chapter 3 and Chapter 4, the approaches developed in this thesis address posterior collapse from complementary perspectives. TS-DSAE improves robustness by restructuring the optimisation process, encouraging the model to establish disentangled representations before prioritising reconstruction. J-DSAE focuses on preserving disentanglement while improving reconstruction quality by constraining the second training stage using outputs from the first. In contrast, pvpGD operates directly on the latent bottleneck by dynamically modulating the effective capacity of the local latent through noise parameterised by the posterior variance of the global latent. As such, the three methods are best interpreted as addressing different aspects of the same underlying robustness problem, rather than as interchangeable alternatives.

Chapter 6

Learning Composable Pitch-Timbre Representations: A Two-Level Disentanglement Framework

Disentangling pitch and timbre from the audio of a musical instrument involves encoding these attributes as distinct latent representations, thereby enabling the synthesis of novel sounds through independent manipulation of either attribute, as reviewed in Section 2.3. While Chapter 3 and Chapter 4 focus on improving the robustness of unsupervised pitch–timbre disentanglement (PTD) for monophonic, single-instrument audio, the present chapter extends PTD to the multi-instrument scenario, a relatively underexplored direction. As a preliminary study, this chapter investigates a fully supervised approach for its conceptual simplicity, which retains the global–local inductive bias developed in the previous chapters.

The objective of multi-instrument PTD is to extract, for each instrument in a

mixture, a source-specific representation that further encodes factorised pitch and timbre representations. These representations serve as modular building blocks that can be recombined to condition a decoder, thereby enabling the synthesis of new mixtures via arbitrary combinations of pitch and timbre both within and across source representations.

To tackle the challenge, this chapter employs a two-level disentanglement framework, which first learns to separate and encode each source in a latent space, and then disentangles pitch from timbre within each individual source representation. To validate the approach, a simple autoencoder is introduced and evaluated on structured, isolated chord mixtures. Subsequently, the method is scaled to the more complex four-part chorale dataset using a more advanced architecture that jointly learns latent variables and a diffusion transformer.

The evaluation highlights the key factors that contribute to successful PTD in the multi-instrument context, and demonstrates the feasibility of mixture transformation through source-level attribute manipulation.[1]

6.1 Introduction

Building on the pitch–timbre disentanglement framework reviewed in Section 2.3, this chapter considers the extension from single-instrument settings to mixtures containing multiple instruments. Existing approaches focus primarily on single-instrument audio and enable attribute transfer through latent manipulation. While effective, they do not address mixtures containing multiple instruments. Although MSI-DIS (Lin et al., 2021) encodes mixtures of two instruments, its decoder still generates one instrument at a time, conditioned only on the instrument's pitch and timbre. This setup limits the ability to model interactions between sources, as the decoder is not conditioned jointly on the pitch and timbre of all constituent sources.

To bridge the gap, this chapter proposes *DisMix*, a fully supervised framework for disentangling musical mixtures and composing new ones via modular manipulation of pitch and timbre representations. Each instrument in a mixture is encoded into

[1] This chapter is based on work published in Luo et al. (2024a).

a source-specific latent representation comprising distinct pitch and timbre variables. Encoding is conditioned on both the input mixture and a set of instrument-specific query clips, which specify the timbral identity of the sources to be extracted. These pitch and timbre latents are encoded in separate subspaces, thereby enabling independent manipulation of each attribute. The decoder then reconstructs a new mixture based on the manipulated latent set, supporting fine-grained editing of musical content.

These representations serve as modular building blocks for music generation and enable transformations such as replacing the flute in a piano–flute mixture with a violin while preserving its melodic line. The decoder assembles these objects in a way that preserves their pitch and timbre. This is reminiscent of "object representation" (Greff et al., 2020; Bizley et al., 2013) motivated by the human ability to understand complex ideas in terms of reusable and primitive components (Fodor et al., 1988).

Two instantiations of DisMix are introduced. The first, described in Section 6.5, offers a proof of concept based on isolated chord mixtures using a simple autoencoder architecture. The second, presented in Section 6.6, extends the framework to the more complex CocoChorale dataset (Y. Wu et al., 2022), which consists of four-part chorales rendered by thirteen orchestral instruments. The latter model admits a diffusion autoencoder framework (Preechakul et al., 2022) that jointly learns the disentangled representations and a diffusion transformer (Peebles et al., 2023).

6.2 Related Work

6.2.1 Pitch and Timbre Disentanglement

A variety of strategies have been proposed for PTD in music audio. These approaches include supervised learning frameworks (Luo et al., 2019; Engel et al., 2017; Bitton et al., 2018; Y. Wu et al., 2023) and self-supervised and metric learning guided by domain knowledge (Esling et al., 2018; Tanaka et al., 2021; Tanaka et al., 2022; Cífka et al., 2021; Luo et al., 2020). Both TS-DSAE (Chapter 3) and J-DSAE (Chapter 4) leverage inductive biases introduced through architectural design and training strategies that are general and domain-agnostic (X. Liu et al., 2023; Y. Wu et al., 2025). As reviewed

in Section 2.3, these methods have demonstrated effective disentanglement for single-instrument audio, supporting applications such as pitch or timbre transfer and fine-grained control over synthesis.

However, these methods are fundamentally constrained to single-source scenarios, and their applicability to mixtures containing multiple instruments remains unclear. To fill this gap, the present chapter aims to disentangle pitch and timbre at the source level from mixtures of instruments, assigning each instrument a distinct latent representation with separable pitch and timbre components. Few studies have explicitly addressed the problem of disentanglement from multi-instrument mixtures. Hung et al. (2019) and Cwitkowitz et al. (2024) both focus on learning a global timbre representation from a mixture, for downstream tasks such as symbolic music rearrangement and transcription, respectively. Cheuk et al. (2023) propose a multi-task framework that improves pitch transcription by conditioning on timbre estimates, but do not explicitly learn disentangled pitch and timbre representations.

MSI-DIS (Lin et al., 2021) is a unified model for transcription, separation, and single-instrument generation from mixtures. While it supports two-instrument scenarios, its design is centred on synthesising individual instrument tracks rather than jointly sampling multiple sources. By contrast, the method proposed in this chapter focuses on mixture transformation by encoding and manipulating per-source pitch and timbre representations, and composing novel audio mixtures via a conditional diffusion transformer. Evaluation is conducted through source-level manipulations, which test the composability and disentanglement of the learnt latent space.

6.2.2 Object-Centric Representation Learning

In computer vision, learning object representations entails encoding entities within a visual scene as distinct representations (Greff et al., 2020; Locatello et al., 2020b). Recent work has also introduced attribute-level disentanglement within object representations, facilitating compositional scene generation (Singh et al., 2023; Z. Wu et al., 2024; Y.-F. Wu et al., 2024).

This paradigm has also been explored in the audio domain. Gha et al. (2023) and Reddy et al. (2023) learn separate representations for different sources from mixtures.

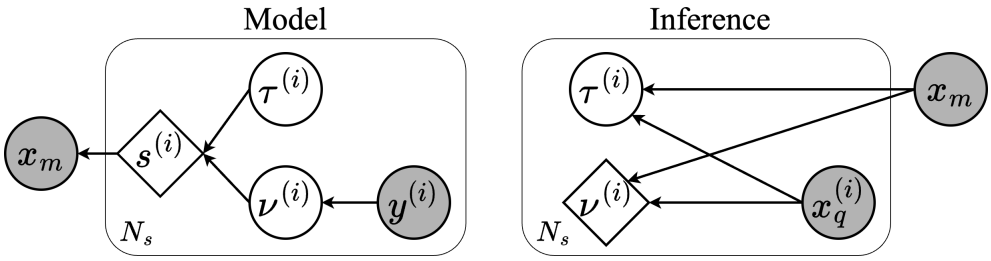

Figure 6.1: Plate notation of the proposed framework, DisMix. Diamond-shaped nodes indicate deterministic transformations of their parent nodes. Shaded nodes denote observed variables, while unshaded nodes represent latent variables. *Left*: The generative process samples, for the i-th source among N_s sources, a source-level latent $s^{(i)}$ composed of a timbre latent $\tau^{(i)}$ and a pitch latent $\nu^{(i)}$. The pitch latent is conditioned on its ground-truth pitch label $y^{(i)}$. All source-level latents $\{s^{(i)}\}_{i=1}^{N_s}$ jointly contribute to generate the observed mixture x_m. *Right*: The inference process independently infers each source-specific timbre latent $\tau^{(i)}$ and pitch latent $\nu^{(i)}$, conditioned on both the mixture x_m and a query $x_q^{(i)}$. The timbre of the query determines which instrument to extract from the mixture.

The approach proposed in this chapter treats each instrument as an audio object and disentangles its pitch and timbre into separable latent subspaces. This formulation supports fine-grained, compositional editing of musical scenes, such as recombining attributes across instruments or reassembling mixtures with novel configurations.

6.3 DisMix: The Proposed Framework

Given x_m, a mixture composed of N_s unique sources of musical instruments $\{x_s^{(i)}\}_{i=1}^{N_s}$, the goal is to extract, for each instrument, a source-level latent $s^{(i)}$ that combines representations of timbre $\tau^{(i)}$ and pitch $\nu^{(i)}$. These N_s source-level latents are used to condition a decoder to reconstruct the mixture. The proposed framework, DisMix, follows a generative process and an inference network depicted in Fig. 6.1.

6.3.1 Generative Process

Fig. 6.1 (left) shows the generative process of a mixture x_m:

1. Sample a timbre latent for the i-th source: $\tau^{(i)} \sim \mathcal{N}\big(0, \mathrm{I}\big)$;

2. Sample a pitch latent for the i-th source, conditioned on the pitch annotation $y^{(i)}$: $\nu^{(i)} \sim p_{\theta_\nu}(\nu^{(i)}|y^{(i)})$;

3. Deterministically set the source-level latent as a function of both latents: $s^{(i)} = \mathrm{f}_{\theta_s}(\tau^{(i)}, \nu^{(i)})$;

4. Generate the mixture: $x_m \sim p_{\theta_m}(x_m|\mathcal{S})$, where $\mathcal{S} = \{s^{(i)}\}_{i=1}^{N_s}$.

The networks with parameters θ_m, θ_s, and θ_ν, corresponding to the sampling of mixtures, source-level latents, and pitch latents, are specified in Sections 6.5 and 6.6 for the AE and LDM variants, respectively. Note that $\nu^{(i)}$ and $\tau^{(i)}$ are sampled independently, which allows pitch and timbre latents to serve as modular building blocks for mixture composition, and facilitates source-level pitch–timbre disentanglement and manipulation.

The joint distribution over the mixture, the timbre latents, and the pitch latents is written as:

$$p_{\theta_m}(x_m|\mathcal{S}) \prod_{i=1}^{N_s} \delta\Big(s^{(i)} - \mathrm{f}_{\theta_s}(\tau^{(i)}, \nu^{(i)})\Big)\, p_{\theta_\nu}(\nu^{(i)}|y^{(i)}) p(\tau^{(i)}), \tag{6.1}$$

where $\delta(\cdot)$ is the Dirac delta function. Here, the delta term encodes the deterministic mapping $s^{(i)} = \mathrm{f}_{\theta_s}(\tau^{(i)}, \nu^{(i)})$, meaning that the density has support only where this equality holds. DisMix is learnt by introducing an inference network, detailed in Section 6.3.2, to approximate posteriors over the latent variables that are otherwise intractable, and by maximising a lower bound on the marginal likelihood $p(x_m|\{y^{(i)}\}_{i=1}^{N_s})$ (Kingma et al., 2014; Rezende et al., 2014). The training objective is described in Section 6.4.

6.3.2 Inference Network

The inference network in Fig. 6.1 (right) extracts the timbre and pitch latent variables given both a mixture and a query, motivated by the query-based source separation framework (Lin et al., 2021; Lee et al., 2019; Y. Wang et al., 2022). The assumption is that the instruments of the constituent sources are known and the queries are available to extract the sources from a given mixture. A query $x_q^{(i)}$ shares the timbre characteristics with a constituent source $x_s^{(i)}$ of the mixture—both are played by the same instrument, regardless of pitch. However, instrument annotations are not

explicitly used in any of the loss functions in Section 6.4.

Each mixture x_m is paired with a set of N_s queries, randomly sampled from the datasets, that match the set of instruments in the mixture. The features $e_m = \mathrm{E}_{\phi_m}(x_m)$ and $e_q^{(i)} = \mathrm{E}_{\phi_q}(x_q^{(i)})$ are extracted, where $\mathrm{E}_{\phi_m}(\cdot)$ and $\mathrm{E}_{\phi_q}(\cdot)$ denote neural networks. The concatenation of the two embeddings, denoted as $e_{m,q}^{(i)}$, is used to extract pitch and timbre features specific to the i-th of N_s sources from the mixture.

As illustrated on the right of Fig. 6.1, both the pitch and timbre encoders take a common factorised form in the inference network:

$$q_{\phi_u}(\mathcal{U}|x_m, \{x_q^{(i)}\}_{i=1}^{N_s}) = \prod_{i=1}^{N_s} q_{\phi_u}(u^{(i)}|e_{m,q}^{(i)}), \tag{6.2}$$

where $u \in \{\nu, \tau\}$ and $\mathcal{U} \in \{\{\nu^{(i)}\}_{i=1}^{N_s}, \{\tau^{(i)}\}_{i=1}^{N_s}\}$. Given $e_{m,q}^{(i)}$, the pitch and timbre latents of the i-th source are encoded independently of each other and of other sources. Fig. 6.2 outlines the implementation of the inference network, and the parameterisation of both the pitch and timbre encoders is described as follows.

Pitch Encoder

The extraction of the pitch latent follows the steps of transcription, binarisation, and translation. The transcription task encourages timbre-invariant outputs, and the binarisation layer introduces a strong bottleneck to further remove information sensitive to timbre. The binarised output is finally transformed to the pitch latent. The three steps can be mathematically written as:

$$\begin{aligned} &\text{Transcription } \hat{y}^{(i)} = \mathrm{E}_{\phi_\nu}(e_{m,q}^{(i)}); \\ &\text{Binarisation } \hat{y}_{\text{bin}}^{(i)} = \mathbf{1}_{\{\text{Sigmoid}(\hat{y}^{(i)})>h\}}; \\ &\text{Translation } \ \hat{\nu}^{(i)} = \mathrm{f}_{\phi_\nu}(\hat{y}_{\text{bin}}^{(i)}). \end{aligned} \tag{6.3}$$

Both $\mathrm{E}_{\phi_\nu}(\cdot)$ and $\mathrm{f}_{\phi_\nu}(\cdot)$ are neural networks. $\mathbf{1}_{\{\cdot\}}$ is the indicator function that defines the stochastic binarisation layer SB and $h \sim \mathrm{U}(0, 1)$ is a threshold sampled for each training step and is fixed at 0.5 during the evaluation. The straight-through estimator (Bengio et al., 2013b) is used to bypass the non-differentiable stochastic operator.

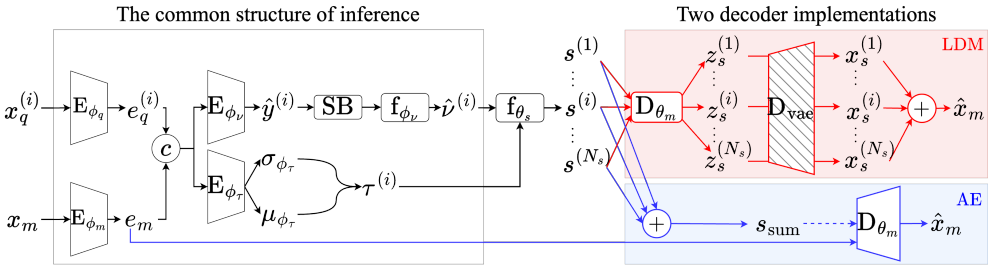

Figure 6.2: Two decoder variants—a simple autoencoder (AE, Section 6.5) and a latent diffusion model (LDM, Section 6.6)—are implemented to validate the proposed framework. While the two variants employ different mixture synthesis pipelines, illustrated by the red and blue boxes and arrows, they share the same architecture for extracting sets of timbre latents $\tau^{(i)}$ and pitch latents $\hat{\nu}^{(i)}$. *Encoders*: Given a mixture x_m comprising N_s sources, a query $x_q^{(i)}$ is used to extract the pitch and timbre latents corresponding to the i-th source. Specifically, the embeddings of x_m and $x_q^{(i)}$ are concatenated and used to derive both latent variables. These two attributes are then combined to form the source-level latent $s^{(i)}$ via f_{θ_s}. This process is repeated N_s times to extract a set of source latents $\{s^{(i)}\}_{i=1}^{N_s}$. *Decoders*: The AE variant reconstructs the mixture from the embedding of the mixture e_m, regularised by a prior distribution parameterised by $\{s^{(i)}\}_{i=1}^{N_s}$. This design is reminiscent of a VAE. In contrast, the LDM variant first predicts the latent representations $\{z_s^{(i)}\}_{i=1}^{N_s}$ of the sources $\{x_s^{(i)}\}_{i=1}^{N_s}$, extracted from a pre-trained model, and reconstructs the mixture using the decoder $\mathrm{D}_{\mathrm{vae}}$ of that model. The details of each module are elaborated in the main text.

Altogether, the pitch encoder is defined as follows:

$$q_{\phi_\nu}(\nu^{(i)}|e_{m,q}^{(i)}) \coloneqq \delta(\nu^{(i)} - \hat{\nu}^{(i)}) = \delta(\nu^{(i)} - \mathrm{f}_{\phi_\nu}(\hat{y}_{\mathrm{bin}}^{(i)})), \tag{6.4}$$

where $\nu^{(i)}$ deterministically takes the value of $\hat{\nu}^{(i)}$ in (6.3). This is reflected in (6.6), a training objective of DisMix—the first term takes the pitch latent $\nu^{(i)} = \hat{\nu}^{(i)}$ to condition the mixture reconstruction, and the second term encourages the value $\hat{\nu}^{(i)}$ to adhere to the pitch prior.

Timbre Encoder

The timbre encoder parameterises the mean and variance of a Gaussian posterior:

$$q_{\phi_\tau}(\tau^{(i)}|e_{m,q}^{(i)}) = \mathcal{N}(\tau^{(i)}; \mu_{\phi_\tau}(e_{m,q}^{(i)}), \sigma_{\phi_\tau}^2(e_{m,q}^{(i)})\mathrm{I}), \tag{6.5}$$

The timbre latent is sampled from the Gaussian by the reparameterisation trick (Kingma et al., 2014; Rezende et al., 2014), i.e., $\tau^{(i)} = \mu_{\phi_\tau}(e^{(i)}_{m,q}) + \epsilon\sigma_{\phi_\tau}(e^{(i)}_{m,q})$, where $\epsilon \sim \mathcal{N}\big(0, \mathrm{I}\big)$. The standard Gaussian prior described in Section 6.3.1 is applied to constrain the information capacity of the timbre latent, which introduces the Kullback–Leibler divergence (KLD) term corresponding to the timbre prior in (6.6).

6.4 Training Objectives

6.4.1 Evidence Lower Bound

Based on the generative process and the inference network described in Section 6.3, a valid evidence lower bound (ELBO) to the marginal log-likelihood $\log p(x_m|\{y^{(i)}\}_{i=1}^{N_s})$ is written as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{ELBO}} = \underbrace{\mathbb{E}_{\prod_{i=1}^{N_s} q_{\phi_\tau}(\tau^{(i)}|e^{(i)}_{m,q})}\big[\log p_{\theta_m}(x_m|\{s^{(i)} = \mathrm{f}_{\theta_s}(\tau^{(i)}, \hat{\nu}^{(i)})\}_{i=1}^{N_s})\big]}_{\text{mixture reconstruction}} \\ + \underbrace{\sum_{i=1}^{N_s} \log p_{\theta_\nu}(\hat{\nu}^{(i)}|y^{(i)})}_{\text{pitch prior}} \\ - \underbrace{\sum_{i=1}^{N_s} \mathcal{D}_{\mathrm{KL}}\big(q_{\phi_\tau}(\tau^{(i)})\|p(\tau^{(i)})\big)}_{\text{timbre prior}}.\end{aligned} \tag{6.6}$$

Intuitively speaking, a pitch and a timbre latent for each source are extracted given a mixture and a source-dependent query, the collection of which is used to reconstruct the mixture. The pitch and timbre latents are constrained by the information bottleneck imposed by their prior distributions to encourage the disentanglement of the two attributes.

Section 6.4.2 further introduces the auxiliary objectives used to explicitly separate each source-level latent from the mixture and enhance the pitch-timbre disentanglement within each source. As illustrated in Fig. 6.2, while the LDM and AE variants share the same inference network described in Section 6.3.2, they have different implementations for the modules corresponding to the parameters θ_m, θ_s, and θ_ν, which are detailed in Section 6.5 and Section 6.6, respectively.

6.4.2 Disentanglement by Auxiliary Supervision

Source Supervision

In addition to observing the mixtures, it is assumed that individual sources, $\{x_s^{(i)}\}_{i=1}^{N_s}$, are also available during training. To facilitate the extraction of source-level latents, source-wise reconstruction terms are introduced to the objective:

$$\mathcal{L}_{\text{Src}} = \sum_{i=1}^{N_s} \mathbb{E}_{q_{\phi_\tau}(\tau^{(i)}|e_{m,q}^{(i)})} \big[\log p_{\theta_m}(x_s^{(i)}|s^{(i)} = \mathrm{f}_{\theta_s}(\tau^{(i)}, \hat{\nu}^{(i)})) \big]. \tag{6.7}$$

Note that the decoder with parameters θ_m is reused for source reconstruction.

Pitch Supervision

Extracting the pitch latent begins with pitch transcription by $\mathrm{E}_{\phi_\nu}(\cdot)$ as specified in (6.3). The transcription is supervised by minimising a binary cross entropy loss:

$$\mathcal{L}_{\text{BCE}} = \text{BCE}(\hat{y}^{(i)} = \mathrm{E}_{\phi_\nu}(e_{m,q}^{(i)}), y^{(i)}), \tag{6.8}$$

where $y^{(i)}$ is the ground-truth pitch annotation of the i-th source.

Timbre Supervision

The query $x_q^{(i)}$ and the constituent source $x_s^{(i)}$ can be seen as two "views", as they have different pitch values or melodies, with a common semantic "object", for them sharing the same instrument. Similar to how common self-supervised learning techniques consider an image of an object and its rotated version as a positive pair that is semantically close (T. Chen et al., 2020), a simplified Barlow Twins loss (Zbontar et al., 2021) is introduced to enhance the correlation between the query and the timbre latent. Specifically, the objective is to minimise:

$$\mathcal{L}_{\text{BT}} = \sum_{i=1}^{N_s} \sum_{d=1}^{D_\tau} (1 - \mathcal{C}_{dd}(e_q^{(i)}, \tau^{(i)}))^2 \tag{6.9}$$

where $\mathcal{C}$ is a cross-correlation matrix, and both $e_q^{(i)}$ and $\tau^{(i)}$ share the same dimensionality D_τ.

Empirically, the standard Gaussian prior can over-regularise and collapse the timbre latent, similar to the effect of posterior collapse discussed in Section 2.2.2 and throughout the thesis. $\mathcal{L}_{\mathrm{BT}}$ counteracts this effect and promotes a discriminative timbre space.

6.4.3 The Final Objective

In summary, the aggregated objective function to be maximised is written as follows, with all loss terms weighted equally in the experiments:

$$\mathcal{L}_{\mathrm{DisMix}} = \mathcal{L}_{\mathrm{ELBO}} + \mathcal{L}_{\mathrm{Src}} - \mathcal{L}_{\mathrm{BCE}} - \mathcal{L}_{\mathrm{BT}}, \tag{6.10}$$

which can be interpreted as a VAE augmented by inductive biases introduced to separate source-level latents by $\mathcal{L}_{\mathrm{Src}}$, and to disentangle pitch and timbre of each source by $\mathcal{L}_{\mathrm{BCE}}$ and $\mathcal{L}_{\mathrm{BT}}$, respectively. The next sections elaborate on the architectural implementations of the two DisMix variants: the AE and the LDM.

6.5 A Simple Case Study Using an Autoencoder

This section first introduces a hierarchical variant of DisMix implemented by an AE with simple model architectures. The experiments and results are then reported which identify the key loss functions for pitch-timbre disentanglement.

6.5.1 A Hierarchical Model

The Gaussian likelihood $p_{\theta_m}(x_m|\mathcal{S})$ in $\mathcal{L}_{\mathrm{ELBO}}$ (6.6) is parameterised by a decoder D_{θ_m}. Although a permutation-invariant decoder such as a transformer (Vaswani et al., 2017) could be employed to reconstruct a mixture from the set of source-level latents (red box in Fig. 6.2), this section first examines an alternative parameterisation that does not require specific architectural choices (blue box in Fig. 6.2).

In this variant of DisMix, the mixture embedding $e_m = \mathrm{E}_{\phi_m}(x_m)$ is used to reconstruct the mixtures, thereby avoiding reliance on specialised network architectures that may be computationally expensive and architecturally complex. Specifically, the

ELBO of this DisMix variant is written as follows:

$$\begin{aligned}\mathcal{L}'_{\text{ELBO}} = &\underbrace{\mathbb{E}_{\delta(z_m - \mathrm{E}_{\phi_m}(x_m))}\big[\log p_{\theta_m}(x_m|z_m)\big]}_{\text{mixture reconstruction}} \\ &+ \underbrace{\mathbb{E}_{\prod_{i=1}^{N_s} q_{\phi_\tau}(\tau^{(i)}|e^{(i)}_{m,q})}\big[\log p(z_m|\{s^{(i)} = \mathrm{f}_{\theta_s}(\tau^{(i)}, \hat{\nu}^{(i)})\}_{i=1}^{N_s})\big]}_{\text{mixture embedding prior}} \\ &+ \underbrace{\sum_{i=1}^{N_s} \log p_{\theta_\nu}(\hat{\nu}^{(i)}|y^{(i)})}_{\text{pitch prior}} \\ &- \underbrace{\sum_{i=1}^{N_s} \mathcal{D}_{\text{KL}}\big(q_{\phi_\tau}(\tau^{(i)})\|p(\tau^{(i)})\big)}_{\text{timbre prior}},\end{aligned} \tag{6.11}$$

where z_m is introduced as an additional latent variable to encode x_m. Note the similarity between the first term from $\mathcal{L}_{\text{ELBO}}$ and the second term from $\mathcal{L}'_{\text{ELBO}}$—the latter can be seen as reconstructing the mixture in the latent space, as the target is z_m instead of x_m.

In the first term of $\mathcal{L}'_{\text{ELBO}}$, the delta function, with respect to which the expectation is taken, suggests that z_m is output by the mixture encoder $\mathrm{E}_{\phi_m}(x_m)$ deterministically. Furthermore, the second term, or the mixture embedding prior, is parameterised by the summation of the source-level latents. Therefore, the first two terms can be re-written as follows:

$$\underbrace{\log p_{\theta_m}(x_m|e_m = \mathrm{E}_{\phi_m}(x_m))}_{\text{mixture reconstruction}}; \quad \text{and} \tag{6.12}$$

$$\underbrace{\mathbb{E}_{\prod_{i=1}^{N_s} q_{\phi_\tau}(\tau^{(i)}|e^{(i)}_{m,q})}\big[\log p(e_m|\sum_{i=1}^{N_s} s^{(i)})\big]}_{\text{mixture embedding prior}}. \tag{6.13}$$

In other words, the summation is used as the permutation-invariant function that takes as input the set of source-level latents to reconstruct e_m in the latent space. Fig. 6.2 illustrates (6.12) with the blue arrow connecting e_m to D_{θ_m}, and (6.13) by the dashed blue arrow drawn from s_{sum} to D_{θ_m}. Concretely, while e_m is used to reconstruct x_m during training, s_{sum} can be used instead during evaluation.

To summarise, the hierarchical variant uses the embedding of mixtures e_m for

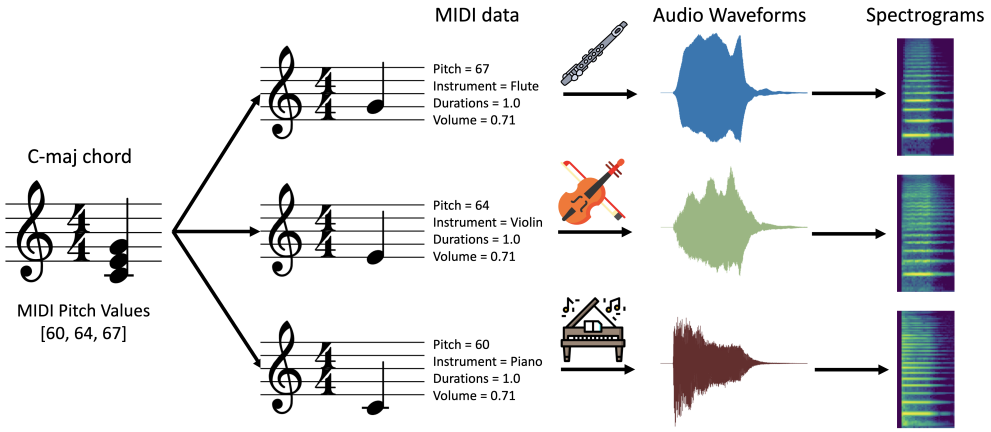

Figure 6.3: The rendition of chords of multiple instruments, reproduced from Gha et al. (2023). In this experiment, the spectrograms shown in the figure correspond to $x_s^{(i)}$ and the spectrogram derived from the summation of the audio waveforms is x_m.

mixture reconstruction, and the sum of the source-level latents to reconstruct e_m. This formulation does not require the decoder D_{θ_m} to be permutation-invariant, unlike the LDM variant which will be elaborated in Section 6.6, and provides a proof-of-concept for DisMix for its architectural simplicity.

An alternative design choice is to simply parameterise the first term of the ELBO (6.6) by

$$\underbrace{\mathbb{E}_{\prod_{i=1}^{N_s} q_{\phi_\tau}(\tau^{(i)}|e_{m,q}^{(i)})}\left[\log p_{\theta_m}(x_m|\sum_{i=1}^{N_s} s^{(i)}])\right]}_{\text{mixture reconstruction}} \tag{6.14}$$

and maintain the original formulation. However, empirical experiments suggest that the optimisation of the source-level latents and the decoder becomes more challenging and struggles to converge without using e_m as a guide, especially during the early training phase. Although more expressive decoders could potentially mitigate the issue, introducing e_m as an intermediary facilitates the training without complicating the architecture.

6.5.2 Dataset

The MusicSlots dataset (Gha et al., 2023) comprises 3,131 unique chords derived from JSB Chorales (Boulanger-Lewandowski et al., 2012). Each chord is rendered as

an audio mixture using the sound fonts of piano, violin, and flute via FluidSynth. For each composite note within a chord, a sound font is randomly sampled with replacement, allowing a single instrument to render multiple notes ($N_s \in \{1, 2, 3\}$). The individual note waveforms are summed to form the final chord mixture. The rendering process is illustrated in Fig. 6.3.

Each source $x_s^{(i)}$ is defined as the set of notes rendered by a single sound font and is annotated with a multi-hot pitch vector $y^{(i)} \in \{0, 1\}^{N_p}$, where $N_p = 52$ denotes the number of pitch classes covered by the dataset. The full dataset consists of 28,179 chord mixtures, partitioned into training, validation, and test sets in a 70/20/10 ratio.

Waveforms are sampled at 16 kHz and transformed into mel spectrograms using 128 mel frequency bands, a window size of 1,024, and a hop length of 512. A 320 ms segment is cropped from the sustain phase of each mixture, corresponding to 10 spectral frames, and is used for both training and evaluation. This results in representations $x_s^{(i)}$, $x_m \in \mathbb{R}^{128\times10}$.

Inp. size	**Out. size**	**Kernel size**	**Stride**	**Padding**	**Normalization**	**Activation**
128	768	3	1	0	Layer	ReLU
768	768	3	1	1	Layer	ReLU
768	768	4	2	1	Layer	ReLU
768	768	3	1	1	Layer	ReLU
768	768	3	1	1	Layer	ReLU
768	64	1	1	1	None	None

Table 6.1: The Conv1D layers used to implement E_{ϕ_m} and E_{ϕ_q} in the AE.

Module	**Input size**	**Output size**	**Normalization**	**Activation**
E_{ϕ_ν}	128	256	Layer	ReLU
	256	256	Layer	ReLU
	256	256	Layer	ReLU
	256	256	Layer	ReLU
	256	256	Layer	ReLU
	256	256	Layer	ReLU
	256	52	None	Sigmoid
Stochastic Binarisation (SB)				
f_{ϕ_ν}	52	64	None	ReLU
	64	64	None	ReLU
	64	64	None	None

Table 6.2: The three modules of the pitch encoder implemented in the AE: the transcriber E_{ϕ_ν}, the SB layer, and the projector f_{ϕ_ν}, as described in Section 6.3.2.

6.5.3 Implementation

This section provides details of the implementation of the AE modules illustrated in Fig. 6.2, including the decoder and the encoders for mixtures, queries, pitch, and timbre. [2]

Mixture and Query Encoders

The mixture encoder E_{ϕ_m} and the query encoder E_{ϕ_q}, introduced in Section 6.3.2, share an architecture comprising a stack of Conv1D layers. These encoders take as input the mel spectrograms x_m and the query $x_q^{(i)}$, respectively, and produce representations $e_m, e_q^{(i)} \in \mathbb{R}^{64}$. The architectural details are summarised in Table 6.1. Each encoder is followed by mean pooling along the temporal dimension, reducing a mel spectrogram of shape $\mathbb{R}^{128\times 10}$ to a 64-dimensional vector. The outputs e_m and $e_q^{(i)}$ are concatenated to form a 128-dimensional embedding, which is used to infer the pitch and timbre latents.

Pitch Encoder

Given this concatenated representation, the pitch encoder first predicts $\hat{y}^{(i)} \in \mathbb{R}^{N_p}$ using $\mathrm{E}_{\phi_\nu} : \mathbb{R}^{128} \to \mathbb{R}^{N_p}$, where $N_p = 52$ denotes the number of pitch classes. A pitch latent $\nu^{(i)} \in \mathbb{R}^{64}$ is then computed by applying $\mathrm{f}_{\phi_\nu} : \mathbb{R}^{N_p} \to \mathbb{R}^{64}$ to the binarised prediction $\hat{y}_{\mathrm{bin}}^{(i)}$. The full architecture of the pitch pathway is provided in Table 6.2.

Timbre Encoder

The timbre encoder shares the same architecture as E_{ϕ_ν}, except that its final layer is replaced by a Gaussian parameterisation layer. This consists of two linear layers projecting the 256-dimensional hidden state to the mean and log-variance of the posterior distribution, from which the timbre latent $\tau^{(i)} \in \mathbb{R}^{64}$ is sampled.

[2] The initial focus of this project was to build upon Takida et al. (2022), which adopt the model architecture of Niekerk et al. (2020b). Although the research direction subsequently changed, the architecture was largely retained without tuning.

Source-Level Latents

The source-level representation is derived using FiLM (Perez et al., 2018; J. W. Kim et al., 2019), formulated as

$$s^{(i)} = \mathrm{f}_{\theta_s}(\tau^{(i)}, \hat{\nu}^{(i)}) = \alpha_{\theta_s}(\tau^{(i)}) \odot \hat{\nu}^{(i)} + \beta_{\theta_s}(\tau^{(i)}), \tag{6.15}$$

where the pitch latent $\hat{\nu}^{(i)}$ is modulated element-wise by scaling and shifting factors α_{θ_s} and β_{θ_s}, both parameterised by the timbre latent $\tau^{(i)}$. The FiLM layer thus integrates pitch and timbre by linearly transforming $\tau^{(i)}$ to produce the modulation parameters, which are applied to $\hat{\nu}^{(i)}$ to yield the source-level latent $s^{(i)}$.

Reconstruction

The decoder $p_{\theta_m} : \mathbb{R}^{64\times 10} \to \mathbb{R}^{128\times 10}$ comprises a combination of RNN and MLP layers and is responsible for reconstructing the mel spectrograms of either x_m (as in (6.12)) or $x_s^{(i)}$ (as in (6.7)). To perform reconstruction, the source-level latent $s^{(i)}$ is temporally broadcast to match the number of spectral frames (10). A two-layer bidirectional gated recurrent unit (GRU) processes the broadcast latent to produce a sequence of shape $\mathbb{R}^{128\times 10}$, which is subsequently passed through a linear layer to yield the final reconstructed spectrogram.

6.5.4 Prior Distributions

While Section 6.3.1 specifies that the timbre prior $p(\tau^{(i)})$ follows a standard normal distribution, this section focuses on the pitch prior $p_{\theta_\nu}(\nu^{(i)}|y^{(i)})$ and the mixture embedding prior $p(e_m|s_{\text{sum}})$, as defined in (6.13), both of which are implemented in the AE variant.

Pitch Prior

Pitch priors are studied at different levels of expressivity. The first, following (6.1), defines a conditional Gaussian:

$$p_{\theta_\nu}(\nu^{(i)}|y^{(i)}) = \mathcal{N}(\nu^{(i)}; \mu^{\text{fac}}_{\theta_\nu}(y^{(i)}), \sigma_\nu^2 \mathrm{I}), \tag{6.16}$$

where $\mu^{\text{fac}}_{\theta_\nu}$ is a neural network conditioned on the ground-truth pitch annotation $y^{(i)}$, and $\sigma_\nu = 0.5$ is a fixed hyperparameter. Specifically, the factorised prior network $\mu^{\text{fac}}_{\theta_\nu}$ first reuses the MLP $\mathrm{f}_{\phi_\nu} : \mathbb{R}^{N_p} \to \mathbb{R}^{64}$ to project $y^{(i)} \in \mathbb{R}^{N_p}$. The remaining layers follow the architecture of E_{ϕ_ν} in Table 6.2 (without reusing any layers), with two key differences: the input size is reduced to 64, and the final layer is replaced by two linear layers that parameterise the mean and log-variance of a 64-dimensional Gaussian distribution.

A more expressive prior is also considered to model inter-source dependencies, defined as a mixture of Gaussians:

$$p_{\theta_\nu}(\nu^{(i)}|\mathcal{Y}_{\setminus i}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\big(\nu^{(i)}; \mu^{\text{rich}}_{\theta_\nu,k}(\mathcal{Y}_{\setminus i}), \sigma_\nu^2 \mathrm{I}\big), \tag{6.17}$$

where $\pi_k = \frac{1}{K}$, $\mathcal{Y}_{\setminus i}$ denotes the set of pitch annotations excluding that of the i-th source (the target pitch), and $\mu^{\text{rich}}_{\theta_\nu,k}$ specifies the mean of the k-th Gaussian component. The motivation is that $\hat{y}^{(i)}$ may be conditionally dependent on the pitches of other sources in the mixture, reflecting musical harmony—i.e., the inference of the pitch latent is contextualised by accompanying sources. Furthermore, K Gaussian components are used to capture multiple possible outcomes of the target pitch value given the accompanying pitches. Note that the cross-entropy loss for the target pitch (6.8) is still applied, and this prior is considered more expressive than its factorised counterpart $\mu^{\text{fac}}_{\theta_\nu}$ due to its contextual information.

The expressive prior network applies f_{ϕ_ν} to each element in $\mathcal{Y}_{\setminus i}$, followed by a transformer to model the set. The architecture consists of three blocks of a post-norm transformer, as illustrated in Fig. 6.4, with four attention heads and an embedding size of 64. The feed-forward block is a two-layer MLP with ReLU activation and a hidden size of 64.

To preserve permutation invariance, no positional encodings are used; the elements in $\mathcal{Y}_{\setminus i}$ are treated as an unordered set of tokens. The transformer output is aggregated via max pooling over the $N_s - 1$ tokens to produce a hidden state, which is then passed to the K Gaussian parameterisation heads, each producing the mean and log-variance of a Gaussian component.

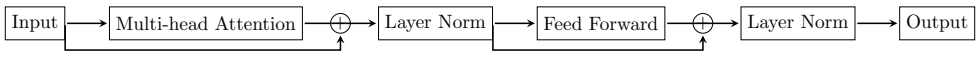

Figure 6.4: A regular post-norm transformer block.

Another motivation for parameterising the prior using only the accompanying sources while excluding the target source is to potentially support iterative sampling of pitch latents conditioned on an available set of pitch latents that have already been sampled. This application is left for future exploration (Parker et al., 2024; Mariani et al., 2024).

Mixture Embedding Prior

The AE variant introduces a latent representation of mixtures and maximises (6.13), the likelihood associated with the mixture embedding e_m. This is defined as a Gaussian:

$$p(e_m \mid \{\tau^{(i)}, \hat{\nu}^{(i)}\}_{i=1}^{N_s}) = \mathcal{N}\left(e_m; \sum_{i=1}^{N_s} s^{(i)}, \sigma_m^2 \mathrm{I}\right), \tag{6.18}$$

where $\sigma_m = 0.25$ is a fixed hyperparameter. Recall that $s^{(i)} = \mathrm{f}_{\theta_s}(\tau^{(i)}, \hat{\nu}^{(i)})$

6.5.5 Optimisation

Training is performed using the Adam optimiser (Kingma et al., 2015) with a batch size of 32, a learning rate of 0.0004, and a gradient clipping threshold of 0.5. Early stopping is applied if the loss fails to improve on the validation set for 260k steps.

6.5.6 Experiments

Pitch–Timbre Disentanglement

Given a mixture x_m and a set of queries $\{x_q^{(i)}\}_{i=1}^{N_s}$ corresponding to the N_s constituent instruments, the timbre and pitch latents $\{\tau^{(i)}, \nu^{(i)}\}_{i=1}^{N_s}$ are first extracted. To evaluate disentanglement, a random permutation is applied such that the pitch latent of source i is replaced with that of source j, while the timbre remains unchanged. This yields a novel source-level latent $\hat{s}^{(i)} = \mathrm{f}_{\theta_s}(\tau^{(i)}, \nu^{(j)})$. A new source $\hat{x}_s^{(i)}$ is then rendered by passing $\hat{s}^{(i)}$ to the decoder D_{θ_m}. Note that sources, rather than mixtures,

are rendered by providing $\hat{s}^{(i)}$ to the decoder. This is made feasible by the auxiliary source supervision that reuses the decoder to reconstruct individual sources, as detailed in Section 6.4.2.

To evaluate whether $\hat{s}^{(i)}$ carries the timbre and pitch attributes of sources i and j, respectively, classifiers are pre-trained using the training set. The instrument and pitch classifiers share the same architecture, differing only in the output layer, which produces either three instrument classes or 52 pitch classes. Both classifiers use the encoder architecture from Table 6.1, followed by a three-layer MLP. The first two MLP layers have input and output dimensions of 64 with ReLU activation functions. The final layer projects from 64 to the number of target classes.

Disentanglement is considered successful if the classifiers predict the pitch of $x_s^{(j)}$ (the swapped pitch) and the instrument of $x_s^{(i)}$ (the preserved timbre) for each generated source $\hat{x}_s^{(i)}$.

The resulting classification accuracy is reported under "Disentanglement" in Table 6.3, which evaluates DisMix without a pitch prior by omitting the second term in $\mathcal{L}_{\text{ELBO}}$ (6.6), as well as the effects of removing individual components that constrain the timbre and pitch latent spaces. Effects of the pitch prior are reported in Table 6.4.

Table 6.3 suggests that the removal of the KLD term in (6.6), or the timbre prior $p(\tau^{(i)})$, results in contamination of the timbre latent $\tau^{(i)}$ with pitch-related information. Consequently, the pitch of the synthesised output $\hat{x}_s^{(i)}$ is influenced not only by $\nu^{(j)}$, but also by $\tau^{(i)}$, leading to a marked decrease in pitch classification accuracy to 69.41%.

In contrast, eliminating the SB layer introduces excessive capacity into the pitch latent representation. This allows the instrument identity of $\hat{x}_s^{(i)}$ to be influenced by both $\tau^{(i)}$ and $\nu^{(j)}$, resulting in a substantial drop in instrument classification accuracy to 46.71%. Although the BCE loss for pitch supervision remains active during training, the model fails to disentangle pitch and timbre without the SB constraint, indicating that supervision alone is insufficient.

	Disentanglement		Mixture Rendering	
	Pitch	Instrument	Pitch	Instrument
DisMix w/o p_{θ_ν}	93.39%	100.00%	90.69%	100.00%
- $\mathcal{L}_{\text{BT}}$	93.18%	99.92%	87.92%	100.00%
- KLD	69.41%	100.00%	35.10%	100.00%
- SB	93.46%	46.71%	40.23%	98.91%

Table 6.3: Ablation study examining the impact of individual loss components in DisMix, evaluated on pitch–timbre disentanglement and the rendering of novel mixtures. Classification accuracy (%) is reported using pre-trained pitch and instrument classifiers. The study isolates the effects of removing the auxiliary loss $\mathcal{L}_{\text{BT}}$, the KLD on the timbre space, and the stochastic binarisation within the pitch encoding pipeline.

	DisMix w/o p_{θ_ν}	+fac	+rich ($K = 1$)	+rich ($K = 10$)
Disentanglement	93.39	93.98	94.01	94.18
Mixture Rendering	90.69	91.15	91.39	92.04

Table 6.4: Pitch classification accuracy (%) using different pitch priors.

Novel Mixture Rendering

To assess the model's ability to synthesise novel mixtures, a new set $\{\hat{s}^{(i)}\}_{i=1}^{N_s}$ is formed by applying the same permutation to the source-level latents. A novel mixture $\hat{x}_m$ is then rendered by decoding the sum $\hat{s}_{\text{sum}} = \sum_{i=1}^{N_s} \hat{s}^{(i)}$ via D_{θ_m}. To verify that the attributes of $\hat{x}_m$ reflect the manipulated $\{\hat{s}^{(i)}\}_{i=1}^{N_s}$, source-level representations are re-extracted using the original queries $\{x_q^{(i)}\}_{i=1}^{N_s}$, and the reconstructed sources are again passed to the pre-trained classifiers.

The results are reported under "Mixture Rendering" in Table 6.3. The removal of the auxiliary loss $\mathcal{L}_{\text{BT}}$ negatively affects pitch accuracy. This may be attributed to the timbre prior $p(\tau^{(i)})$, which can over-regularise the posterior distribution, yielding high-variance, noisy samples of $\tau^{(i)} \sim q_{\phi_\tau}(\tau^{(i)}|e_m, e_q^{(i)})$. Such variance may hinder the optimisation of (6.13), amplifying the mismatch between mixture representations used in training (e_m) and evaluation (s_{sum}). As illustrated on the left of Fig. 6.5, the removal of $\mathcal{L}_{\text{BT}}$ disrupts the instrument-related structure in the PCA projection of the timbre latent, suggesting degraded organisation due to noisy sampling.

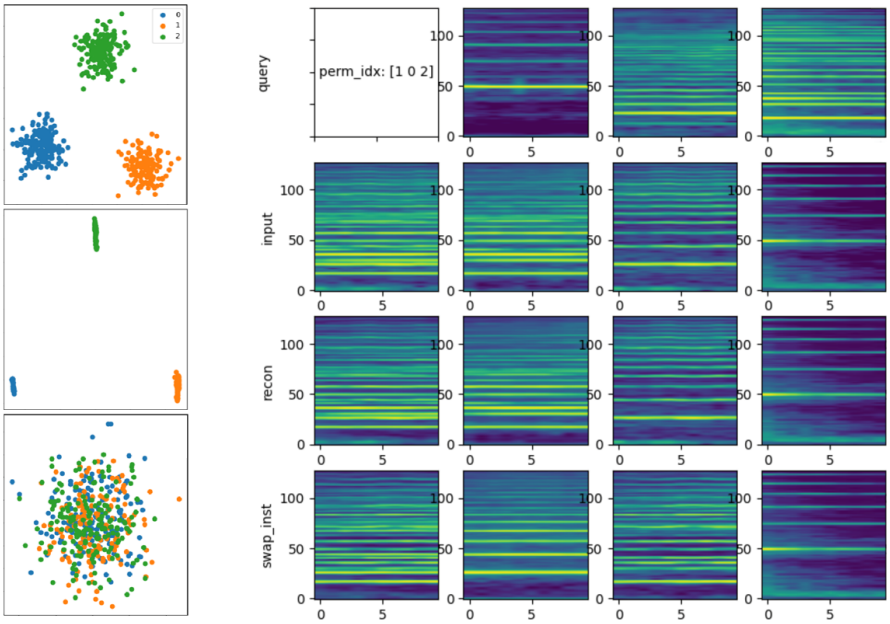

Figure 6.5: *Left*: PCA of the timbre latent space. Top: DisMix with full objective, showing distinct clustering of $\tau^{(i)} \sim q_{\phi_\tau}(\tau^{(i)}|e^{(i)}_{m,q})$ by instrument class. Middle: mean of the posterior distribution $q_{\phi_\tau}(\tau^{(i)}|e^{(i)}_{m,q})$. Bottom: samples from the same distribution exhibit high variance and loss of structure in the absence of $\mathcal{L}_{\text{BT}}$. *Right*: mel spectrograms illustrating novel mixture rendering. First row: query examples for three instruments. Second row: input mixture x_m and the corresponding reconstructed sources. Third row: reconstructed mixture and sources from s_{sum} and individual $s^{(i)}$. Final row: manipulated mixture rendered from $\hat{s}_{\text{sum}}$, followed by sources re-extracted using the original queries. The first two sources reflect successful pitch–timbre swapping; the third source remains unchanged.

Pitch Prior

Table 6.4 reports the performance of the model under different choices of pitch priors. The variants fac and rich correspond to the factorised and expressive priors described in Section 6.5.4. For the expressive prior, experiments are conducted with $K = 1$ and $K = 10$ Gaussian components.

A progressive improvement is observed with increasing prior complexity. The performance gain from fac to rich with $K = 1$ is modest compared to the gain achieved by using rich with $K = 10$. This suggests that the pitch distribution, when conditioned on contextual pitch information, is inherently multimodal, and a single Gaussian component may not offer substantial advantage over the factorised form.

Visualisation

Fig. 6.5 shows mel spectrograms on the right side. The first row shows queries for different instruments. The second row shows an input mixture x_m and the corresponding sources. The timbre characteristics are consistent across rows for the last three columns. In the third row, the leftmost is the reconstructed mixture given s_{sum}, and the rest are the reconstructed sources given $s^{(i)}$.

The first column of the last row refers to the manipulated mixture rendered by $\hat{s}_{\text{sum}}$, and the rest are the sources extracted from the mixture using the original queries. Attributes are successfully manipulated. In the last row, the second column combines the first source's timbre with the second's pitch, and the third column combines the second's timbre with the first's pitch. The third source is unchanged.

6.6 A Latent Diffusion Framework

This section elaborates on the LDM variant of DisMix, where the likelihood $p_{\theta_m}(x_m|\mathcal{S})$ is modelled by a diffusion transformer (DiT) (Peebles et al., 2023). Unlike the AE variant, the LDM counterpart does not introduce additional latent variables and objective functions, preserving the original DisMix formulation. The DiT directly reconstructs a mixture from a set of source-level latents $\mathcal{S} = \{s^{(i)}\}_{i=1}^{N_s}$ using an expressive permutation-invariant decoder.

6.6.1 Data Representation

The LDM framework (Rombach et al., 2022) enhances the efficiency of diffusion models (DMs) (Sohl-Dickstein et al., 2015; Ho et al., 2020) by conducting generation in a learnt latent space. A pre-trained VAE from AudioLDM2 (H. Liu et al., 2024) is employed to map audio waveforms to latent representations. Each source $x_s^{(i)}$ is encoded as $z_s^{(i)} = \text{E}_{\text{vae}}(x_s^{(i)})$ and reconstructed using $x_s^{(i)} = \text{D}_{\text{vae}}(z_s^{(i)})$, where E_{vae} and D_{vae} denote the pre-trained encoder and decoder, respectively.

6.6.2 Latent Diffusion Models

To contextualise the implementation of the DiT model, the fundamentals of DMs are briefly reviewed. LDMs operate DMs in a compressed latent space, where the target latent feature z_0 is sampled via a Markov chain defined as

$$p(z_T) \prod_{t=1}^{T} p_\theta(z_{t-1} \mid z_t), \tag{6.19}$$

with

$$p(z_T) = \mathcal{N}(z_T; 0, \mathrm{I}), \tag{6.20}$$

and

$$p_\theta(z_{t-1} \mid z_t) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t), \Sigma_\theta(z_t, t)), \tag{6.21}$$

where μ_θ and Σ_θ are parameterised by neural networks. This defines an iterative denoising procedure—also referred to as the reverse process—which progressively refines Gaussian noise into a structured latent variable through T steps.

The forward (noising) process is given by a linear Gaussian posterior:

$$q(z_t \mid z_{t-1}) = \mathcal{N}(z_t; \sqrt{\alpha_t} z_{t-1}, (1 - \alpha_t)\mathrm{I}), \tag{6.22}$$

where α_t is a diffusion hyperparameter defined by a pre-specified noise schedule. Since the forward process is analytically tractable and fixed, Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020) define μ_θ and Σ_θ to match the form of the posterior and simplify training to the minimisation of a denoising objective:

$$\|\mathrm{f}_\theta(z_t, t) - z_0\|_2^2, \tag{6.23}$$

which corresponds to learning a decoder f_θ that predicts the clean latent z_0 from its noisy counterpart z_t.

6.6.3 Mixture Reconstruction

As illustrated by the red arrows in Fig. 6.2, the mixture x_m is reconstructed by summing its constituent sources, each recovered from the latent representation $z_s^{(i)}$ using the VAE decoder $\mathrm{D}_{\mathrm{vae}}$. To generate the set $\{z_s^{(i)}\}_{i=1}^{N_s}$, a DiT is trained to predict the clean source latents from their noised versions using source-level representations $\mathcal{S} = \{s^{(i)}\}_{i=1}^{N_s}$ as condition. This reframes the likelihood $p_{\theta_m}(x_m|\mathcal{S})$ in the ELBO (6.6) as:

$$p_{\theta_m}(\{z_{s,0:T}^{(i)}\}_{i=1}^{N_s}|\mathcal{S}) = p(\{z_{s,T}^{(i)}\}_{i=1}^{N_s}) \prod_{t=1}^{T} p_{\theta_m}(\{z_{s,t-1}^{(i)}\}_{i=1}^{N_s}|\{z_{s,t}^{(i)}\}_{i=1}^{N_s}, \mathcal{S}). \tag{6.24}$$

Here, $z_{s,t}^{(i)}$ denotes the noisy version of $z_s^{(i)}$ at diffusion step t, with $z_{s,0}^{(i)} = z_s^{(i)}$. The DiT is trained to denoise the entire set $\{z_{s,t}^{(i)}\}_{i=1}^{N_s}$ in a single forward pass, while evaluation proceeds via an iterative denoising trajectory over T steps. This design enables the model to capture interdependencies between sources, which is shown to be beneficial: the set-conditioned configuration (M1) outperforms its single-source-conditioned counterpart (M0), as reported in Table 6.7.

6.6.4 Dataset

The CocoChorale dataset (Y. Wu et al., 2022) provides realistic generative music in the style of Bach chorales (Boulanger-Lewandowski et al., 2012), featuring 13 orchestral instruments rendered within the pitch ranges of a standard four-part chorale (i.e., Soprano, Alto, Tenor, Bass, or SATB), with $N_s = 4$ sources per mixture.

The random ensemble subset is used, comprising 60,000 four-part chorales and totalling approximately 350 hours of audio. For each example, the instruments for the four parts are independently drawn from predefined pools associated with each voice part. The pitch annotation $y^{(i)} \in \{0,1\}^{N_p \times T_p}$ is a time sequence of one-hot vectors, where $N_p = 129$ denotes the number of discrete pitch bins. The dataset is split into training, validation, and test sets using a ratio of 80/10/10.

Each audio file is sampled at 16 kHz and converted into mel spectrograms using 64 mel frequency banks, with a window size of 1,024 and a hop length of 160. A

four-second segment is randomly selected from each audio example, corresponding to $T_p = 400$ time frames.

6.6.5 Implementation

Encoders

Mel spectrograms of 64 frequency bands and 400 time frames are extracted. The mixture encoder E_{ϕ_m} processes the mel spectrogram of x_m and outputs $e_m \in \mathbb{R}^{8\times100\times16}$, where the dimensions denote channels, time frames, and feature size. The encoder E_{ϕ_m} is instantiated and frozen from $\mathrm{E}_{\mathrm{vae}}$, the pre-trained VAE encoder in AudioLDM2 (H. Liu et al., 2024).

The same architecture is reused for the query encoder E_{ϕ_q}, which is trained from scratch and augmented with a temporal pooling layer, yielding $e_q^{(i)} \in \mathbb{R}^{8\times1\times16}$. To integrate $e_q^{(i)}$ and e_m for latent extraction, $e_q^{(i)}$ is broadcast along the time dimension and concatenated with e_m along the feature axis. A 1×1 Conv2D layer projects the result back to the original 16-dimensional feature space.

Given the concatenated representation, the pitch encoder $\mathrm{E}_{\phi_\nu} : \mathbb{R}^{8\times100\times16} \to \mathbb{R}^{129\times400}$ predicts the pitch transcription. Its architecture is adapted from the decoder $\mathrm{D}_{\mathrm{vae}}$ of AudioLDM2, with the output dimension modified from 64 to 129 to accommodate the pitch class count. The output is binarised by an SB layer and passed to $\mathrm{f}_{\phi_\nu} : \mathbb{R}^{129\times400} \to \mathbb{R}^{8\times100\times16}$ to produce the pitch latent $\nu^{(i)}$, where f_{ϕ_ν} adopts the architecture of $\mathrm{E}_{\mathrm{vae}}$ and is trained from scratch.

The timbre encoder transforms an input of shape $\mathbb{R}^{8\times100\times16}$ into the timbre latent $\tau^{(i)} \in \mathbb{R}^{8\times1\times16}$, using the architecture specified in Table 6.5, followed by temporal pooling. The timbre latent is then broadcast along the time axis and concatenated with the pitch latent to form the source-level latent $s^{(i)} \in \mathbb{R}^{8\times100\times32}$.

Partition

DiTs are transformers that operate on sequences of patches (Peebles et al., 2023). The input sequences are constructed from $\{z_s^{(i)}\}_{i=1}^{N_s}$ and $\{s^{(i)}\}_{i=1}^{N_s}$. Sinusoidal positional

Inp. size	Out. size	Kernel size	Stride	Padding	Normalization	Activation
128	128	5	2	0	Group(1)	ReLU
128	128	5	2	0	Group(1)	ReLU
128	128	5	2	0	Group(1)	ReLU
128	256	1	1	0	None	None

Table 6.5: The Conv1D layers used to implement the timbre encoder for the LDM in Section 6.6. The last row refers to a Gaussian layer where the 256-dimensional output is split to represent the mean and log-variance of the posterior. The number in the parenthesis indicates the number of groups divided for normalisation.

embeddings are applied to each $z_s^{(i)}$, followed by partitioning:

$$\mathrm{Par}(z_s^{(i)}) : \mathbb{R}^{T_z \times (D_z \times C)} \to \mathbb{R}^{L \times D_z'}, \tag{6.25}$$

where $T_z = 100$, $D_z = 16$, $C = 8$, and $L = 25$. The patches are created along the time axis, and each is flattened to size $D_z' = \frac{T_z}{L} \times D_z \times C$. This is repeated for all N_s sources in $\{z_s^{(i)}\}_{i=1}^{N_s}$ to yield $z_m \in \mathbb{R}^{(N_s \times L) \times D_z'}$, representing the aggregated sequence of patches.

Similarly, the set of source-level representations $\mathcal{S} = \{s^{(i)}\}_{i=1}^{N_s}$ is partitioned into $s_c \in \mathbb{R}^{(N_s \times L) \times D_s'}$, where $D_s = 32$ and $D_s' = \frac{T_z}{L} \times D_s \times C$. Thus, z_m and s_c share structural alignment in terms of sequence length, with differing patch dimensions.

The decoder is then redefined as:

$$p_{\theta_m}(z_{m,0:T} \mid \mathcal{S}) = p(z_{m,T}) \prod_{t=1}^{T} p_{\theta_m}(z_{m,t-1} \mid z_{m,t}, s_c). \tag{6.26}$$

To ensure permutation invariance with respect to the N_s sources, no additional positional embeddings are added to z_m or s_c.

Conditioning

The conditioning mechanism for reconstructing $z_{m,t-1}$ from $z_{m,t}$ and s_c (as defined in (6.26)) is realised using adaptive layer normalisation (adaLN) (Peebles et al., 2023) within transformer blocks. Each block comprises a multi-head self-attention module and a feedforward network, both followed by skip connections and layer normalisation (Vaswani et al., 2017), with standard LN replaced by adaLN for conditioning.

To modulate the transformer with diffusion step information, the condition s_c is

augmented by adding a learnt embedding of the step index t. This condition is used to regress the scaling and shifting factors of adaLN, which are applied to the transformer activations. Consequently, $z_{m,t}$ is adaptively modulated using source-level pitch and timbre information.

Decoder

The DiT comprises three standard transformer blocks as shown in Fig. 6.4, with four attention heads per block. The embedding size matches the patch size D'_z, and all layer normalisation operations are replaced with adaLN to enable conditional decoding based on s_c.

6.6.6 Optimisation

Training uses the Adam optimiser (Kingma et al., 2015) with a batch size of 8 and gradient clipping at 0.5. The learning rate is linearly warmed up to 0.0001 over 308k steps, followed by cosine decay (Loshchilov et al., 2017) for up to 4,092k training steps.

6.6.7 Experiments

Setup

Given a reference mixture and its extracted set of pitch and timbre latents, the timbre latents $\{\tau^{(i)}\}_{i=1}^{N_s}$ are arbitrarily replaced with those from a separate target mixture. The goal is for the four instruments in the reference mixture to be substituted with those from the target mixture, while retaining the original melodic content. To ensure musical plausibility, each replacement is constrained such that the instrument from the target mixture corresponds to the same SATB part (Soprano, Alto, Tenor, or Bass) as in the reference. This guarantees that the substituted instruments perform melodies aligned with their characteristic pitch ranges. An example is shown in Fig. 6.6.

The modified latent set is then used to sample $\{\hat{z}_s^{(i)}\}_{i=1}^{N_s}$ from the reverse diffusion process defined in (6.24), over $T = 1000$ steps, starting with a standard Gaussian prior. The generated audio sources $\{\hat{x}_s^{(i)}\}_{i=1}^{N_s}$ are then evaluated using pre-trained

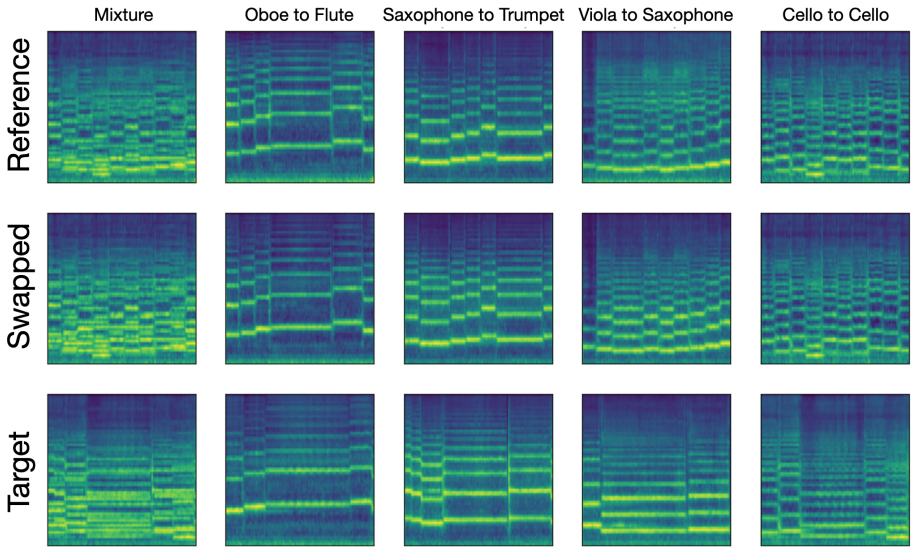

Figure 6.6: Instrument replacement in a reference mixture using the timbre latents from a target mixture. Each instrument is replaced with its counterpart in the target while preserving the original pitch content.

pitch and instrument classifiers. Each source $\hat{x}_s^{(i)}$ is expected to reflect the instrument identity from the target mixture and preserve the melody of the corresponding source in the reference mixture.

The resulting mixture is obtained by summing the rendered sources in the waveform domain:

$$\hat{x}_m = \sum_{i=1}^{N_s} \hat{x}_s^{(i)}. \tag{6.27}$$

Hence, the evaluation of pitch–timbre disentanglement corresponds to assessing whether the reconstructed mixture reflects the intended instrument identities and pitch content.

The architecture of the instrument classifier is summarised in Table 6.6, where the final layer projects to logits over 13 instrument classes. For pitch evaluation, the Crepe model (J. W. Kim et al., 2018) is employed—a state-of-the-art monophonic pitch estimator.[3] Fréchet Audio Distance (FAD) (Kilgour et al., 2019) is also computed to assess audio quality. The FAD compares VGG-style features extracted from the generated audio against those from the original sources and averages the results over

[3] `https://github.com/maxrmorrison/torchcrepe`

Inp. size	Out. size	Kernel size	Stride	Padding	Normalization	Activation
64	64	5	2	0	Group(1)	ReLU
64	64	5	2	0	Group(1)	ReLU
64	64	5	2	0	Group(1)	ReLU
64	16	1	1	0	None	None

Table 6.6: Instrument classifier used to evaluate the disentanglement performance of the proposed LDM. The number in parentheses indicates the number of groups used for group normalisation.

the 13 instrument classes using a public implementation.[4]

Results

The results are presented in Table 6.7, comparing three variants of the proposed LDM-based model: M0, M1, and M2. Model M0 reconstructs one source at a time, conditioned on its source-level latent, and iterates N_s times to generate the complete mixture. This serves as a proxy for MSI-DIS (Lin et al., 2021), albeit with differing architectures and disentanglement strategies to ensure comparability. Model M1 is the proposed set-conditioned diffusion model trained with pitch supervision via (6.8), while M2 is identical to M1 but without pitch supervision.

Despite their distinct training setups, all models support both "Single" (source-wise) and "Set" (set-wise) inference, as the DiT component can operate on variable-length sequences. In particular, although M0 is trained in the "Single" configuration, it generalises to "Set" mode at inference; the inverse is true for M1 and M2.

Model M1 achieves the best overall performance in the "Set" mode, which aligns with its training configuration. This supports the hypothesis that modelling inter-dependencies among sources improves mixture reconstruction (Mariani et al., 2024; Manilow et al., 2022). Its high instrument classification accuracy also suggests that the timbre latent space is both discriminative and structured, as visualised in Fig. 6.7 (left).

Interestingly, M0 performs competitively under the "Set" mode, despite being trained solely in the "Single" configuration. This reflects the flexibility of the DiT architecture in adapting across inference modes. Meanwhile, M2 demonstrates reasonable performance without pitch supervision, indicating that the SB layer acts as

[4] `https://github.com/gudgud96/frechet-audio-distance/tree/main`

	Instrument (↑)		Pitch (↑)		FAD (↓)	
	Set	Single	Set	Single	Set	Single
M0	89.94	**96.43**	94.37	**97.15**	2.27	**2.13**
M1	**97.49**	96.05	97.15	97.02	**2.10**	2.14
M2	82.83	80.16	**97.19**	97.03	2.14	2.14

Table 6.7: Disentanglement and audio quality results. Instrument and pitch classification accuracies (%) assess the effectiveness of the timbre and pitch latents, respectively. FAD evaluates the perceptual quality of the reconstructed audio. "Set" and "Single" refer to set-conditioned and single-source generation modes.

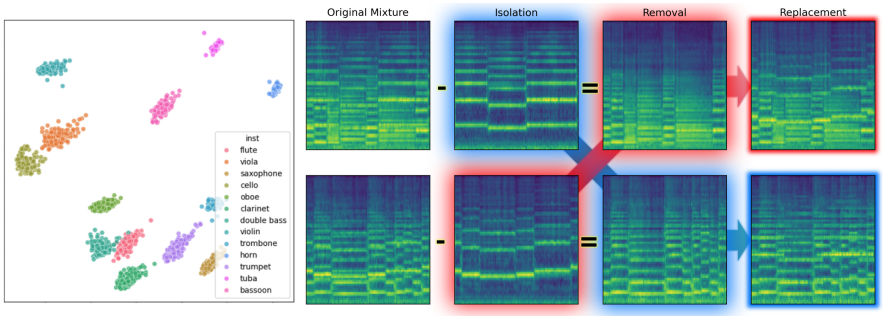

Figure 6.7: *Left*: PCA visualisation of the timbre latent space, showing clustering by instrument class. *Right*: Example of source-level swaps between two mixtures, demonstrating controllable pitch–timbre recombination.

an effective information bottleneck. Nevertheless, the lack of pitch guidance permits leakage of pitch information into the timbre latent, slightly degrading instrument classification accuracy.

Visualisation

While Fig. 6.6 provides an example of instrument swapping between two mixtures, Fig. 6.7 (right) swaps source-level latents between two mixtures and effectively exchanges their constituent instruments and melodies. Together, they illustrate the compositional nature of the learnt latent space.

6.7 Summary

This chapter introduces DisMix, a probabilistic framework for disentangling multi-instrument music mixtures into source-specific pitch and timbre representations (Section 6.3). DisMix enables modular and compositional editing of musical content by

manipulating these representations independently. Two instantiations of the framework are presented: a simple AE-based version for proof-of-concept on isolated chord mixtures (Section 6.5), and a more advanced implementation using an LDM to handle the complex four-part Bach chorales in the CocoChorale dataset (Section 6.6).

A key architectural insight is the use of a binarisation layer, which proves crucial for achieving effective pitch–timbre disentanglement. Additionally, results indicate that jointly reconstructing all sources in a mixture better captures source interdependencies than iterative single-source generation, leading to improved disentanglement and perceptual audio quality.

Despite these advances, DisMix relies on query-based conditioning for source extraction, which presumes access to examples closely matching the instrument identities or timbral properties of the target sources. To mitigate this reliance, future work may investigate clustering-based post-hoc query representations or adopt few-shot query relaxation methods (Y. Wang et al., 2022). Another promising direction is the integration of self-supervised learning frameworks for query-free source separation (Pasini et al., 2023) or unsupervised approaches such as MusicSlots (Gha et al., 2023), which may expand the framework's applicability without strong supervision.

Chapter 7

Conclusion

7.1 Thesis Summary

This thesis has focused on the development of robust and composable disentangled representations for sequential audio data, with applications to music and speech. Building on the framework of the disentangled sequential autoencoder (DSAE), training strategies and regularisation techniques have been proposed to improve robustness against model configurations to which unsupervised disentanglement is most sensitive: the expressivity of decoders, the size of the latent space, and the strength of the information bottleneck. Empirical evidence has been provided to support the proposed techniques, demonstrating their effectiveness in mitigating the persistent challenges of posterior collapse and addressing the trade-off between reconstruction fidelity and disentanglement.

To further illustrate the benefits of disentanglement for creative applications, this thesis also proposed scaling the complexity of pitch–timbre disentanglement beyond single-instrument sounds, enabling fine-grained manipulation and composition of musical attributes in an arbitrary manner. The contributions can be summarised in two aspects: *how to achieve robust disentanglement of sequential data without any explicit form of supervision* and *how to improve the generalisability of creative applications with disentangled representations*.

7.1.1 How to Achieve Robust Disentanglement of Sequential Data Without Explicit Forms of Supervision

The first part of this thesis addressed the lack of constraints in DSAE and its sensitivity to model configurations when learning from raw sequential data without supervision. Chapter 3 proposed TS-DSAE, which separates the optimisation of the global latent variable from the full reconstruction objective, thereby mitigating posterior collapse without the need for adversarial training or domain-specific augmentation. This staged approach proved robust across different decoder architectures and latent dimensionalities, and was empirically validated on both artificial and real-world datasets.

To further improve the balance between disentanglement and reconstruction quality, Chapter 4 introduced a constraint in the observation space rather than the latent space. By explicitly enforcing consistency on the Jacobian of the decoder output with respect to the latent variables between the first- and second-stage model components from TS-DSAE, the method maintained strong global–local disentanglement while reducing the perceptual degradation often associated with excessive regularisation. These two contributions established that robust, domain-agnostic disentanglement of global and local factors of variation can be achieved without any explicit form of supervision.

By corrupting the local latents with multiplicative Gaussian noise, Chapter 5 presented a simple yet effective strategy that dynamically balances the bottleneck between the global and local latent spaces, without introducing additional hyperparameters. Evaluated on the task of zero-shot voice conversion, the approach was shown to be effective across a range of hyperparameters and model architectures. Altogether, Chapters 3, 4, and 5 demonstrated complementary strategies to improve the robustness of unsupervised global–local disentanglement for both music and speech.

7.1.2 How to Improve the Generalisability of Creative Applications With Disentangled Representations

This thesis also demonstrated how disentangled representations can serve as modular and reusable building blocks for creative applications. Chapters 3, 4, and 6 showed generalisability to unseen combinations of pitch and timbre, while Chapter 5 demonstrated the ability to generalise to unseen speakers during training.

In particular, Chapter 6 extended the task of pitch–timbre disentanglement from single-source scenarios to mixtures containing multiple instruments in a supervised setup. This was achieved through a two-level factorisation into per-source pitch and timbre latents, enabling arbitrary within-source and cross-source attribute manipulation, thereby allowing for novel composition of mixtures.

7.2 Limitations

7.2.1 Experimental and Evaluation Limitations

The experimental evaluation presented in this thesis is primarily comparative in nature. As described in Section 3.4, no separate held-out test set is used, and performance is assessed on validation splits. While this limits the ability to draw conclusions about generalisation performance, the use of a consistent evaluation protocol enables meaningful comparison across models.

In addition, the evaluation relies on pitch extraction and classification procedures described in Section 3.5, and is therefore subject to the limitations discussed therein, including the dependence on pitch extraction accuracy and threshold-based classification metrics.

7.2.2 Theoretical Limitations

The methods proposed in this thesis primarily address posterior collapse as an optimisation issue. However, posterior collapse is closely related to the more fundamental problem of non-identifiability in latent-variable models, where multiple latent representations can explain the same observations. In this setting, a collapsed latent

corresponds to a degenerate but valid solution under the model.

While the proposed approaches improve the utilisation and stability of latent variables in practice, they do not resolve this underlying ambiguity and therefore do not provide guarantees of recovering true generative factors. A more detailed discussion of identifiability and potential approaches to addressing it is provided in Section 7.3.1.

7.2.3 Failure Modes and Practical Constraints

The effectiveness of the proposed methods relies on the assumption that the underlying factors of variation operate at different temporal scales. In scenarios where such structure is weak or absent, the distinction between global and local factors may become ill-defined. For example, for audio signals with little or no temporal variation (e.g., constant pitch throughout a sequence), factors such as pitch and instrument identity may both behave as time-invariant attributes, making their separation ambiguous.

More generally, the factors that are disentangled in unsupervised settings are determined by the statistical structure of the training data. Even when the model is designed to separate global and local factors, the semantics of these latents depend on the variability present in the dataset. For instance, if the training data contains only a single instrument, the global latent is unlikely to correspond to instrument identity, regardless of the model architecture.

In addition, the supervised components introduced in Chapter 6 rely on query-based conditioning, which assumes access to examples that reflect the target source characteristics, including instrument identity and acoustic conditions. As discussed in Section 6.7, this reliance limits applicability in scenarios where the characteristics of the available queries at inference time do not match those observed during training.

7.3 Future Work

Building on the empirical robustness and composability demonstrated in this thesis, several directions emerge for extending the proposed approaches. While the methods developed here improve the stability of unsupervised disentanglement and enable flex-

ible manipulation of latent factors, further work is needed to strengthen their theoretical foundations, broaden their applicability to more complex data distributions, and integrate them with modern generative modelling frameworks. In particular, questions of identifiability, factor dependence, and scalable generative modelling present natural avenues for future research.

7.3.1 Identifiability in Unsupervised Disentanglement

A key motivation for this work came from the finding of Locatello et al. (2019b) that, in the absence of suitable inductive biases, the unsupervised learning of disentangled representations is not identifiable: different parameterisations of the latent space can explain the same observations equally well, and the resulting models are highly sensitive to random seeds and hyperparameter settings. The training strategies in Chapter 3 and the Jacobian regularisation in Chapter 4 reduce this sensitivity in practice, leading to more stable factor separation. Yet, these approaches stop short of establishing formal identifiability guarantees. Bridging this gap would not only strengthen the scientific validity of the methods but also increase their transferability across domains and datasets.

The identifiability literature provides several promising pathways. Weakly supervised approaches such as Locatello et al. (2020a) demonstrate that minimal pairwise information—such as whether two sequences share certain latent attributes—can be sufficient for identifiability without compromising reconstruction quality. In the self-supervised setting, Kügelgen et al. (2021) show that carefully chosen data augmentations can provably separate content from style, framing the problem as an invariance–equivariance decomposition over different views of the same semantic object. This view-based reasoning extends naturally to sequential data: Simon et al. (2024) treat different frames of a sequence as distinct views of a shared time-invariant (static) factor while modelling the frame-specific variation as a time-varying (dynamic) factor conditioned on the static one. Their formulation provides sufficient conditions for identifiability in the sequential domain, directly aligning with the global–local factorisation used in Chapter 3 and Chapter 4. Establishing such theoretical guarantees for the empirical techniques proposed in this thesis would help ensure that the learnt

factors are not only robust in practice but also recoverable in principle.

7.3.2 Factor Correlation in Disentanglement

Real-world data often contain generative factors that are not statistically independent. Temporal sequences, in particular, exhibit correlations across the time axis (Klindt et al., 2021), and in pitch–timbre disentanglement (PTD) such correlations are inherent: each instrument has a constrained pitch range, and changes in pitch often influence timbre. The current models in this thesis do not consider these correlations and assume independent priors for factors. While this assumption simplifies optimisation and encourages factor separation, it neglects structured dependencies present in the data, which can limit the ability to extrapolate to out-of-distribution factor combinations.

When correlations between factors are ignored, disentanglement methods tend to produce latent representations in which the correlated factors remain statistically dependent, effectively reproducing the correlation structure of the training data in the latent space. This "mirroring" of dependencies reduces the capacity for independent factor manipulation and degrades performance under correlation shifts (Träuble et al., 2021; Funke et al., 2022). Methods that relax the independence assumption and explicitly model statistical dependencies between factors offer a way forward. For example, Roth et al. (2023) introduce the Hausdorff factorised support criterion, which allows arbitrary correlations within factors as long as their supports remain separable. Similarly, Kügelgen et al. (2021) provide a framework that does not require factors to be statistically independent, achieving separation through augmentations that selectively alter one factor while preserving another. Incorporating such approaches into the DSAE and DisMix frameworks could improve generalisation under realistic correlation structures, thereby enhancing the fidelity of pitch–timbre manipulation and other creative applications explored in this thesis.

7.3.3 Composable Representations for Generative Modelling

Modern generative models often operate on compact latent spaces rather than raw data, achieving scalability and controllability by focusing modelling capacity on se-

mantically meaningful aspects of the data. Examples include latent diffusion models, which condition the generation of VAE latents on explicit controls, and autoregressive models trained on quantised VQ-VAE (Oord et al., 2017) codes for next-token prediction. Such approaches not only reduce computational cost but also make the distribution easier to model by removing nuisance factors.

The disentangled and compositional representations developed in this thesis align naturally with this paradigm, serving as the output domain in which a generative model can operate. Rather than modelling raw audio waveforms, one could learn the distribution of these compact, semantically structured latents, benefiting from both their reduced dimensionality and their focus on salient factors of variation. Per-source factorisation, as in Chapter 6, enables this process to be carried out at the level of individual sources within a mixture. Y.-F. Wu et al. (2024) exemplify this idea in their work, in which a VQ-VAE first learns discrete, structured latents aligned with semantic concepts, and an autoregressive prior is then trained to model the distribution of these latents. This allows the generator to operate entirely in a compositional latent space rather than directly on the raw data. In the audio domain, Bindi et al. (2024) present an unsupervised framework for composable audio representations that encodes signals into a structured latent space before training a diffusion model to generate directly in that space. Their approach supports tasks such as source separation, variation generation, and unconditional sampling, all without reverting to waveform-domain modelling. Applying such latent-space generative modelling to the disentangled representations from this thesis could enable scalable synthesis with fine-grained control over multiple factors of variation.

7.4 Broader Impact

This thesis has been scoped to music and speech, but the challenges it addresses—stability in unsupervised disentanglement, separation of static and dynamic factors, and compositional control—are not unique to audio. Disentangled and object-centric representations are increasingly recognised as valuable across domains where temporal or structured data must be modelled. To illustrate this broader context, three

application areas are highlighted below.

7.4.1 Important Applications in Audio

Within the audio domain, disentangled representations have become central to modern codec design. Recent speech systems separate content, prosody, and speaker characteristics into distinct latent factors, which improves coding efficiency and allows controllable resynthesis of speech with different styles or voices (Zheng et al., 2025; Ren et al., 2024; Guo et al., 2025; Ju et al., 2024). Voice conversion likewise benefits from separating linguistic content from timbre to reduce speaker leakage (Mehta et al., 2025). In music, compact latent spaces are being developed both for compression and creative manipulation: some approaches introduce summary embeddings that capture global musical structure (Pasini et al., 2025), while others learn composable audio representations that support generation and source separation directly in latent space (Bindi et al., 2024). These examples underline the practical importance of temporally structured, factorised representations, reinforcing the broader relevance of the inductive biases studied in this thesis.

7.4.2 Disentangled Representations in Video

Video is a natural extension of this work, as it shares with audio the sequential evolution of data over time. Recent methods in video representation learning have shown that scenes can be decomposed into separate object representations and tracked across frames without supervision (Kipf et al., 2022; Singh et al., 2022; Elsayed et al., 2022; Zadaianchuk et al., 2023; Bae et al., 2024). Work in this area further suggests that explicitly separating stable object identity from time-varying motion leads to improved performance on tasks such as segmentation, tracking, and video understanding. Together, these studies highlight that the inductive biases of this thesis—distinguishing static from dynamic factors—are equally valuable for visual data.

7.4.3 Reinforcement Learning with Disentangled Representations

Reinforcement learning agents must also reason over temporally evolving states, and recent work has shown that disentangled, object-centric latents improve efficiency and generalisation. Pre-training such representations has been found to improve relational reasoning and sample efficiency in decision-making tasks (Yoon et al., 2023). Separating entities into interaction-relevant and irrelevant components has been shown to support more accurate multi-step planning (Nakano et al., 2023), while attribute-factored representations enable policies to adapt to new combinations of objects without retraining (Feng et al., 2023). Hierarchical abstractions over entities further support generalisation to novel rearrangement tasks (Chang et al., 2023). These advances suggest that the inductive biases explored in this thesis—stabilising the separation of global and local factors in time-series data—resonate with principles that benefit decision-making in temporally structured environments more broadly.

Bibliography

Acoustical Society of America (1994). *Acoustical Terminology*. ANSI S1.1-1994 (ASA 111-1994).

A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy (2017). "Deep Variational Information Bottleneck". In: *Proceedings of the International Conference on Learning Representations (ICLR)*.

A. A. Alemi, B. Poole, I. J. Fischer, J. V. Dillon, R. A. Saurous, and K. P. Murphy (2018). "Fixing a Broken ELBO". In: *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 159–168.

K. Bae, G. Ahn, Y. Kim, and J. Choi (2024). "DEVIAS: Learning Disentangled Video Representations of Action and Scene". In: *Proceedings of the European Conference on Computer Vision (ECCV)*.

J. Bai, W. Wang, and C. P. Gomes (2021). "Contrastively Disentangled Sequential Variational Autoencoder". In: *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pp. 10105–10118.

Y. Bengio, A. Courville, and P. Vincent (2013a). "Representation learning: A review and new perspectives". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8, pp. 1798–1828.

Y. Bengio, N. Léonard, and A. Courville (2013b). "Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation". In: *arXiv preprint arXiv:1308.3432*.

N. Berman, I. Naiman, I. Arbiv, G. Fadlon, and O. Azencot (2024). "Sequential Disentanglement by Extracting Static Information from a Single Sequence Element". In: *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

G. Bindi and P. Esling (2024). “Unsupervised Composable Representations for Audio”. In: *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 1037–1045.

A. Bitton, P. Esling, and A. Chemla-Romeu-Santos (2018). “Modulated Variational Auto-Encoders for Many-to-Many Musical Timbre Transfer”. In: *arXiv preprint arXiv:1810.00222*.

J. K. Bizley and Y. N. S. Cohen (Oct. 2013). “The what, where and how of auditory-object perception”. In: *Nature Reviews Neuroscience* 14.10, pp. 693–707. DOI: 10.1038/nrn3565.

N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent (2012). “Modeling Temporal Dependencies in High-Dimensional Sequences: Application to Polyphonic Music Generation and Transcription”. In: *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pp. 1881–1888.

S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio (2016). “Generating Sentences from a Continuous Space”. In: *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL)*. Berlin, Germany: Association for Computational Linguistics, pp. 10–21.

Y. Burda, R. B. Grosse, and R. Salakhutdinov (2016). “Importance Weighted Autoencoders”. In: *International Conference on Learning Representations (ICLR)*.

C. P. Burgess, R. Kabra, I. Higgins, L. Matthey, M. Botvinick, and C. Blundell (2019). “MONet: Unsupervised Scene Decomposition and Representation”. In: *arXiv preprint arXiv:1901.11390*.

M. Chang, A. L. Dayan, F. Meier, T. L. Griffiths, S. Levine, and A. Zhang (2023). “Neural Constraint Satisfaction: Hierarchical Abstraction for Combinatorial Generalization in Object Rearrangement”. In: *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*.

T. Q. Chen, X. Li, R. Grosse, and D. Duvenaud (2018). “Isolating sources of disentanglement in variational autoencoders”. In: *Advances in Neural Information Processing Systems*.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton (2020). "A Simple Framework for Contrastive Learning of Visual Representations". In: *Proceedings of the 37th International Conference on Machine Learning*. PMLR, pp. 1597–1607.

X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel (2017). "Variational lossy autoencoder". In: *International Conference on Learning Representations*.

Y.-H. Chen, D.-Y. Wu, T.-H. Wu, and H.-Y. Lee (2021). "AGAIN-VC: A One-Shot Voice Conversion Using Activation Guidance and Adaptive Instance Normalization". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5954–5958. DOI: 10.1109/ICASSP39728.2021.9414257.

P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin (2020). "CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information". In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*. PMLR, pp. 1779–1788.

K. W. Cheuk, K. Choi, Q. Kong, B. Li, M. Won, J.-C. Wang, Y.-N. Hung, and D. Herremans (2023). "Jointist: Simultaneous Improvement of Multi-instrument Transcription and Music Source Separation via Joint Training". In: *arXiv preprint arXiv:2302.00286*.

J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord (2019). "Unsupervised Speech Representation Learning Using WaveNet Autoencoders". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.12, pp. 2041–2053. DOI: 10.1109/TASLP.2019.2938863.

J.-c. Chou, C.-c. Yeh, H.-y. Lee, and L.-s. Lee (2019). "One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization". In: *Proceedings of Interspeech*. ISCA, pp. 664–668. DOI: 10.21437/Interspeech.2019-1936.

C.-Y. Chuang, J. W. Robinson, Y.-C. Lin, A. Torralba, and S. Jegelka (2020). "Debiased Contrastive Learning". In: *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pp. 8765–8775.

O. Cífka, A. Ozerov, U. Şimşekli, and G. Richard (2021). "Self-Supervised VQ-VAE for One-Shot Music Style Transfer". In: *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 306–310. DOI: 10.1109/ICASSP39728.2021.9414235.

T. M. Cover and J. A. Thomas (1991). *Elements of Information Theory*. New York: Wiley.

E. Creager, D. Madras, J.-H. Jacobsen, M. Weis, K. Swersky, T. Pitassi, and R. Zemel (2019). "Flexibly Fair Representation Learning by Disentanglement". In: *International Conference on Machine Learning*, pp. 1436–1445.

F. Cwitkowitz, K. W. Cheuk, W. Choi, M. A. Martínez-Ramírez, K. Toyama, W.-H. Liao, and Y. Mitsufuji (2024). "Timbre-Trap: A Low-Resource Framework for Instrument-Agnostic Music Transcription". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

N. Demerlé, P. Esling, G. Doras, and D. Genova (2024). "Combining Audio Control and Style Transfer Using Latent Diffusion". In: *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*. Milan, Italy.

N. Dilokthanakul, P. A. M. Mediano, M. Garnelo, M. C. H. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan (2016). "Deep Unsupervised Clustering with Gaussian Mixture Variational Autoencoders". In: *arXiv preprint arXiv:1611.02648*.

D. C. Dowson and B. V. Landau (1982). "The Fréchet distance between multivariate normal distributions". In: *Journal of Multivariate Analysis* 12.3, pp. 450–455. DOI: 10.1016/0047-259X(82)90077-X.

G. F. Elsayed, T. Kipf, S. van Steenkiste, F. Locatello, T. Rückstieß, G. Singh, K. Greff, J. Schmidhuber, and S. Mohamed (2022). "SAVi++: Towards End-to-End Object-Centric Learning from Real-World Videos". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi (2017). "Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 1068–1077.

P. Esling, A. Chemla-Romeu-Santos, and A. Bitton (2018). "Bridging Audio Analysis, Perception and Synthesis with Perceptually-Regularized Variational Timbre Spaces". In: *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*. Paris, France, pp. 175–181.

F. Feng and S. Magliacane (2023). "Learning Dynamic Attribute-Factored World Models for Efficient Multi-Object Reinforcement Learning". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

J. A. Fodor and Z. W. Pylyshyn (1988). "Connectionism and Cognitive Architecture: A Critical Analysis". In: *Cognition* 28.1-2, pp. 3–71. DOI: 10.1016/0010-0277(88)90031-5.

M. Fraccaro, S. K. Sønderby, U. Paquet, and O. Winther (2016). "Sequential Neural Models with Stochastic Layers". In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3989–3997.

C. M. Funke, P. Vicol, K.-C. Wang, M. Kümmerer, R. Zemel, and M. Bethge (2022). "Disentanglement and Generalization Under Correlation Shifts". In: *Proceedings of The 1st Conference on Lifelong Learning Agents (CoLLAs)*. Vol. 199. PMLR, pp. 116–141.

J. Gha, V. Herrmann, B. F. Grewe, J. Schmidhuber, and A. Gopalakrishnan (2023). "Unsupervised Musical Object Discovery from Audio". In: *NeurIPS Workshop on Machine Learning for Audio*.

L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber, and X. Alameda-Pineda (2021). "Dynamical Variational Autoencoders: A Comprehensive Review". In: *Foundations and Trends in Machine Learning* 15.1-2, pp. 1–175. DOI: 10.1561/2200000089.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014). "Generative Adversarial Networks". In: *Advances in Neural Information Processing Systems 27 (NeurIPS)*, pp. 2672–2680.

K. Greff, S. van Steenkiste, and J. Schmidhuber (2020). "On the Binding Problem in Artificial Neural Networks". In: *CoRR* abs/2012.05208. DOI: 10.48550/arXiv.2012.05208. arXiv: 2012.05208.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola (2007). "A Kernel Method for the Two-Sample-Problem". In: *Advances in Neural Information Processing Systems 19*. MIT Press, pp. 513–520.

A. Gretton, B. Sriperumbudur, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, and K. Fukumizu (2012). "Optimal Kernel Choice for Large-Scale Two-Sample Tests". In: *Advances in Neural Information Processing Systems 25 (NeurIPS)*. Curran Associates, Inc., pp. 1205–1213.

I. Gulrajani, K. Kumar, F. Ahmed, A. Taiga, F. Visin, D. Vazquez, and A. Courville (2017). "PixelVAE: A latent variable model for natural images". In: *International Conference on Learning Representations*.

Y. Guo, Y. Sun, Q. Li, X. Chen, X. Huang, and H. Zhao (2025). "LSCodec: Low-Bitrate and Speaker-Decoupled Discrete Speech Codec". In: *Proceedings of Interspeech 2025*. DOI: 10.21437/Interspeech.2025-1106.

M. Gutmann and A. Hyvärinen (2010). "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 9. Proceedings of Machine Learning Research. PMLR, pp. 297–304.

J. Han, M. R. Min, L. Han, L. E. Li, and X. Zhang (2021). "Disentangled Recurrent Wasserstein Autoencoder". In: *Proceedings of the 9th International Conference on Learning Representations (ICLR)*.

B. Hayes, C. Saitis, and G. Fazekas (2021). "Neural Waveshaping Synthesis". In: *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*. Online, pp. 254–261.

G. Heigold, I. Moreno, S. Bengio, and N. Shazeer (2016). "End-to-End Text-Dependent Speaker Verification". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5115–5119. DOI: 10.1109/ICASSP.2016.7472652.

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

I. Higgins, A. Pal, A. A. Rusu, L. Matthey, C. Burgess, A. Pritzel, M. Botvinick, C. Blundell, and A. Lerchner (2017b). "DARLA: Improving Zero-Shot Transfer in Reinforcement Learning". In: *International Conference on Machine Learning*.

J. Ho, A. Jain, and P. Abbeel (2020). "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6840–6851.

E. Hoffer and N. Ailon (2015). "Deep Metric Learning Using Triplet Network". In: *Proceedings of the International Conference on Learning Representations (ICLR) Workshop*.

M. D. Hoffman and M. J. Johnson (2016). "ELBO Surgery: Yet Another Way to Carve Up the Variational Evidence Lower Bound". In: *Proceedings of the NeurIPS 2016 Workshop on Advances in Approximate Bayesian Inference*.

W.-N. Hsu and J. Glass (2018a). "Extracting Domain Invariant Features by Unsupervised Learning for Robust Automatic Speech Recognition". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5614–5618. DOI: 10.1109/ICASSP.2018.8462037.

— (2018b). "Scalable Factorized Hierarchical Variational Autoencoder Training". In: *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1462–1466. DOI: 10.21437/Interspeech.2018-1034.

W.-N. Hsu, Y. Zhang, and J. Glass (2017). "Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data". In: *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 1878–1889.

W.-N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, P. Nguyen, and R. Pang (2019). "Hierarchical Generative Modeling for Controllable Speech Synthesis". In: *Proceedings of the International Conference on Learning Representations (ICLR)*.

T.-H. Huang, J.-H. Lin, and H.-Y. Lee (2021). "How Far Are We from Robust Voice Conversion: A Survey". In: *Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 514–521. DOI: 10.1109/SLT48900.2021.9383614.

W.-C. Huang, S.-W. Yang, T. Hayashi, and T. Toda (2022). "A Comparative Study of Self-supervised Speech Representation Based Voice Conversion". In: *IEEE Journal of Selected Topics in Signal Processing* 16.6, pp. 1216–1230. DOI: 10.1109/JSTSP.2022.3193761.

X. Huang and S. Belongie (2017). "Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, pp. 1501–1510. DOI: 10.1109/ICCV.2017.167.

Y.-N. Hung, I.-T. Chiang, Y.-A. Chen, and Y.-H. Yang (2019). "Musical Composition Style Transfer via Disentangled Timbre Representations". In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*. International Joint Conferences on Artificial Intelligence (IJCAI) Organization, pp. 4697–4703. DOI: 10.24963/ijcai.2019/652.

E. Jang, S. Gu, and B. Poole (2017). "Categorical Reparameterization with Gumbel-Softmax". In: *Proceedings of the International Conference on Learning Representations (ICLR)*.

Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu (2018). "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis". In: *Advances in Neural Information Processing Systems*. Vol. 31, pp. 4485–4495.

Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou (2017). "Variational Deep Embedding: An Unsupervised and Generative Approach to Clustering". In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*. International Joint Conferences on Artificial Intelligence Organization, pp. 1965–1972. DOI: 10.24963/ijcai.2017/273.

Z. Ju, Y. Ren, J. Liu, Y. He, and Z. Zhao (2024). "NaturalSpeech 3: Zero-Shot Speech Synthesis with Factorized Codec and Diffusion Models". In: *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235. Proceedings of Machine Learning Research, pp. 22605–22623.

K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi (2019). "Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms". In: *Pro-*

ceedings of the Annual Conference of the International Speech Communication Association (Interspeech), pp. 2350–2354. DOI: 10.21437/Interspeech.2019-2219.

H. Kim and A. Mnih (2018). "Disentangling by factorising". In: *Proceedings of the 35th International Conference on Machine Learning*, pp. 2649–2658.

J. W. Kim, R. Bittner, A. Kumar, and J. P. Bello (2019). "Neural Music Synthesis for Flexible Timbre Control". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 176–180. DOI: 10.1109/ICASSP.2019.8683596.

J. W. Kim, J. Salamon, P. Li, and J. P. Bello (2018). "CREPE: A Convolutional Representation for Pitch Estimation". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 161–165. DOI: 10.1109/ICASSP.2018.8461329.

D. P. Kingma and J. Ba (2015). "Adam: A Method for Stochastic Optimization". In: *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

D. P. Kingma and M. Welling (2014). "Auto-encoding variational Bayes". In: *International Conference on Learning Representations*.

T. Kipf, G. F. Elsayed, A. Mahendran, A. Stone, S. Sabour, G. Heigold, R. Jonschkowski, A. Dosovitskiy, and K. Greff (2022). "Conditional Object-Centric Learning from Video". In: *Proceedings of the Tenth International Conference on Learning Representations (ICLR)*.

D. A. Klindt, L. Schott, Y. Sharma, I. Ustyuzhaninov, W. Brendel, M. Bethge, and D. Paiton (2021). "Towards Nonlinear Disentanglement in Natural Data with Temporal Sparse Coding". In: *International Conference on Learning Representations (ICLR)*.

N. Kodali, J. Abernethy, J. Hays, and Z. Kira (2018). "On Convergence and Stability of GANs". In: *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

J. von Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello (2021). "Self-Supervised Learning with Data Augmentations Prov-

ably Isolates Content from Style". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 34, pp. 16443–16457.

A. Kumar, P. Sattigeri, and A. Balakrishnan (2018). "Variational inference of disentangled latent concepts from unlabeled observations". In: *International Conference on Learning Representations*.

J. H. Lee, H.-S. Choi, and K. Lee (2019). "Audio Query-based Music Source Separation". In: *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*. Delft, The Netherlands, pp. 878–885.

J. Lezama (2019). "Overcoming the Disentanglement vs Reconstruction Trade-Off via Jacobian Supervision". In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma (2019). "Creating a Multitrack Classical Music Performance Dataset for Multimodal Music Analysis: Challenges, Insights, and Applications". In: *IEEE Transactions on Multimedia* 21.2, pp. 522–535. DOI: 10.1109/TMM.2018.2856090.

Y. Li and S. Mandt (2018). "Disentangled Sequential Autoencoder". In: *Proceedings of the 35th International Conference on Machine Learning (ICML)*. PMLR, pp. 5670–5679.

Y. Li, K. Swersky, and R. Zemel (2015). "Generative Moment Matching Networks". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by F. Bach and D. Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, pp. 1718–1727.

Z. Li, J. V. Murkute, P. K. Gyawali, and L. Wang (2020). "Progressive Learning and Disentanglement of Hierarchical Representations". In: *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

J. Lian, C. Zhang, G. K. Anumanchipalli, and D. Yu (2022a). "Towards Improved Zero-shot Voice Conversion with Conditional DSVAE". In: *Proceedings of Interspeech 2022*. ISCA, pp. 2598–2602. DOI: 10.21437/Interspeech.2022-11225.

J. Lian, C. Zhang, and D. Yu (2022b). "Robust Disentangled Variational Speech Representation Learning for Zero-Shot Voice Conversion". In: *Proceedings of the IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP). DOI: 10.1109/ICASSP43922.2022.9747272.

L. Lin, Q. Kong, J. Jiang, and G. Xia (2021). "A Unified Model for Zero-shot Music Source Separation, Transcription and Synthesis". In: *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*. Online, pp. 385–392. DOI: 10.5281/zenodo.5690654.

A. H. Liu, T. Tu, H.-y. Lee, and L.-s. Lee (2020). "Towards Unsupervised Speech Recognition and Synthesis with Quantized Speech Representation Learning". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7254–7258. DOI: 10.1109/ICASSP40776.2020.9054564.

F. Liu, W. Xu, J. Lu, G. Zhang, A. Gretton, and D. J. Sutherland (2020). "Learning Deep Kernels for Non-Parametric Two-Sample Tests". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. Daumé III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 6316–6326.

H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley (2024). "AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining". In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 32, pp. 2871–2883. DOI: 10.1109/TASLP.2024.3399607.

S. Liu, O. Bousquet, and K. Chaudhuri (2017). "Approximation and Convergence Properties of Generative Adversarial Learning". In: *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 372–381.

X. Liu, D. Chin, Y. Huang, and G. Xia (2023). "Learning Interpretable Low-dimensional Representation via Physical Symmetry". In: *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*.

F. Locatello, G. Abbati, T. Rainforth, S. Bauer, B. Schölkopf, and O. Bachem (2019a). "On the Fairness of Disentangled Representations". In: *Advances in Neural Information Processing Systems* 32.

F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem (2019b). "Challenging Common Assumptions in the Unsupervised Learning of Dis-

entangled Representations". In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*. PMLR, pp. 4114–4124.

F. Locatello, B. Poole, G. Rätsch, B. Schölkopf, O. Bachem, and M. Tschannen (2020a). "Weakly-Supervised Disentanglement Without Compromises". In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*. Vol. 119. PMLR, pp. 6348–6359.

F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf (Dec. 2020b). "Object-Centric Learning with Slot Attention". In: *Advances in Neural Information Processing Systems 33 (NeurIPS)*. Virtual, pp. 11528–11539.

I. Loshchilov and F. Hutter (2017). "SGDR: Stochastic Gradient Descent with Warm Restarts". In: *International Conference on Learning Representations (ICLR)*.

H. Lu, D. Wang, X. Wu, Z. Wu, X. Liu, and H. Meng (2022). "Disentangled Speech Representation Learning for One-Shot Cross-Lingual Voice Conversion Using β-VAE". In: *Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 814–821. DOI: 10.1109/SLT54892.2023.10022787.

H. Lu, X. Wu, H. Guo, S. Liu, Z. Wu, and H. Meng (2024). "Unifying One-Shot Voice Conversion and Cloning with Disentangled Speech Representations". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 11141–11145. DOI: 10.1109/ICASSP48485.2024.10446296.

Y.-J. Luo, K. Agres, and D. Herremans (2019). "Learning Disentangled Representations of Timbre and Pitch for Musical Instrument Sounds Using Gaussian Mixture Variational Autoencoders". In: *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*. Delft, The Netherlands, pp. 746–753.

Y.-J. Luo, K. W. Cheuk, W. Choi, W.-H. Liao, K. Toyama, T. Uesaka, K. Saito, C.-H. Lai, Y. Takida, S. Dixon, and Y. Mitsufuji (2024a). "Disentangling Multi-instrument Music Audio for Source-level Pitch and Timbre Manipulation". In: *NeurIPS 2024 Workshop on Audio Imagination: AI-Driven Speech, Music, and Sound Generation*.

Y.-J. Luo, K. W. Cheuk, T. Nakano, M. Goto, and D. Herremans (2020). "Unsupervised Disentanglement of Pitch and Timbre for Isolated Musical Instrument Sounds". In: *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*. Montréal, Canada, pp. 700–707.

Y.-J. Luo and S. Dixon (2024b). "Posterior Variance-Parameterised Gaussian Dropout: Improving Disentangled Sequential Autoencoders for Zero-Shot Voice Conversion". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Y.-J. Luo, S. Ewert, and S. Dixon (2022). "Towards Robust Unsupervised Disentanglement of Sequential Data – A Case Study Using Music Audio". In: *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3299–3305. DOI: 10.24963/ijcai.2022/455.

— (2024c). "Unsupervised Pitch-Timbre Disentanglement of Musical Instruments Using a Jacobian Disentangled Sequential Autoencoder". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1036–1040. DOI: 10.1109/ICASSP48485.2024.10447564.

M. Luong and V. A. Tran (2021). "Many-to-Many Voice Conversion Based Feature Disentanglement Using Variational Autoencoder". In: *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 851–855. DOI: 10.21437/Interspeech.2021-2086.

C. J. Maddison, A. Mnih, and Y. W. Teh (2017). "The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables". In: *Proceedings of the International Conference on Learning Representations (ICLR)*.

A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey (2016). "Adversarial Autoencoders". In: *International Conference on Learning Representations (ICLR)*.

E. Manilow, C. Hawthorne, C.-Z. A. Huang, B. Pardo, and J. Engel (2022). "Improving Source Separation by Explicitly Modeling Dependencies Between Sources". In: *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 291–295. DOI: 10.1109/ICASSP43922.2022.9747076.

G. Mariani, I. Tallini, E. Postolache, M. Mancusi, L. Cosmo, and E. Rodolà (2024). "Multi-Source Diffusion Models for Simultaneous Music Generation and Separation". In: *Proceedings of the International Conference on Learning Representations (ICLR)*.

S. McAdams, B. Giordano, P. Susini, G. Peeters, and V. Rioux (2005). "A Meta-Analysis of Acoustic Correlates of Timbre Dimensions". In: *Proceedings of the 150th Meeting of the Acoustical Society of America*. Minneapolis, Minnesota.

S. McAdams and M. Goodchild (2017). "The Visual Representation of Timbre". In: *Organised Sound* 22.2, pp. 129–136. DOI: 10.1017/S1355771817000082.

S. Mehta, Y. Liu, Z. Tang, K. Peng, V. Manohar, S. Zhang, M. Seltzer, Q. He, and M. Ma (2025). "SemAlignVC: Enhancing Zero-Shot Timbre Conversion Using Semantic Alignment". In: *Proceedings of the 13th ISCA Speech Synthesis Workshop (SSW)*.

L. Mescheder, A. Geiger, and S. Nowozin (2018). "Which Training Methods for GANs do actually Converge?" In: *Proceedings of the 35th International Conference on Machine Learning*. PMLR, pp. 3481–3490.

N. Mor, L. Wolf, A. Polyak, and Y. Taigman (2019). "A Universal Music Translation Network". In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

I. Naiman, N. Berman, and O. Azencot (2023). "Sample and Predict Your Latent: Modality-free Sequential Disentanglement via Contrastive Estimation". In: *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

A. Nakano, M. Suzuki, and Y. Matsuo (2023). "Interaction-Based Disentanglement of Entities for Object-Centric World Models". In: *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*.

B. Nguyen and F. Cardinaux (2022). "NVC-Net: End-to-End Adversarial Voice Conversion". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6572–6576. DOI: 10.1109/ICASSP43922.2022.9747020.

B. van Niekerk, L. Nortje, and H. Kamper (2020a). "Vector-Quantized Neural Networks for Acoustic Unit Discovery in the ZeroSpeech 2020 Challenge". In: *Pro-*

ceedings of Interspeech 2020, pp. 4836–4840. DOI: 10.21437/Interspeech.2020-1693.

— (2020b). "Vector-Quantized Neural Networks for Acoustic Unit Discovery in the ZeroSpeech 2020 Challenge". In: *Interspeech*.

A. van den Oord, Y. Li, and O. Vinyals (2018). "Representation Learning with Contrastive Predictive Coding". In: *arXiv preprint arXiv:1807.03748*.

A. van den Oord, O. Vinyals, and K. Kavukcuoglu (2017). "Neural Discrete Representation Learning". In: *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 6306–6315.

A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu (2016). "WaveNet: A Generative Model for Raw Audio". In: *Proceedings of the 9th ISCA Workshop on Speech Synthesis (SSW 9)*. ISCA, p. 125.

J. D. Parker, J. Spijkervet, K. Kosta, F. Yesiler, B. Kuznetsov, J.-C. Wang, M. Avent, J. Chen, and D. Le (2024). "StemGen: A Music Generation Model That Listens". In: *Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. DOI: 10.1109/ICASSP48485.2024.10446088.

M. Pasini, S. Lattner, and G. Fazekas (2023). "Self-Supervised Music Source Separation Using Vector-Quantized Source Category Estimates". In: *NeurIPS 2023 Workshop on Machine Learning for Audio*.

— (2025). "Music2Latent2: Audio Compression with Summary Embeddings and Autoregressive Decoding". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

A. Pati, S. Gururani, and A. Lerch (2020). "dMelodies: A Music Dataset for Disentanglement Learning". In: *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*. Montréal, Canada, pp. 125–133.

W. Peebles and S. Xie (Oct. 2023). "Scalable Diffusion Models with Transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205. DOI: 10.1109/ICCV51070.2023.00387.

G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams (2011). "The Timbre Toolbox: Extracting Audio Descriptors from Musical Signals". In: *Journal*

of the Acoustical Society of America 130.5, pp. 2902–2916. DOI: 10.1121/1.3642604.

E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville (2018). "FiLM: Visual Reasoning with a General Conditioning Layer". In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3942–3951. DOI: 10.1609/AAAI.V32I1.11671.

K. Preechakul, N. Chatthee, S. Wizadwongsa, and S. Suwajanakorn (2022). "Diffusion Autoencoders: Toward a Meaningful and Decodable Representation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10619–10629. DOI: 10.1109/CVPR52688.2022.01055.

K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson (2019). "AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 5210–5219.

P. Reddy, S. Wisdom, K. Greff, J. R. Hershey, and T. Kipf (2023). "AudioSlots: A Slot-Centric Generative Model For Audio Separation". In: *arXiv preprint arXiv:2305.05591*.

Y. Ren, R. Huang, J. Liu, Z. Ye, X. Zhang, Z. Zhao, Z. Chen, C. Li, Z. Ma, Z. Xu, and Z. Zhu (2024). "TiCodec: Fewer-token Neural Speech Codec with Time-invariant Codes". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12737–12741.

M. Rey and A. Mnih (2021). "Gaussian Dropout as an Information Bottleneck Layer". In: *Bayesian Deep Learning Workshop at NeurIPS*.

D. J. Rezende, S. Mohamed, and D. Wierstra (2014). "Stochastic backpropagation and approximate inference in deep generative models". In: *Proceedings of the 31st International Conference on Machine Learning*, pp. 1278–1286.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer (2022). "High-Resolution Image Synthesis with Latent Diffusion Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695.

K. Roth, M. Ibrahim, Z. Akata, P. Vincent, and D. Bouchacourt (2023). "Disentanglement of Correlated Factors via Hausdorff Factorized Support". In: *International Conference on Learning Representations (ICLR)*. Poster.

K. Roth, A. Lucchi, S. Nowozin, and T. Hofmann (2017). "Stabilizing Training of Generative Adversarial Networks through Regularization". In: *Advances in Neural Information Processing Systems 30 (NeurIPS)*. Curran Associates, Inc., pp. 2018–2028.

J. Salamon, E. Gómez, D. P. Ellis, and G. Richard (2014). "Melody Extraction from Polyphonic Music Signals: Approaches, Applications, and Challenges". In: *IEEE Signal Processing Magazine* 31.2, pp. 118–134. DOI: 10.1109/MSP.2013.2271648.

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen (2016). "Improved Techniques for Training GANs". In: *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 2234–2242.

V. Saxena, J. Ba, and D. Hafner (2021). "Clockwork Variational Autoencoders". In: *Advances in Neural Information Processing Systems*. Vol. 34, pp. 20966–20978.

I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio (2017). "A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues". In: *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3295–3301. DOI: 10.1609/aaai.v31i1.10983.

M. C. Simon, P. Frossard, and C. De Vleeschouwer (2024). "Sequential Representation Learning via Static-Dynamic Conditional Disentanglement". In: *Proceedings of the European Conference on Computer Vision (ECCV)*.

G. Singh, Y. Kim, and S. Ahn (2023). "Neural Systematic Binder". In: *International Conference on Learning Representations (ICLR)*. Poster.

G. Singh, Y.-F. Wu, and S. Ahn (2022). "Simple Unsupervised Object-Centric Learning for Complex and Naturalistic Videos". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli (2015). "Deep Unsupervised Learning using Nonequilibrium Thermodynamics". In: *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 2256–2265. DOI: 10.5555/3045118.3045402.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole (2021). "Score-Based Generative Modeling through Stochastic Differential Equations". In: *International Conference on Learning Representations*.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014). "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15, pp. 1929–1958.

D. Stanton, M. Shannon, S. Mariooryad, R. Skerry-Ryan, E. Battenberg, T. Bagby, and D. Kao (2021). "Speaker Generation". In: *arXiv preprint arXiv:2111.05095*.

S. van Steenkiste, F. Locatello, J. Schmidhuber, and O. Bachem (2019). "Are Disentangled Representations Helpful for Abstract Visual Reasoning?" In: *Advances in Neural Information Processing Systems*.

Y. Takida, T. Shibuya, W.-H. Liao, C.-H. Lai, J. Ohmura, T. Uesaka, N. Murata, S. Takahashi, T. Kumakura, and Y. Mitsufuji (2022). "SQ-VAE: Variational Bayes on Discrete Representation with Self-annealed Stochastic Quantization". In: *Proceedings of the International Conference on Machine Learning (ICML)*.

K. Tanaka, R. Nishikimi, Y. Bando, K. Yoshii, and S. Morishima (2021). "Pitch-Timbre Disentanglement of Musical Instrument Sounds Based on VAE-Based Metric Learning". In: *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 111–115. DOI: 10.1109/ICASSP39728.2021.9413914.

K. Tanaka, K. Yoshii, S. Dixon, and S. Morishima (2022). "Unsupervised Disentanglement of Timbral, Pitch, and Variation Features from Musical Instrument Sounds with Random Perturbation". In: *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. APSIPA, pp. 709–716.

— (2025). "Unsupervised Pitch-Timbre-Variation Disentanglement of Monophonic Music Signals Based on Random Perturbation and Re-entry Training". In: *APSIPA Transactions on Signal and Information Processing* 14.1, e4. DOI: 10.1561/116.20240072.

H. Tang, X. Zhang, J. Wang, N. Cheng, and J. Xiao (2022). "AVQVC: One-Shot Voice Conversion by Vector Quantization with Applying Contrastive Learning".

In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 4613–4617. DOI: 10.1109/ICASSP43922.2022.9747020.

Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola (2020). "What Makes for Good Views for Contrastive Learning?" In: *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pp. 1–13.

N. Tishby, F. C. Pereira, and W. Bialek (1999). "The Information Bottleneck Method". In: *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pp. 368–377.

A. Tjandra, Y. Zhang, and S. Karita (2021). "Unsupervised Learning of Disentangled Speech Content and Style Representation". In: *Proceedings of Interspeech*. ISCA, pp. 3277–3281. DOI: 10.21437/Interspeech.2021-1936.

I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schölkopf (2018). "Wasserstein auto-encoders". In: *International Conference on Learning Representations*.

F. Träuble, E. Creager, N. Kilbertus, F. Locatello, A. Dittadi, A. Goyal, B. Schölkopf, and S. Bauer (2021). "On Disentangled Representations Learned from Correlated Data". In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. Vol. 139. PMLR, pp. 10401–10412.

M. Tschannen, O. Bachem, and M. Lucic (2018). "Recent advances in autoencoder-based representation learning". In: *arXiv preprint arXiv:1812.05069*.

E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez (2014). "Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4052–4056. DOI: 10.1109/ICASSP.2014.6854363.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017). "Attention Is All You Need". In: *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*. Long Beach, CA, USA, pp. 5998–6008.

C. Veaux, J. Yamagishi, and S. King (2013). "The Voice Bank Corpus: Design, Collection and Data Analysis of a Large Regional Accent Speech Database". In: *2013*

International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE). IEEE, pp. 1–4. DOI: 10.1109/ICSDA.2013.6709856.

W. Verhelst and M. Roelands (1993). "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech". In: *Proceedings of the 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 554–557. DOI: 10.1109/ICASSP.1993.319366.

B. Vowels, F. Herrmann, and C. Rother (2021). "VDSM: Unsupervised Video Disentanglement with State-Space Modeling and Deep Mixtures of Experts". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16668–16677.

M. Vowels, N. C. Camgoz, and R. Bowden (2020). "NestedVAE: Isolating Common Factors via Weak Supervision". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14181–14190. DOI: 10.1109/CVPR42600.2020.01420.

L. Wan, Q. Wang, A. Papir, and I. Lopez Moreno (2018). "Generalized End-to-End Loss for Speaker Verification". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 4879–4883. DOI: 10.1109/ICASSP.2018.8462665.

D. Wang, L. Deng, Y. T. Yeung, X. Chen, X. Liu, and H. Meng (2021). "VQMIVC: Vector Quantization and Mutual Information-Based Unsupervised Speech Representation Disentanglement for One-Shot Voice Conversion". In: *Proceedings of Interspeech*. ISCA, pp. 1614–1618. DOI: 10.21437/Interspeech.2021-1936.

S. Wang and C. D. Manning (2013). "Fast Dropout Training". In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by S. Dasgupta and D. McAllester. Vol. 28. Proceedings of Machine Learning Research 2. Atlanta, Georgia, USA: PMLR, pp. 118–126.

W. Wang, R. Arora, K. Livescu, and J. Bilmes (2016). "Deep Variational Canonical Correlation Analysis". In: *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 1088–1097.

X. Wang, H. Chen, S. Tang, Z. Wu, and W. Zhu (2024). "Disentangled Representation Learning". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Accepted.

Y. Wang, D. M. Blei, and J. P. Cunningham (2021). "Posterior Collapse and Latent Variable Non-identifiability". In: *Advances in Neural Information Processing Systems 35 (NeurIPS)*.

Y. Wang, D. Stoller, R. M. Bittner, and J. P. Bello (2022). "Few-Shot Musical Source Separation". In: *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 121–125. DOI: 10.1109/ICASSP43922.2022.9747536.

P. Wei, X. Yin, C. Wang, Z. Li, X. Qu, Z. Xu, and Z. Ma (2023). "S2CD: Self-heuristic Speaker Content Disentanglement for Any-to-Any Voice Conversion". In: *Proceedings of Interspeech 2023*. ISCA, pp. 2288–2292. DOI: 10.21437/Interspeech.2023-215.

Y.-F. Wu, M. Lee, and S. Ahn (2024). "Neural Language of Thought Models". In: *International Conference on Learning Representations (ICLR)*.

D.-Y. Wu and H.-y. Lee (2020). "One-Shot Voice Conversion by Vector Quantization". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7734–7738. DOI: 10.1109/ICASSP40776.2020.9053854.

Y. Wu, J. Gardner, E. Manilow, I. Simon, C. Hawthorne, and J. Engel (2022). "The Chamber Ensemble Generator: Limitless High-Quality MIR Data via Generative Modeling". In: *arXiv preprint arXiv:2209.14458*.

Y. Wu, Y. He, X. Liu, Y. Wang, and R. B. Dannenberg (2023). "TransPlayer: Timbre Style Transfer with Flexible Timbre Control". In: *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. DOI: 10.17023/81yw-jy50.

Y. Wu, Z. Wang, B. Raj, and G. Xia (2025). "Unsupervised Disentanglement of Content and Style via Variance-Invariance Constraints". In: *Proceedings of the 13th International Conference on Learning Representations (ICLR)*.

Z. Wu, Y. Rubanova, R. Kabra, D. A. Hudson, I. Gilitschenski, Y. Aytar, S. van Steenkiste, K. R. Allen, and T. Kipf (2024). "Neural Assets: 3D-Aware Multi-Object Scene Synthesis with Image Diffusion Models". In: *Advances in Neural Information Processing Systems (NeurIPS) 37*. NeurIPS 2024 Spotlight, Spotlight paper.

J. Yoon, Y.-F. Wu, H. Bae, and S. Ahn (2023). "An Investigation into Pre-Training Object-Centric Representations for Reinforcement Learning". In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research, pp. 40147–40174.

S. Yuan, P. Cheng, R. Zhang, W. Hao, Z. Gan, and L. Carin (2021). "Improving Zero-Shot Voice Style Transfer via Disentangled Representation Learning". In: *Proceedings of the International Conference on Learning Representations (ICLR)*.

A. Zadaianchuk, M. Seitzer, and G. Martius (2023). "Object-Centric Learning for Real-World Videos by Predicting Temporal Feature Similarity". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny (2021). "Barlow Twins: Self-Supervised Learning via Redundancy Reduction". In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. Vol. 139. PMLR, pp. 12310–12320.

J. Zhang and K. Ma (2022). "Rethinking the Augmentation Module in Contrastive Learning: Learning Hierarchical Augmentation Invariance with Expanded Views". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16649–16659. DOI: 10.1109.

Y. Zhao, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda (2020). "Voice Conversion Challenge 2020: Intra-lingual Semi-Parallel and Cross-Lingual Voice Conversion". In: *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, pp. 80–98. DOI: 10.21437/VCC_BC.2020-14.

Y. Zheng, T. Wei, Z. Zhang, X. He, S. Deng, Z. Xu, Y. Pan, Y. Sun, X. Chen, and D. Yu (2025). "FreeCodec: A Self-Supervised Disentangled Neural Speech Codec". In: *Proceedings of Interspeech 2025*. DOI: 10.21437/Interspeech.2025-2229.

Y. Zhu, M. R. Min, A. Kadav, and H. P. Graf (2020). "S3VAE: Self-Supervised Sequential VAE for Representation Disentanglement and Data Generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 6538–6547. DOI: 10.1109/CVPR42600.2020.00657.