Audiovisual speech representation learning applied to emotion recognition

Samir Sadok

COLLEGE DOCTORAL BRETAGNE
MATHS, TELECOMS
INFORMATIQUE, SIGNAL
SYSTEMES, ELECTRONIQUE

CentraleSupélec

THÈSE DE DOCTORAT DE

CENTRALESUPÉLEC

ÉCOLE DOCTORALE N° 601
Mathématiques, Télécommunications, Informatique, Signal, Systèmes, Électronique
Spécialité : *Signal, Image, Vision*

Par

Samir Sadok

Audiovisual speech representation learning applied to emotion recognition

Thèse présentée et soutenue à CentraleSupélec, le 08/03/2024
Unité de recherche : IETR UMR CNRS 6164
Thèse N° : 2024CSUP0003

Rapporteurs avant soutenance :

Slim Ouni	Professeur à l'Université de Lorraine - LORIA
Dominique Vaufreydaz	Professeur à l'Université Grenoble Alpes - LIG

Composition du Jury :

Président :	Céline Hudelot	Professeur à CentraleSupélec - MICS
Examinateurs :	Slim Essid	Professeur à Télécom Paris - IPP - LTCI
	Céline Hudelot	Professeur à CentraleSupélec - MICS
	Slim Ouni	Professeur à l'Université de Lorraine - LORIA
	Dominique Vaufreydaz	Professeur à l'Université Grenoble Alpes - LIG
Dir. de thèse :	Renaud Séguier	Professeur à CentraleSupélec - IETR
Encadrant de thèse :	Simon Leglaive	Maître de conférences à CentraleSupélec - IETR

Acknowledgement

Embarking on the journey of a doctoral thesis is an emotionally charged venture, leading you into the depths of a research laboratory where you become part of a dedicated team, fueled by the aspiration to make a meaningful scientific contribution. There are moments of joy throughout this adventure—like the elation of uncovering a solution after days of contemplation. Yet, it is also a path that demands resilience during challenging times, marked by uncertainties and occasional setbacks. Along this thesis journey, I have realized that there are no true failures; each setback becomes a valuable experience, propelling us toward continuous improvement. In the following paragraphs, I will make a few dedications to thank important individuals who have contributed greatly to the success of my thesis.

`To ...`

`My Family`

I deeply thank my family, particularly my parents, for their unwavering support and encouragement. Their constant belief in me has been a wellspring of strength and motivation throughout my accomplishments. I sincerely appreciate my brothers—Romain, Lamine, and Zakary—for their enduring encouragement and understanding during this challenging yet rewarding thesis journey. Their unwavering faith in me has provided consistent support, and their presence has brought immense joy and positivity to my life.

`My supervisors`

I also want to express my heartfelt appreciation to Simon & Renaud for their guidance and unwavering support. Simon, your presence and dedication have been instrumental, guiding me every step of the way with your expertise and availability. Renaud, your encouragement and confidence in my abilities have inspired me to strive for excellence. I am truly grateful for your mentorship and guidance throughout my doctoral journey.

Furthermore, I am grateful to Catherine for entrusting me with the opportunity to contribute to teaching at CentraleSupélec.

`My colleagues`

I cannot overlook the camaraderie of my colleagues, who have also become friends, accompanying me throughout these three years. Guénolé Fiche (Guénolé let's meet whenever you are available for a game of foosball), Adrien Llave, Paul Desombre, Jérémy Guillaume, Sen Yan, Amdjed Belaref, Ali, Gwendal Fouche, Neil Farmer, Matthieu Delmas, Marc Tournadre & George Konstantopoulos—your friendship has enriched this journey in countless ways. Your unwavering support, stimulating discussions, and shared experiences have been invaluable.

`Technical staff & Friends`

Lastly, I sincerely thank all those who supported me during my doctoral studies, including the administrative staff and technical support from centraleSupélec (Rennes). Your assistance has been invaluable. I also want to thank all my friends who stood by me throughout my thesis journey.

`Thanks a lot ···`

Table of Contents

Abstract

Emotions are vital in our daily lives, becoming a primary focus of ongoing research. Automatic emotion recognition has gained considerable attention owing to its wide-ranging applications across sectors such as healthcare, education, entertainment, and marketing. This advancement in emotion recognition is pivotal for fostering the development of human-centric artificial intelligence. Supervised emotion recognition systems have significantly improved over traditional machine learning approaches. However, this progress encounters limitations due to the complexity and ambiguous nature of emotions. Acquiring extensive emotionally labeled datasets is costly, time-intensive, and often impractical. Moreover, the subjective nature of emotions results in biased datasets, impacting the learning models' applicability in real-world scenarios. Motivated by how humans learn and conceptualize complex representations from an early age with minimal supervision, this approach demonstrates the effectiveness of leveraging prior experience to adapt to new situations. Unsupervised or self-supervised learning models draw inspiration from this paradigm. Initially, they aim to establish a general representation learning from unlabeled data, akin to the foundational prior experience in human learning. These representations should adhere to criteria like invariance, interpretability, and effectiveness. Subsequently, these learned representations are applied to downstream tasks with limited labeled data, such as emotion recognition. This mirrors the assimilation of new situations in human learning. In this thesis, we aim to propose unsupervised and self-supervised representation learning methods designed explicitly for multimodal and sequential data and to explore their potential advantages in the context of emotion recognition tasks. The main contributions of this thesis encompass: (i) *Developing* generative models via unsupervised or self-supervised learning for audiovisual speech representation learning, incorporating joint temporal and multimodal (audiovisual) modeling. (ii) *Structuring* the latent space to enable disentangled representations, enhancing interpretability by controlling human-interpretable latent factors. (iii) *Validating* the effectiveness of our approaches through both qualitative and quantitative analyses, in particular on emotion recognition task. Our methods facilitate signal *analysis, transformation, and generation*.

Chapter 1

Introduction

"In the beginning was emotion", by Louis-Ferdinand Céline.

Contents

Summary

Emotions are intricate phenomena that significantly impact our daily experiences. Automating their recognition stands as a pivotal domain in research, fostering the convergence of human interactions with technology across various fields like healthcare, education, and advertising. Over the past decade, supervised emotion recognition systems have made notable strides by learning emotional representations from annotated databases. Despite their remarkable accuracy compared to conventional methods, these systems face constraints that limit their practical real-world applicability. This chapter aims to contextualize my research thesis, analyze the constraints of supervised emotion recognition systems, and explore research avenues aimed at overcoming these limitations.

Figure 1.1 – An image generated using DALL.E-2 depicting a person in a stressful job interview situation.

Imagine yourself in a job interview or videoconference session with your potential future employers. You confront a question that leaves you perplexed. In an instant, you find yourself grappling with physical manifestations like sweating, an elevated heart rate, high blood pressure, hot flashes, and more. These physical responses can be accompanied by various emotions, such as anxiety or anger, making it challenging to remain focused and perform at your best during the interview. It is important to remember that these physiological and emotional responses are entirely normal. They are the natural result of your brain and body reacting to a potentially uncomfortable situation. These responses alert you that something might be wrong and motivate you to protect your well-being. However, in some situations, like a job interview, these reactions can interfere with your ability to communicate effectively and showcase your skills and qualifications.

Dealing with the emotional and physiological responses that arise in a stressful environment can be challenging. However, imagine accessing a tool capable of mitigating such responses in a videoconference or similar situation. This tool could be designed to propose appropriate interactions through any interface, considering your emotional states, the context, and the environment, allowing you to regain control. Like a digital coach, this tool

can help restore confidence by familiarizing the candidate with the stressful environment. For instance, this tool might suggest relaxation exercises or breathing techniques tailored to the individual's stress levels before the presentation. During the presentation, it could provide real-time feedback on pacing, tone, or body language, helping the candidate stay composed and engaged with the audience. By providing such assistance, the tool empowers job seekers to feel more confident during the interview process.

This tool is currently under development through a collaborative with Randstad, a recognized leader in human resources services. The project is a joint effort between Randstad and CentraleSupélec, which aims to revolutionize recruitment by encouraging equal opportunities for all candidates. This collaboration resulted in several research theses, including my own, which I will present in this manuscript.

1.1 Introduction to affective computing

To ensure the tool's effectiveness, accurate *emotion recognition* that could disrupt the interview process is essential. This is where my thesis plays a role, primarily focusing on recognizing and analyzing emotions. Specifically, my work falls within the field of *affective computing*, which encompasses the development of technologies that can accurately recognize, interpret, and respond to human emotional states.

Affective computing is an umbrella term for human emotions, feelings, and attitudes (Fleckenstein, 1991). The concept of affective computing initiated and proposed by R. Picard, 1997, has been guiding the development of computers in recognizing, expressing, and intelligently responding to human emotions. This thesis considers and focuses only on the *recognition* part. Automatic emotion recognition is an important aspect across diverse domains, profoundly influencing human communication and interactions. The applications and implications of this recognition are discussed in the following section.

Emotions, feelings, and attitudes are interconnected yet distinct psychological constructs that collectively shape human experiences and behaviors.

Emotion is the most fundamental of the three concepts. It refers to a complex, automatic response to a specific event, situation, or stimulus. Emotions are brief and intense, often triggered by external factors or internal thoughts (Izard, 1971). They are "universal" and biologically wired in humans, as well as in many animals. Common

emotions include happiness, sadness, anger, fear, disgust, and surprise (Ekman, 1973).

Feelings are the conscious experiences or subjective states that arise as a result of emotions. In other words, emotions are the underlying processes, while feelings are the conscious awareness of those processes. Feelings are more specific than emotions (i.e., unique to each individual).

Attitudes are more stable (days and longer) and enduring evaluations or predispositions toward various objects, ideas, people, or groups. They are shaped by a combination of emotions and feelings, as well as cognitive processes, such as beliefs, thoughts, and experiences (Van der Pligt et al., 1997).

Throughout the remainder of this manuscript, we will consistently employ the term **emotion** to remain consistent with the literature.

1.1.1 Applications of affective computing

In such a short time, affective computing has gained popularity due to its wide range of application domains (B. W. Schuller, 2018). In many practical scenarios, there is a need to develop intelligent systems that can accurately distinguish and understand people's emotions and provide appropriate and friendly responses quickly (Scheutz, 2012). In social media, affective computing can aid in understanding the opinions being expressed on different platforms (Balazs & Velásquez, 2016). In healthcare, affective computing can help diagnose and treat mental health conditions, monitor patients' emotions and stress levels, and improve communication between doctors and patients (Mano et al., 2016). In education, affective computing can create personalized learning experiences based on students' emotions, attention levels, and cognitive abilities (C.-H. Wu et al., 2016). In automotive, affective computing can improve driver safety and comfort by detecting drowsiness or stress and adjusting the driving environment accordingly. In entertainment, affective computing can be used in the gaming industry to create more immersive and engaging experiences by adapting game difficulty or storylines based on the player's emotional state. In marketing, affective computing can analyze consumer emotions and behavior and develop targeted advertising and product recommendations. Emotion recognition stands as the foundational pillar towards achieving this objective (Latif et al., 2021). By accurately discerning and understanding emotions, machines can navigate the complexity of human affect, paving the way for a future where technology harmoniously interacts with and responds to our emotional states. Thus, R. W. Picard et al., 2001 believe that affective computing is the key to promoting and advancing the development

of human-centric AI [1]. But How can an effective automatic emotion recognition system be developed? We will delve into this query in the subsequent section.

1.1.2 Emotion recognition system

According to Pantic et al., 2005, an ideal emotion recognition system should encompass five fundamental functionalities: *multimodality, robustness and accuracy, generality, sensitivity to dynamics, and contextual awareness.*

These attributes collectively shape the framework for an effective emotion recognition system. Firstly, being multimodal implies using multiple modalities, such as facial expressions, vocal intonations, and body language, enabling a more comprehensive understanding of emotions. Secondly, the system should exhibit robustness and accuracy, ensuring consistent and reliable performance across various conditions and individuals. Thirdly, it should be generic and capable of recognizing emotions across different cultures, contexts, and demographics. Additionally, sensitivity to dynamics is essential, as emotions are often transient and evolve. Finally, emotions are profoundly affected by situational factors and context, making it necessary to grasp the broader context in which they manifest. By encompassing these five qualities, an emotion recognition system can significantly enhance its effectiveness in real-world applications.

In the following sections, we will explore why prevalent emotion recognition systems, primarily reliant on *supervised learning*, currently struggle to satisfy the second and third outlined criteria.

1.2 Context of the thesis

To highlight the context of my thesis, I will briefly recall the emotion recognition pipeline. This latter provides a foundation for understanding the following sections.

Figure 1.2 illustrates the pipeline for emotion recognition using deep learning which can be summarized in four steps. (i) It begins with collecting and preprocessing diverse data, including audio, video, or text samples that exhibit various emotions. Typically, human experts gather labeled data using self-reporting, physiological sensors, or facial expression analysis. During data preprocessing, noise removal and standardization

1. Human-centric AI refers to the design, development, and deployment of artificial intelligence systems and technologies with a primary focus on benefiting and empowering humans.

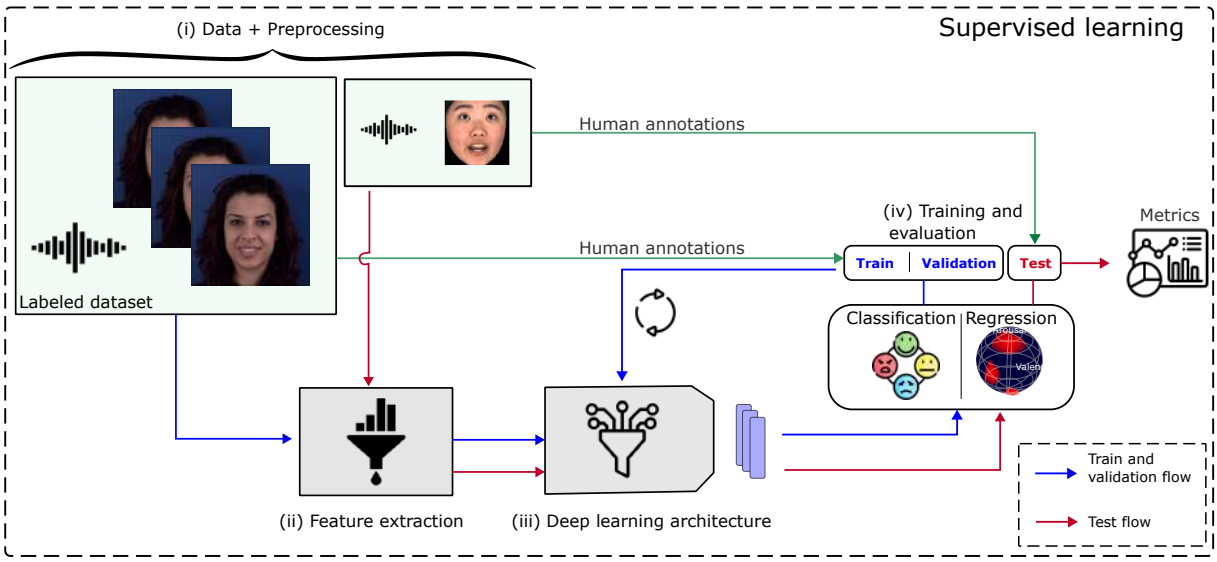


Figure 1.2 – An overview of a deep representation learning-based emotion recognition system.

are performed, (ii) The next step is feature extraction, where relevant features like Mel-Frequency Cepstral Coefficients (MFCCs) (Davis & Mermelstein, 1980) and pitch for audio and facial landmarks for video are extracted. For text data, conversion into numerical vectors through techniques like word embeddings or advanced methods like BERT embeddings (Devlin et al., 2019) is employed, (iii) Then comes the choice of the deep learning architecture: Convolutional Neural Networks (CNNs) (LeCun et al., 1989) may be preferred for images or spectrograms, while Recurrent Neural Networks (RNNs) (Elman, 1990) or Transformers (Vaswani et al., 2017) are suitable for sequential data like text or audio. In supervised learning, the training phase involves feeding the model with input features and corresponding labeled emotions, enabling it to adjust its internal parameters to minimize prediction errors, (iv) Finally, validation and hyperparameter tuning are performed to ensure model performance. Datasets are split into training and validation sets, and hyperparameters are adjusted based on validation metrics. Following training, model evaluation is conducted using a separate test dataset, assessing various metrics such as accuracy, precision, recall, and F1-score in the case of a classification task.

In affective computing, supervised learning is a common approach for emotion recognition across different data sources. While supervised learning has been quite successful in emotion recognition, it has limitations, as noted in many studies (Abate et al., 2023;

Y. Chen & Joo, 2021; Latif et al., 2022; Ouali, 2023; Sebe et al., 2005). A significant concern is the potential for unfair outcomes stemming from biases in both the *data* and the *model*. These points will be developed in Section 1.4.

To tackle these challenges, our studies explore alternatives to traditional supervised learning involving less supervised or unsupervised approaches, thus reducing the reliance on human annotations. These points will be developed in Section 1.5.

But before that, it is important to note that emotions are complex phenomena, often characterized by ambiguity and difficulties in their description (Tran et al., 2022). Different individuals might express similar emotions in varying ways, making it challenging to categorize or identify emotions accurately. The same emotion can manifest differently based on cultural, social, or personal contexts, leading to ambiguity in interpretation (Gendron et al., 2014). An introduction to the understanding of emotions and the inherent challenges in accurately representing them is introduced in the next section.

1.3 Ambiguity in the representation of emotions

To explain the idea of ambiguity in the representation of emotion, consider a scenario presented by Tran et al., 2022 within Shannon's communication model (Shannon, 1948). This model includes three key components: the source, the message, and the recipient. In machine communication, the source corresponds to the person expressing emotions, the message includes their verbal and non-verbal cues, and the recipient interprets this message. Emotion recognition in human communication involves complex challenges.

Ambiguity in emotions can be observed across all three stages of Shannon's model. First, it can stem from the source, as emotions are often complex and not easily understood, even by the person experiencing them. Second, ambiguity can be found in the message itself, which uses various channels (verbal and non-verbal) to convey emotions, making interpretation challenging. Third, ambiguity can occur at the recipient level, where people may interpret the same emotional message differently. This subjectivity in emotion perception is evident when annotators disagree on perceived emotions, especially when emotions are not strongly expressed.

Individuals rely on their judgment when interpreting emotions, giving varying importance to different communication signals. Perceived emotions can be seen as a mathematical combination of observed signals, each with a different weight influenced by personal experiences. For instance, one annotator might emphasize facial expressions more in their

emotional assessment, while another might focus on speech prosody.

In affective computing, researchers commonly employ categorical labels or numerical scales to represent emotions (Gunes & Schuller, 2013; Schroder et al., 2007). Categorical labels classify emotions into distinct categories, such as anger, happiness, sadness, and fear, providing a discrete representation of emotions. On the other hand, numerical scales continuously measure emotional dimensions, such as arousal and valence (see Figure 1.3). Arousal signifies the activation level linked to an emotion, spanning from low arousal or calmness to high arousal or excitement. Valence, on the other hand, captures the positive or negative nature of an emotion. Nevertheless, these approaches to emotion representation provide single-valued point estimates that cannot fully encapsulate emotions ambiguity, encompassing both their expression and perception (Sethu et al., 2019).

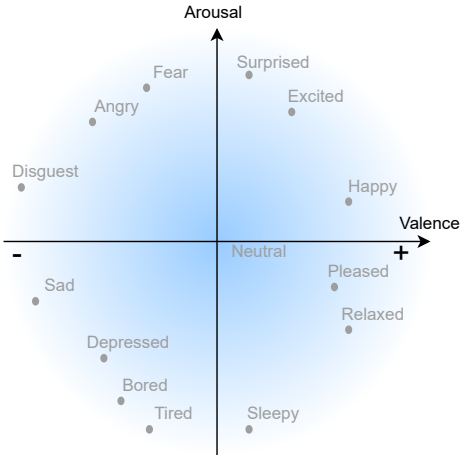


Figure 1.3 – Arousal represents the level of activation of an emotion, ranging from low to high. Conversely, valence represents the pleasantness or unpleasantness of emotion, ranging from negative to positive.

As explained in this section, the representation of emotions inherently carries ambiguity, which subsequently introduces biases across the entire supervised learning pipeline (Section 1.2), spanning from data acquisition to model validation. The following section will delve into an exploration of the limitations intrinsic to supervised learning.

1.4 Limitations of supervised methods

In terms of data Biases in data can lead to unfairness in learning tasks and take various forms. The literature on emotion recognition distinguishes three types of datasets (Abbaschian et al., 2021; Sebe et al., 2005): *simulated, semi-simulated, and natural emotion databases.*

Simulated datasets use recordings by actors in a controlled studio environment, making comparing results across different studies easy. However, the models trained on these datasets tend to overfit and may struggle to generalize to real-world conversations with

slightly different emotions. On the other hand, semi-simulated databases involve placing speakers in emotionally charged situations to evoke near-authentic emotions, based on scenarios that elicit emotions similar to natural utterances of speech. However, since these emotions are artificially created, speakers may express them differently than in real-life situations. Besides, the scenarios may not encompass all the emotions in natural conversations (Ververidis et al., 2004). Natural databases are mainly obtained from talk shows, call center recordings, radio talks, and similar sources and are sometimes called spontaneous speech. Nevertheless, collecting these datasets can be difficult due to ethical and legal concerns, and annotating them poses challenges; annotating each example in the database with a particular emotion can be contentious among annotators, resulting in variations and a lack of agreement in many instances (Y. Chen & Joo, 2021).

In terms of model Real-world datasets inherently contain biases due to the finite nature of the collected data. When we train models using standard supervised learning techniques, they tend to lack robustness to shifts in the data distribution (Vapnik, 1999). For instance, let us consider an emotion recognition model trained on facial expressions captured in a controlled studio environment characterized by well-regulated demographics and lighting conditions. Although such a model might excel on this specific training data (possibly overfitting), its performance could drastically deteriorate when applied to real-world scenarios, such as images or videos taken outdoors, under varying lighting conditions, or featuring diverse populations.

The model's ability to uphold its performance across different domains is known as out-of-distribution generalization (Shen et al., 2021). Out-of-distribution generalization, often called OOD generalization, is a critical facet of machine learning, including applications like emotion recognition. It refers to the model's capacity to make accurate predictions not just on the exact dataset it was trained on but also on novel, unseen data from various sources or domains.

1.5 Leveraging unlabeled data for emotion recognition: challenges

The limitations associated with supervised learning, coupled with the potential for biases in labeled datasets, emphasize the need to explore alternative approaches that require minimal or no supervision. Throughout my thesis, we have delved into other

training paradigms that aim to reduce the need for labels and address the aforementioned challenges.

1.5.1 Towards other learning paradigms for emotion recognition

In recent years, machine and deep learning have witnessed an increasing interest in exploring unsupervised and self-supervised learning paradigms. These methods present a promising alternative to the fully supervised learning approach, especially when labeled data is limited or prone to bias (Ouali, 2023). An illustrative example is that of autonomous driving. Consider the contrast between humans mastering driving within roughly 20 hours of practice with minimal guidance, while even advanced AI systems, trained on extensive human driving data, struggle with achieving fully autonomous driving. The crux lies in humans' reliance on their accumulated background knowledge of how the world operates.

Unsupervised learning involves training models on unlabeled data, intending to identify patterns and structures in the data without explicit guidance on what the model should look for, often referred to as *knowledge discovery* (Murphy, 2012). This can include techniques such as clustering, generative modeling, and dimensionality reduction (Ghahramani, 2003). *Self-supervised* learning involves training models on pretext tasks, such as predicting missing parts of an image or reconstructing a corrupted image, to learn a general representation that can be used for downstream tasks. By learning to solve these pretext tasks, the model can indirectly recognize patterns and structures in the data without explicit labels (Ericsson et al., 2022).

The following subsection explains why unsupervised and self-supervised learning can have significant advantages in affective computing.

1.5.2 Learning representations is (all) you need

Unsupervised and self-supervised methods usually involve a separate two-step learning process (Ericsson et al., 2022; Ghahramani, 2003; Obin, 2023). In the first step, the model learns to represent the data in an unsupervised or self-supervised way, through a pretext task that does not require labeling data. This approach enables the model to extract relevant features from the vast amount of unlabeled data available. In the second step, the model transfers the knowledge learned in the first step to an auxiliary task, such as emotion recognition, thereby using the learned representation for various downstream tasks, even with limited labeled data. In contrast, supervised methods combine these

two-step processes simultaneously in a single training session. This can result in the learned features being heavily reliant on the labeled input data, which may introduce potential biases that could affect the model's performance on unseen data.

The concept of leveraging information from the unsupervised or self-supervised phase to improve performance in the supervised learning stage has become more widely recognized. The basic idea is that some relevant features for unsupervised tasks may also be useful for supervised learning tasks. For instance, a generative model trained to produce images of emotional faces must identify the patterns that differentiate each emotion. If the model can accurately represent these patterns, the learned representation can be effectively used for supervised learning tasks. While there is to be a clear mathematical or theoretical understanding of this process, predicting which tasks will benefit from unsupervised learning is often difficult. Additionally, many aspects of this approach are model-dependent. For example, if we want to apply a linear classifier to the pre-trained features, the features must make the underlying classes linearly separable, which may only sometimes be the case. While these properties *often* occur naturally, they are not always guaranteed (Goodfellow et al., 2016).

Self-supervised and unsupervised learning for emotion recognition share a common goal: To learn a robust latent representation of the data that can extract relevant features (which satisfies some criteria we will discuss in the next chapter) without relying on explicit emotion labels. This latent representation can then be transferred to the task of emotion recognition, leading to more accurate and robust models, particularly in scenarios where labeled data is scarce or biased.

1.5.3 Towards multimodal learning for affective computing

Mehrabian, 2017 suggests that only 7% of feelings and emotions are conveyed through the words used in oral communication. In contrast, 38% are conveyed through tone and voice, and the remaining 55% are conveyed through facial expressions and body language. These findings, drawn from experiments related to the communication of feelings and attitudes (Mehrabian & Ferris, 1967; Mehrabian & Wiener, 1967), highlight the significance of incorporating nonverbal communication and multiple modalities to reduce the uncertainty in unimodal systems for accurate emotion recognition.

Besides improving interpretability, emotions are not experienced in isolation but are influenced by the surrounding context. The multimodal input allows for capturing the contextual factors that shape emotional expressions. Imagine a situation where someone

makes a sarcastic comment during a conversation. He/she says, "Wow, what a brilliant idea!" with a smirk on the face. In this case, relying solely on the statement's textual content may appear as a positive affirmation. However, we can detect the underlying sarcasm by considering the visual modality, precisely their facial expression. The smirk on their face, coupled with the sarcastic tone of their voice, provides additional cues that indicate their true intention. Combining the textual content, facial expression, and vocal tone forms a multimodal representation that enables us to identify the sarcasm or irony in the statement accurately. Another example is when someone is giving a presentation on a serious topic. He maintains a composed and serious facial expression during his speech while his tone remains calm and measured. However, their body language, such as tapping their foot or fidgeting with their hands, indicates nervousness or anxiety. In this scenario, analyzing only the audio speech might indicate that the person is delivering their message in a profound and controlled manner. However, incorporating the visual modality allows us to observe additional cues that reveal the underlying emotions. Despite the composed speech, their subtle signs of nervousness suggest an underlying sense of anxiety or discomfort. The emotion recognition system can capture the full range of expressed emotions by combining audio speech and visual modality.

Moreover, using multiple modalities offers a solution to the limitations associated with individual modalities (Abdullah et al., 2021; Sebe et al., 2005). Acoustic features may be vulnerable to background noise or speaker accents, potentially impacting emotion recognition accuracy. Similarly, the visual modality can be hindered by factors like lighting and background clutter, camera quality and angle, and the potential for movement and occlusions during data capture. The fusion of audio and visual modalities can significantly improve the accuracy and robustness of emotion recognition systems. Multimodal approaches effectively tackle the impact of noise or inconsistencies present in a single modality. This capability allows the models to handle real-world scenarios where emotions vary across individuals or challenging environments, making them more suitable for real-world deployments.

Overall, emotions are complex and subjective, leading to ambiguity in their representation, which is translated into the data and methods based on supervised learning. To address this issue, we seek to investigate whether integrating unsupervised or self-supervised representation learning methods, specifically when dealing with multimodal *and* sequential data, can improve emotion recognition.

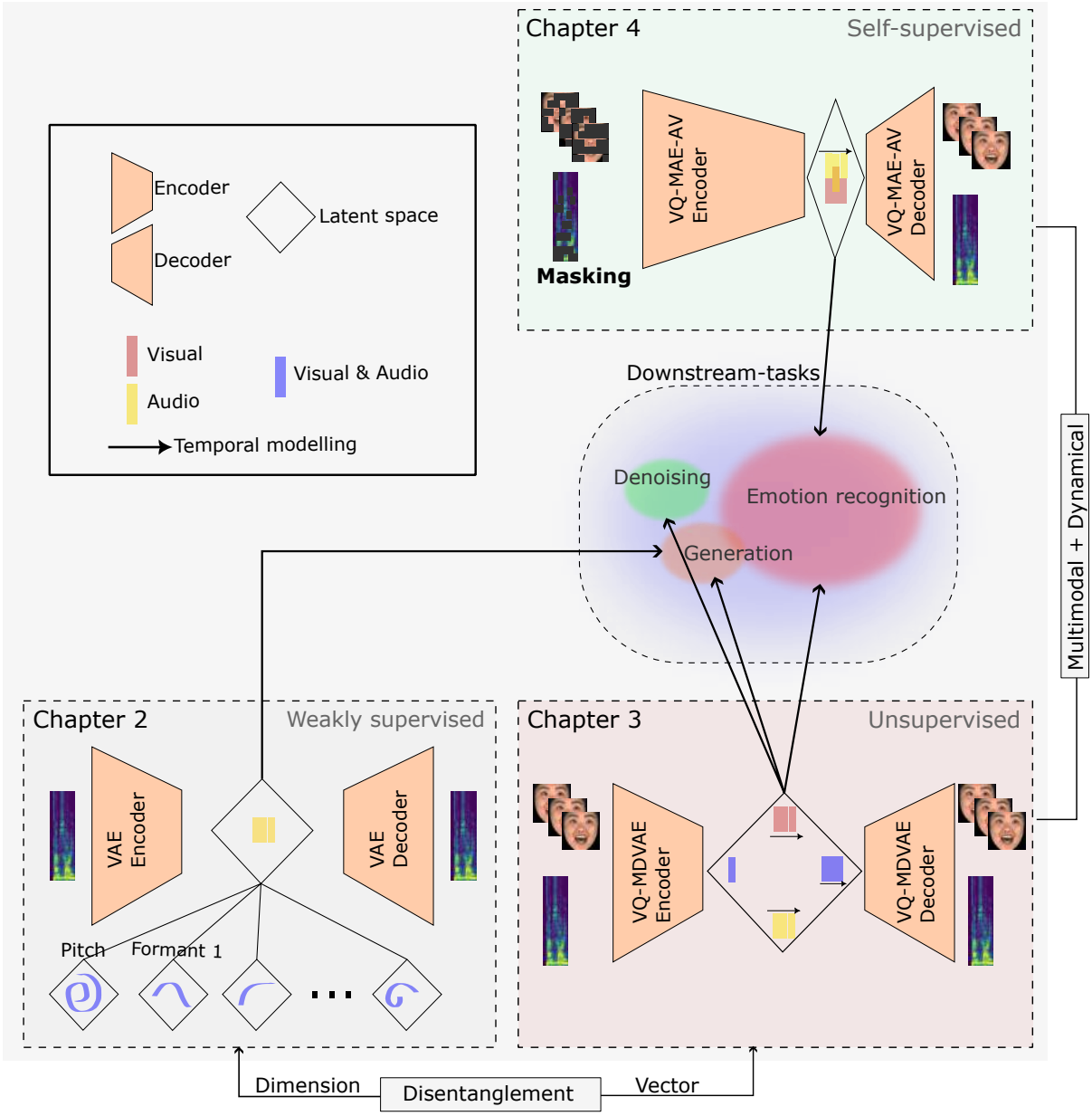


Figure 1.4 – This figure divides our contributions into 3 blocks, each of which is developed in the next chapters.

1.6 Thesis organization and contributions

Our contributions can be summarized in three main points as illustrated in Figure 1.4. These contributions and the organization of the manuscript are summarized in the following paragraphs:

Deep representation learning, Chapter 2, Page 27 We provide a review of the general state-of-the-art in deep representation learning, focusing on unsupervised and self-supervised approaches.

Learning and controlling the source-filter representation of speech with a variational autoencoder, Chapter 3, Page 75 We aim to improve the understanding and control of latent representations in deep generative models. We can analyze, transform, and generate diverse data types by exploring these representations. In speech processing, the source-filter model (Fant, 1970) is inspired by the physiological aspects of speech production, suggesting that speech signals arise from a few independent and meaningful latent factors. These factors include the fundamental frequency f_0 and formants, which play a crucial role in emotion recognition, particularly f_0. According to this model, speech is generated by the vibration of the vocal folds (source) and filtered by the vocal tract (filter) before being emitted. This work investigates the relationship between the source-filter model and the latent space of a variational autoencoder (VAE) (Kingma & Welling, 2014; Rezende et al., 2014) trained on a large dataset of unlabeled speech signals. We find that the VAE naturally captures the source-filter model as orthogonal subspaces within its latent space. We identify the subspaces corresponding to f_0 and the first three formant frequencies using a small amount of labeled data generated by an artificial speech synthesizer. These subspaces are orthogonal, allowing us to accurately and independently control the source-filter factors within them. Our method does not require additional information or human-labeled data, enabling the creation of a deep generative model for speech spectrograms conditioned on f_0 and formant frequencies. This model can effectively transform speech signals.

A multimodal and dynamical variational autoencoder, Chapter 4, Page 103 We present a multimodal and dynamical variational autoencoder (MDVAE) for unsupervised learning of audiovisual speech representations. The MDVAE's latent space is designed to separate dynamical factors shared across modalities from those specific to each modality. A static latent variable is introduced to capture constant information within an audiovisual speech sequence. The MDVAE is trained in two

stages on an audiovisual emotional speech dataset. In the first stage, separate vector quantized VAEs (VQ-VAEs) (Van den Oord et al., 2017) are trained for each modality without temporal modeling. In the second stage, the MDVAE is trained using the intermediate representation of the VQ-VAEs before quantization. In this second stage occurs the disentanglement of static versus dynamic and modality-specific versus modality-common information. The proposed approach is validated through extensive experiments involving resynthesis, transformation-synthesis, image denoising, and emotion recognition to demonstrate the effectiveness of the proposed model.

A vector quantized masked autoencoder, Chapter 5, Page 137 We build upon the growing interest in self-supervised learning methods, which offer promising solutions to the limitations of supervised learning. These approaches enable learning from vast amounts of unlabeled data, often readily available in various domains. In this context, we propose the VQ-MAE-AV model, a vector quantized masked autoencoder designed explicitly for self-supervised representation learning of audiovisual speech. Unlike existing multimodal MAEs that process raw audiovisual speech data, the VQ-MAE-AV model adopts a self-supervised paradigm based on discrete audio and visual speech representations learned by two pre-trained VQ-VAEs (Van den Oord et al., 2017). To evaluate the effectiveness of the proposed approach, the VQ-MAE-AV model is pre-trained on the VoxCeleb2 (Chung et al., 2018) database and fine-tuned on standard emotional audiovisual speech datasets. The experimental results demonstrate that the proposed method outperforms state-of-the-art audiovisual speech emotion recognition methods. These results underscore the potential of self-supervised learning approaches and showcase the efficacy of the VQ-MAE-AV model in learning robust and effective representations of audiovisual speech for emotion recognition.

Conclusion, ethical concerns and perspective, Chapter 6, Page 159 In this final chapter, we summarize the key contributions of this thesis and provide an overview of the findings. We reflect on the implications and significance of the research conducted, highlighting its potential impact in the field. Additionally, we discuss future perspectives and potential directions for further exploration and development, identifying areas for future research and potential applications of the findings.

PUBLICATIONS

1. Learning and controlling the source-filter representation of speech with a variational autoencoder.
   Sadok Samir, Leglaive Simon, Girin Laurent, Alameda-Pineda Xavier, & Séguier Renaud (2023).
   Speech Communication, 148, 53-65

2. A vector quantized masked autoencoder for speech emotion recognition.
   Sadok Samir, Leglaive Simon, & Séguier Renaud (2023).
   IEEE ICASSP 2023 Workshop on Self-Supervision in Audio, Speech and Beyond (SASB).

3. A multimodal dynamical variational autoencoder for audiovisual speech representation learning.
   Sadok Samir, Leglaive Simon, Girin Laurent, Alameda-Pineda Xavier, & Séguier Renaud (2023).
   In minor revision for Neural Networks journal.

4. A vector quantized masked autoencoder for audiovisual speech emotion recognition
   Sadok Samir, Leglaive Simon, & Séguier Renaud.
   Submitted to Face and Gesture (2024).

Chapter 2

Deep representation learning

Contents

It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without surrendering the adequate representation of a single datum of experience. Albert Einstein, in a 1933 lecture, "On the Method of Theoretical Physics,"

Summary

This chapter examines the current state-of-the-art in deep representation learning, encompassing various approaches, techniques, and paradigms. This chapter aims to provide an overview of the existing literature, emphasizing essential concepts and emerging trends in the field, focusing explicitly on unsupervised and self-supervised learning methods. This chapter establishes the groundwork for the following chapters, which will delve into more specific aspects of representation learning.

2.1 Introduction

According to Bengio, Courville, and Vincent, 2013, representation learning (RL) is defined as the process of "Learning representations of the data that facilitate the extraction of valuable information during the construction of classifiers or other predictive models". RL involves various methods that enable systems to automatically uncover the needed representations for feature detection or classification directly from raw data. The term *automatically* emphasizes the distinction between traditional feature engineering approaches and deep representation learning methods.

The process of manually designing a conversion of raw data into meaningful information using domain knowledge is referred to as *feature engineering*. In the context of emotion recognition (ER), as in many other applications, feature engineering and machine learning models for classification or regression are often treated as separate problems. Much ER research has focused on feature engineering pipelines to construct emotional representations that facilitate machine learning algorithms. While feature engineering techniques can improve ER performance, the downside is that they are labor-intensive, time-consuming and potentially sub-optimal (Latif et al., 2021). On the other hand, representation learning refers to the process of learning representations through the automatic transformation of input data using deep neural networks (DNNs) [1]. The primary objective of representation

1. DNNs consist of multiple layers of artificial neurons and are renowned for their ability to learn

learning is to generate abstract and meaningful representations that can be used for machine learning tasks like classification and regression (Bengio, Courville, & Vincent, 2013; Obin, 2023). Representation learning is a less time-consuming process requiring minimal human domain knowledge to produce better results than hand-engineered features (Latif et al., 2021).

Reducing dependency on engineered features is desirable to expand machine learning's applicability. By doing so, new applications can be developed more efficiently and effectively, leading to the advancement of artificial intelligence (AI). According to Bengio, Courville, and Vincent, 2013, AI system must deeply understand our world. This can only be accomplished if it can discern and distinguish the underlying factors that explain the environment it observes while relying on low-level sensory data.

In this chapter, we delve into the following key points:

— **Understanding representation learning, Section 2.2**: We introduce representation learning from an information theory perspective, allowing us to bring together various learning paradigms, including supervised, unsupervised, and self-supervised learning.

— **Effective representation learning, Section 2.3**: We thoroughly examine the criteria that a good learned representation should satisfy. Our exploration takes us into disentangled representation learning, where we delve into its formalization and explore its properties.

— **Understanding the generative variational autoencoder model, Section 2.4**: We delve into the generative variational autoencoder model, followed by a discussion on the methodologies for acquiring disentangled representations. Furthermore, we explore techniques for extending the generative model's capabilities to process sequential or multimodal data effectively.

— **Exploring the masked autoencoder, Section 2.5**: We introduce another generative model based on self-supervised learning, the masked autoencoder, where each part of the model is explained and analyzed.

Remark. The purpose of this chapter is to present methods for unsupervised and self-supervised representation learning while abstracting from the foundation of deep learning techniques. Therefore, this chapter does not cover basic deep learning techniques and assumes that the reader already has a background in deep learning.

complex patterns. A deep neural network can approximate any function through its learned parameters (universal approximation theorem (T. Kim & Adalı, 2003)).

2.2 Exploring representation learning through the lens of information theory

Before delving into the subsequent sections, let us introduce some key definitions that will be important for our discussion.

2.2.1 Notations and background

Notations In our discussions, we will use different symbols to represent different types of mathematical entities. We use x, $\mathbf{x}$, and $\mathbf{X}$ to denote scalars, vectors, and higher-dimensional tensors respectively. We use $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2 \ldots\}$ to denote the observed variables of the input domain, $\mathcal{Z} = \{\mathbf{z}_1, \mathbf{z}_2 \ldots\}$ for latent representations and $\mathcal{Y}$ for the output domain. For example, if we solve a classification task, $\mathcal{Y}$ is a set of discrete classes.

Background Mutual information (MI) measures the statistical dependence or information shared between two random variables. Let $(\mathbf{x}, \mathbf{y})$ be a pair of random variables. If their joint distribution is $p(\mathbf{x}, \mathbf{y})$ and the marginal distributions are $p(\mathbf{x})$ and $p(\mathbf{y})$, the mutual information is defined as

$$I(\mathbf{x}; \mathbf{y}) = D_{KL}\Big(p(\mathbf{x}, \mathbf{y}) \parallel p(\mathbf{x}) \cdot p(\mathbf{y})\Big), \tag{2.1}$$

where D_{KL} is the Kullback–Leibler divergence, which is a mathematical measure used to quantify the difference between two probability distributions. D_{KL} is formalized by the following equation $D_{KL}(p \parallel q) = \mathbb{E}_{p(\mathbf{x})}\left[\log\left(p(\mathbf{x})\right) - \log\left(q(\mathbf{x})\right)\right] \geq 0$.

MI quantifies the shared information or dependence between the two random variables. If $\mathbf{x}$ and $\mathbf{y}$ are independent, their MI is zero, indicating that knowing one variable provides no information about the other. Conversely, a higher MI value indicates a stronger dependence or shared information between the two variables.

The conditional mutual information between three random variables, $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$, with the expectation term, denoted as $I(\mathbf{x}; \mathbf{y}|\mathbf{z})$, is a measure of the amount of information shared between $\mathbf{x}$ and $\mathbf{y}$ while conditioning on the information provided

by $\mathbf{z}$. In mathematical terms, it is defined as:

$$I(\mathbf{x};\mathbf{y}|\mathbf{z}) = \mathbb{E}_{p(\mathbf{z})}\left[D_{KL}\Big(p(\mathbf{x},\mathbf{y}|\mathbf{z}) \parallel p(\mathbf{x}|\mathbf{z}) \cdot p(\mathbf{y}|\mathbf{z})\Big)\right], \tag{2.2}$$

where $\mathbb{E}[.]$ is the expectation operator.

2.2.2 Representation learning, information theory and information bottleneck

A concept underlies RL is that of "minimal sufficient statistics". To understand this concept, we need to decompose it:

Fisher, 1925 introduced the concept of a *sufficient statistic*. This means there are specific statistics that hold all the information we can gather from data related to a particular distribution. In other words, once we know the value of the sufficient statistic, the additional information provided by the remaining data does not add any new information about the estimated parameter. A theorem arises from the concept of sufficient statistics:

Theorem 2.2.1. *Let T be a probabilistic function of $\mathbf{x}$. Then, T is a sufficient statistic for a target variable $\mathbf{y}$ if and only if $I(T(\mathbf{x});\ \mathbf{y}) = I(\mathbf{x};\ \mathbf{y})$.*

This theorem states that a sufficient statistic captures all the information about $\mathbf{y}$ in $\mathbf{x}$. If this sufficient statistic effectively captures all the relevant information about $\mathbf{y}$ in the most concise or compact way possible, we refer to it as a *minimal sufficient statistic*.

This latter definition resonates with a fundamental concept in deep learning known as the *information bottleneck* (IB) (Tishby & Zaslavsky, 2015). The core idea behind IB is to extract valuable insights from observed signals linked to a target $\mathbf{y}$ by discovering a representation $\mathbf{z}$ that optimizes the quantity of information about $\mathbf{y}$ (enhanced performance) while minimizing the information required for representing $\mathbf{x}$ (maximal compression). This principle can be translated as the following optimization problem (Fischer, 2020; Shwartz-Ziv & Tishby, 2017):

$$\min_{\mathbf{z}} \Big\{ I(\mathbf{x};\mathbf{z}) - \beta I(\mathbf{z};\mathbf{y}) \Big\}. \tag{2.3}$$

The $I(\mathbf{z};\mathbf{y})$ measures the amount of target information $\mathbf{y}$ that is accessible through the compressed representation $\mathbf{z}$ and is an indicator of the model's ability to perform well on the task of interest. On the other hand, $I(\mathbf{x};\mathbf{z})$ represents the amount of information that $\mathbf{z}$ carries about the input $\mathbf{x}$, which is the information that we aim to compress. Therefore, there is a trade-off between compressing the representation and preserving the

relevant information about the target variable $\mathbf{y}$, and the hyperparameter β controls this compromise.

Let us analyse the relationship between the IB and the minimal sufficient statistic T. These two concepts are closely related. An efficient IB is characterized by the encoder, which can be any function T (possibly stochastic) that maps the observed data $\mathbf{x}$ to a compressed representation $\mathbf{z}$, and whose information content in $\mathbf{z}$ is maximized for a given task. Moreover, if $I(\mathbf{x}; \mathbf{y}|\mathbf{z}) = 0$, then we say that $\mathbf{z}$ satisfies the sufficiency property concerning the target variable $\mathbf{y}$. If this condition is satisfied, we can declare that the IB is the exact minimal sufficient statistic. However, if the condition is unsatisfied, the IB approximates the minimal sufficient statistic.

The concepts we have discussed so far have been based on the assumption of a single source of data $\mathbf{x}$. However, these concepts can be generalized to multiple sources of input $\{\mathbf{x^1}, \mathbf{x^2}, \mathbf{x^3} \ldots\}$, which is often referred to as multiview or multimodal representation learning. This extension is significant for self-supervised learning (SSL) but applies to both supervised and unsupervised learning paradigms.

2.2.3 Supervised representation learning

Definition Supervised representation learning involves the task of learning a meaningful representation $\mathbf{z} = T(\mathbf{x})$ of the input data $\mathbf{x}$ that facilitates accurate predictions of the corresponding labels y. In the context of the formalism described earlier in Equation 2.3, the variable y corresponds to the labels associated with the data, such as different emotions or categories. The goal is to find a representation that captures the relevant information in $\mathbf{x}$ and allows the model to make predictions about the labels y.

Connection to information theory Shwartz-Ziv and Tishby, 2017 present quantitative evidence that supervised learning models undergo two distinct phases: empirical error minimization (fitting phase) and representation compression (compression phase). The authors visualized the dynamic of training a neural network by plotting the values of $I(\mathbf{x};\ \mathbf{z})$ and $I(\mathbf{z};\ y)$ against each other (called the information plane). During the fitting phase, the model extracts information from the input and converts it into learned representations, leading to an increase in MI between inputs and hidden representations $I(\mathbf{x};\ \mathbf{z})$. In contrast, the compression phase focuses on discarding unnecessary information for target prediction, resulting in a decrease in MI between learned representations and inputs $I(\mathbf{x};\ \mathbf{z})$, while the MI between representations and targets $I(\mathbf{z};\ y)$ increases. The connection between the

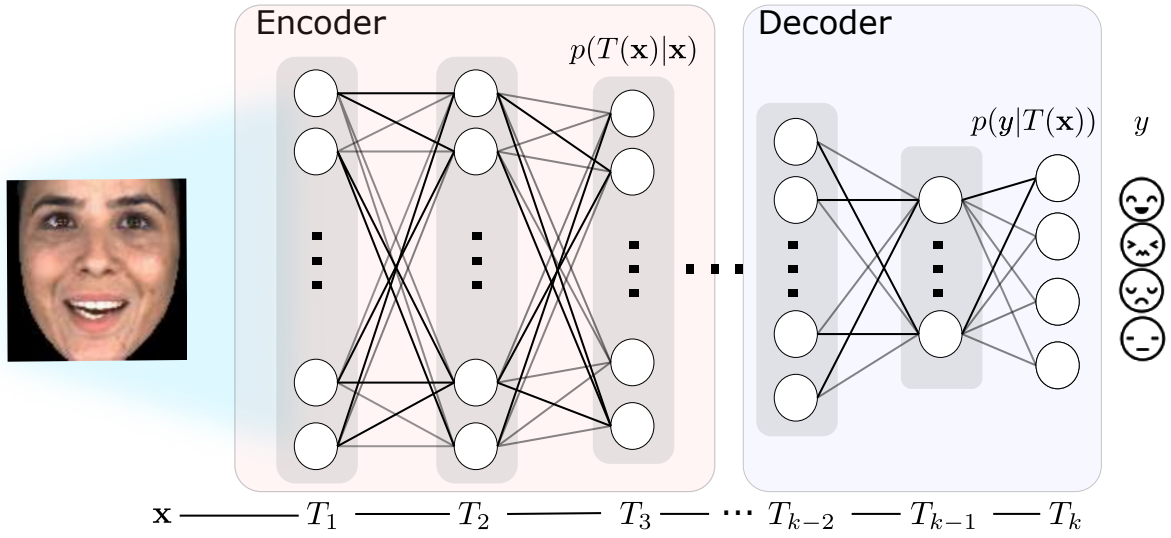


Figure 2.1 – Example of a feedforward DNNs with k hidden layers, an input layer $\mathbf{x}$, and an output layer $\hat{y}$. The target output, y, is available only during the learning phase and is obtained from a finite sample of the joint distribution, $p(\mathbf{x}, y)$.

compression phase and generalization has been emphasized by Shwartz-Ziv and Tishby, 2017, and subsequent studies have further supported their findings (Chelombiev et al., 2019; Shwartz-Ziv et al., 2018).

Based on this analysis, we can define the supervised approach with two components (see Figure 2.1): the encoder, which maps the raw data to a meaningful representation, and the decoder, which uses the encoded representation to make predictions or generate the desired categories y.

Suppose we consider the encoder and decoder as DNNs. In that case, we can harness the power of DNNs in learning hierarchical representations that benefit a wide range of machine learning tasks. In DNNs, each layer processes inputs solely from the preceding layer, forming a Markov chain within the network. This Markov property gives rise to the data processing inequality, which states that each subsequent layer in the network can only access and extract information from the input data equal to or less than the information processed by the preceding layer. In other words, as we progress deeper into the network, the representation becomes increasingly compressed and abstract, capturing the most salient and relevant features of the input (Tishby & Zaslavsky, 2015):

$$I(\mathbf{x};\ y) \geq I(T_1(\mathbf{x});\ y) \geq I(T_2(\mathbf{x});\ y) \geq \cdots \geq I(T_k(\mathbf{x});\ y) \geq I(\hat{y};\ y). \tag{2.4}$$

2.2.4 Unsupervised representation learning

Definition Unlike supervised representation learning, where the variable y is the target label, unsupervised representation learning takes a different approach by replacing the target labels with the reconstruction performance of the input. By replacing the labels with the unlabeled input data, unsupervised representation learning becomes a special case of supervised representation learning. The focus shifts from predicting specific labels to reconstructing or capturing the input data's properties. This allows the model to extract relevant features and representations that can later be used for various downstream tasks such as classification, clustering, or anomaly detection.

Connection to information theory The information theory perspective of unsupervised learning is characterized by two components: $I(\mathbf{x};\mathbf{z})$, which represents the MI resulted encoded by the *encoder*, and $I(\mathbf{z}; y = \hat{\mathbf{x}})$, which represents the MI resulted by the *decoder*. This definition encapsulates a paradoxical nature that arises from the inherent trade-off between two conflicting objectives. On the one hand, our goal is to minimize $I(\mathbf{x};\mathbf{z})$. Doing so encourages the representation to capture only the most discriminative information, effectively filtering out irrelevant or redundant details. This process promotes the discovery of compact representations that generalize well to unseen data. On the other hand, we also desire the representation $\mathbf{z}$ to retain sufficient information about the original input $\mathbf{x}$ to enable accurate reconstruction $\hat{\mathbf{x}}$. By preserving this information, the decoder can effectively reconstruct the input from the learned representation, facilitating tasks that rely on faithful data reconstruction or generation.

In the unsupervised learning paradigm, the concept of compression assumes a nuanced perspective (Voloshynovskiy et al., 2020). In the supervised case, the latent variable $\mathbf{z}$ is a sufficient statistic for the target variable y, leading to a lower entropy [2] than the input variable $\mathbf{x}$. This reduction in entropy is a consequence of the compression phase. However, in the unsupervised setting, the IB principle suggests a different approach to compression. Here, the objective is to encode the input into a compressed representation that allows for unique decoding of each input sequence. In this scenario, the entropy of the latent space needs to match the entropy of the input space, making compression considerably more challenging. Unlike the supervised case, where the latent space can capture the essential information required for prediction, the unsupervised setting requires the latent space

2. The entropy of a random variable is the average level of information or uncertainty inherent to the variable's possible outcomes. The entropy is expressed as $H(\mathbf{x}) = \mathbb{E}_{p(\mathbf{x})}[-\log p(\mathbf{x})]$.

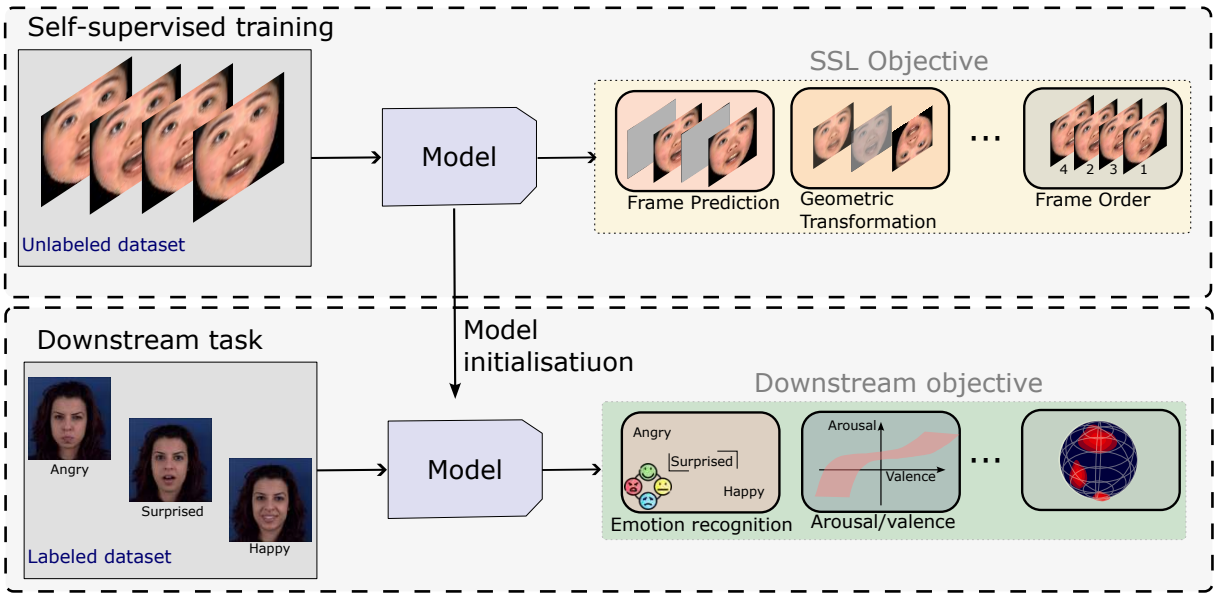


Figure 2.2 – The workflow initiates with the pre-training of a model on an unlabeled dataset, using a Self-supervised Learning (SSL) objective. Following this, the acquired parameters serve as the starting point for the model configuration in a downstream task involving a smaller labeled dataset.

to capture all the information necessary for faithful reconstruction. This places a higher burden on the compression process, as it necessitates encoding the input data in a way that preserves its essential characteristics without discarding too much important information.

2.2.5 Self-supervised representation learning

Definition Piaget's theory of cognitive development explains how children acquire knowledge through sensory and motor experiences, from birth until around 18 months (Piaget, 2000). This sensorimotor stage involves basic actions like sucking, grasping, looking, and listening, leading to the emergence of early representational thought. As children progress through different developmental stages, their reasoning abilities advance toward abstract thinking and deductive logic.

Learning is most effective when complex structures are built upon simpler ones, promoting natural development rather than relying solely on external reinforcement. Education aims to create dynamic structures that enable generalization, allowing learned knowledge and skills to be applied in different contexts (transfer of learning).

Like cognitive development, self-supervised learning aims to create models capable of

generating universal representations. SSL originated in robotics, where training data is labeled using relationships between input sensor signals. More formally, in contrast to supervised learning that depends on labeled data, SSL defines pretext tasks based on unlabeled inputs to generate descriptive representations. These pretext tasks serve as a proxy for the actual task of interest, allowing the model to learn meaningful representations. Pretext tasks can take various forms depending on the type of data and the specific learning objectives. In computer vision, pretext tasks, as shown in Figure 2.2, encompass various challenges, including adjusting image colors, geometric transformations, solving jigsaw puzzles, ordering frames in videos, and predicting future data. These tasks train models to understand images independently of color, recognize global-to-local view changes, reconstruct images from scrambled patches, establish temporal connections in videos, and capture high-level temporal patterns. In natural language processing, pretext tasks involve language modeling, word prediction, and coherent sentence generation.

Connection to information theory Integrating SSL into the information theory framework is a complex challenge requiring substantial research efforts (Garrido et al., 2022; W. Huang et al., 2021; Shwartz-Ziv & LeCun, 2023). Numerous studies have endeavored to bridge this gap (Bachman et al., 2019; Hjelm et al., 2018). For instance, Dubois et al., 2021 conducted a theoretical analysis of the IB principle in the field of SSL, focusing on determining the minimum bit rate required to store the input while achieving high performance on downstream tasks. This problem can be framed as a trade-off between rate and distortion, aiming to find a compressed representation that yields accurate predictions for each task. To ensure bounded distortion, they impose a condition that the difference between the conditional entropy of $\mathbf{y}$ given $\mathbf{z}$ and $\mathbf{y}$ given $\mathbf{x}$ remains below a specified threshold δ.

Other methods in this field rely on what is called the *multiview assumption* (Sridharan & Kakade, 2008). Which considers the case when we have several input sources (e.g., $\mathbf{x}_1, \mathbf{x}_2$) and their respective representations (e.g., $\mathbf{z}_1, \mathbf{z}_2$).

Assumption 2.2.2 (the multiview assumption)**.** *There exists an* $\epsilon_{info} \geq 0$ *such that:* $I(\mathbf{y}; \mathbf{x}_2|\mathbf{x}_1) \leq \epsilon_{info}$ and $I(\mathbf{y}; \mathbf{x}_1|\mathbf{x}_2) \leq \epsilon_{info}$.

An intuitive understanding of this assumption suggests that, on average, if we already know $\mathbf{x}_1$, learning $\mathbf{x}_2$ would not significantly enhance our understanding of $\mathbf{y}$ (and vice versa). This marginal potential gain is ϵ_{info}. In simpler terms, this assumption implies that both $\mathbf{x}_1$ and $\mathbf{x}_2$ are (approximately) carrying redundant information about $\mathbf{y}$. Embracing

this assumption allows us to distinguish between what is essential and not in the information. Using this insight, we can selectively compress unimportant details that do not contribute to subsequent tasks. The goal is twofold: we aim to maximize the relevant information captured by $I(\mathbf{x}_2; \mathbf{z}_1)$ while minimizing the information $I(\mathbf{x}_1; \mathbf{z}_1 | \mathbf{x}_2)$. This information should not be more useful for the task and can thus be discarded.

To illustrate the mildness of this assumption, consider self-supervised visual contrastive learning (Hjelm et al., 2018; Tsai et al., 2018), where the input and the self-supervised signal are the same images with different augmentations. This can be associated with altering the *style* of an image while preserving its *content*. The information required for downstream tasks should be retained in the content, not the style. However, the assumption may falter in cases of failure (or when ϵ_{info} is large): when the input and the self-supervised signal contain vastly different task-relevant information. For instance, drastic image augmentation (e.g., adding substantial noise) could alter the image's content (e.g., the noise completely occludes the objects).

2.3 Learning good representation

We have previously defined representation learning, *but what makes one representation better than another?* Defining an effective representation remains challenging. For example, a valid representation in probabilistic models captures the posterior distribution of explanatory factors under the observed input (Bengio, Courville, & Vincent, 2013). Others suggest that an ideal representation disentangles the underlying factors of variation that generated the data (Goodfellow et al., 2016). Before delving into this matter, let us establish the criteria for effective representation learning.

2.3.1 Challenges of representation learning

Invariance and equivariance To illustrate the concept of invariance and equivariance, let us consider the task of emotion recognition from audio data. In this task, we aim to build a model to accurately recognize emotions conveyed in speech data ($\mathbf{x} \in \mathcal{X}$). These audio samples may come from different accents, languages, and recording qualities. We aim to learn a representation ($\mathbf{z} \in \mathcal{Z}$) that suits this task. Ideally, if the emotional content in the audio changes, we want our representation to adapt accordingly; our representation should be emotion-equivariant. Equivariance means that if a transformation affects the

input $\mathbf{x}$, the output representation $\mathbf{z}$ is affected similarly. However, we also desire our representation to be invariant to variations caused by factors like background noise or recording conditions.

Generating factors In the context of a data distribution representing the domain $\mathcal{X}$, the generating factors $\mathcal{S}$ are the fundamental variables that comprehensively capture the data's variability, whether it is observed or not. Recent research suggests that representations should aid in the *decomposition* or *disentanglement* of input data into distinct factors (Bengio, Courville, & Vincent, 2013; Schölkopf et al., 2021). These factors should correspond to significant variables involved in the underlying data-generation process. Attempting to enumerate and isolate every possible combination of these factors in a dataset would be impractical due to the sheer number of potential variations.

Out-of-distribution generalization Considering an independent and identically distributed (i.i.d.) data distribution is a convenient yet often overly simplistic assumption. Real-world datasets inherently carry biases due to the finite and context-specific nature of the data they represent. When we train learning algorithms using standard supervised learning approaches, without accounting for these biases, the resulting models tend to struggle when confronted with changes in the data distribution (see more details in Section 1.4).

To illustrate this in the context of multimodal ER, consider scenarios where we collect speech and visual data from various sources. Each source might introduce biases related to varying recording equipment, speaking styles, facial expressions, or lighting conditions. These biases could affect the model's ability to consistently recognize emotions across these different sources. However, understanding the data generation process and the underlying relationships between variables can help us mitigate these biases. We can develop more robust multimodal ER systems by explicitly modeling these domain shifts and defining the changes we want our model to be invariant or equivariant.

2.3.2 Disentangled representation

Disentangled representations offer a promising solution to the challenges we have discussed. These representations aim to learn features that remain unaffected or exhibit well-defined changes for some data transformations (invariant/equivariant). They achieve this by considering the data generation process and potential domain shifts. While there

is no universally accepted definition for disentangled representations, the fundamental concept revolves around separating the primary factors that drive variations in our data distribution.

Generic definition Bengio, Courville, and Vincent, 2013 propose an intuitive definition of a disentangled representation. Disentangled representation refers to the capacity of a data representation system to separate the different underlying generative factors of variation within the data. In such a representation, individual latent variables or components correspond to specific and distinct generative factors, and they are sensitive to changes in one particular factor while remaining relatively unaffected by changes in other factors.

Formal definition Higgins et al., 2018 provide a rigorous mathematical definition of disentangled representations (from the perspective of group theory) without any assumptions regarding dimensionality or basis: "*A vector representation is called a disentangled representation with respect to a particular decomposition of a symmetry group into subgroups if it decomposes into independent subspaces, where each subspace is affected by the action of a single subgroup, and the actions of all other subgroups leave the subspace unaffected.*"

Consider a symmetry group[3] G, world state vector $s \in \mathcal{S}$ (i.e., ground truth factors which generate observations), data vector $\mathbf{x} \in \mathcal{X}$, and representation vector $\mathbf{z} \in \mathcal{Z}$. We refer to the group action as $(\cdot)$. Assume G can be decomposed as a direct product $G = g_1 \times g_2 \times \ldots \times g_n$. Representation $\mathbf{z}$ is disentangled with respect to G if:

- There is an action of G on $Z : G \times \mathbf{z} \to \mathbf{z}$
- There exists a mapping from s to $\mathbf{z}$, i.e., $f : s \to \mathbf{z}$ which is equivariant between the action of G on s and $\mathbf{z}$. This condition can be formulated as follows: $g.f(s) = f(g.s),\ \forall g \in G,\ \forall s \in \mathcal{S}$.
- The action of G on $\mathbf{z}$ is disentangled with respect to the decomposition of G. In other words, there is a decomposition $\mathbf{z} = \mathbf{z}_1 \times \ldots \times \mathbf{z}_n$ such that each $\mathbf{z}_i$ is affected only by g_i and invariant to g_j, $\forall j \neq i$.

This rigorous definition substitutes the term "data generative factors" from the generic definition with "disentangled actions of the symmetry group."

Properties of disentangled representation Recently, some properties of the disentangled representation have been introduced (Higgins et al., 2018; X. Liu et al., 2022;

3. A symmetry group refers to a set of transformations or operations that leave an object or a system invariant.

Locatello et al., 2019):

- *Compactness:* This property measures whether a single latent dimension encodes each generative factor of the data. According to Higgins's definition, each disentangled subspace can indeed be *multi-dimensional*. However, numerous disentangled representation methods and metrics promote the idea of a single latent dimension for each generative factor (R. T. Chen et al., 2018; X. Chen et al., 2016; Higgins et al., 2017a).
- *Identifiability:* Learning disentangled representations without guidance is incredibly challenging (Locatello et al., 2019). This is because numerous models could have generated the same observed data. In other words, when we have an observation $\mathbf{x}$, countless generative models could have produced a sample from the same general distribution (Locatello et al., 2019);
- *A causal perspective:* Learning disentangled representations becomes problematic when the factors we seek to disentangle are not truly independent but are interconnected through causal relationships. Causal relationships are directional, meaning that changing the cause will alter the effect but not vice versa. For example, a person's vocal tone may be influenced by their underlying emotional state, which, in turn, affects their facial expressions. Therefore, causal representation learning extends disentangled representation learning by incorporating additional constraints on the relationships between latent variables.

Among the various methods for learning disentangled representations, generative models such as generative adversarial networks (GANs) (Goodfellow et al., 2014) and variational autoencoders (VAEs) (Kingma & Welling, 2014; Rezende et al., 2014) have emerged as powerful contenders. These methods operate in an unsupervised manner and have achieved notable success in disentangled representation learning (R. T. Chen et al., 2018; X. Chen et al., 2016; Higgins et al., 2017a). However, this chapter exclusively focuses on VAEs. This choice is justified because autoencoders constitute a primary research focus of our AIMAC team. Additionally, various issues associated with GANs have been highlighted (Creswell et al., 2018). One prominent issue with GANs is their training instability, often leading to difficulties in convergence and mode collapse, where the generator produces limited outputs. Additionally, GANs are known to demand more meticulous parameter tuning and can be more computationally expensive compared to VAEs. Moreover, VAEs, as we will delve into later, possess properties that render this model *flexible* for expansion into sequential or multimodal data contexts.

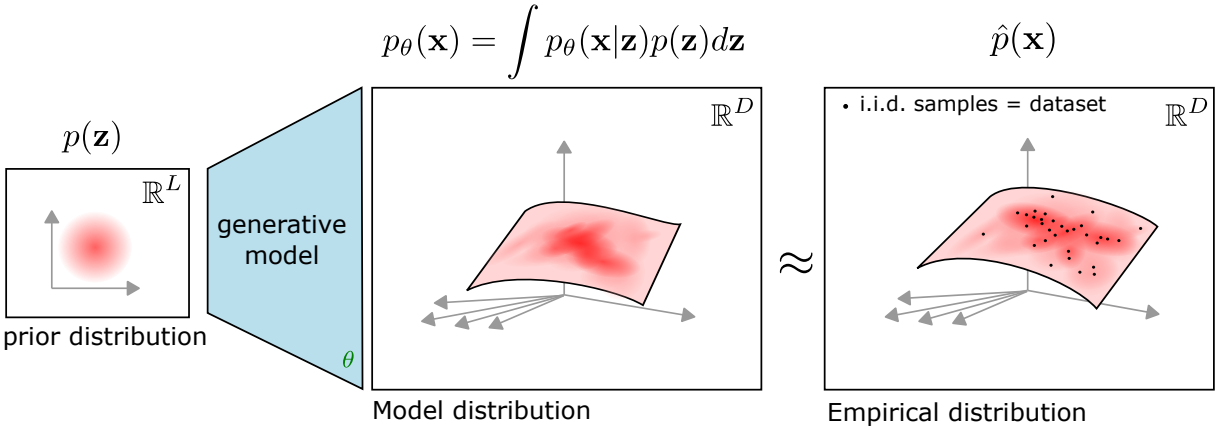


Figure 2.3 – The goal of generative modeling is to estimate the parameters θ such that the model distribution $p_\theta(\mathbf{x})$ closely approximates the empirical distribution $\hat{p}(\mathbf{x})$.

2.4 Variational autoencoder

Real-world data often exhibit complex patterns among their dimensions. Pixels in an image, for example, are not independent but somewhat influenced by their spatial relationships and semantic meanings. This has led to the hypothesis that a lower-dimensional latent representation exists from which the observed high-dimensional data is generated (Fefferman et al., 2016).

Dimensionality reduction techniques aim to capture and leverage these structural dependencies, enabling us to work with a more concise data representation while preserving essential information. In the following Section 2.4.2, we will explore dimensionality reduction methods, starting with a linear method like principal component analysis (PCA). Then, we will progress to discuss probabilistic and generative extensions before introducing the versatile framework known as the VAE.

In Section 2.4.3, we will delve into VAE training, covering topics such as the variational inference, the evidence lower bound, and tricks for efficient training.

Following that, in Sections 2.4.5, 2.4.6, and 2.4.7, we will expand upon VAE's capabilities, exploring its applications in disentanglement, dynamical modeling, and multimodal modeling, respectively.

Background Bayesian rules are fundamental probability theory and statistics principles. These rules govern updating and manipulating probability distributions based on new evidence or information. One of the foundational Bayesian rules is **Bayes the-**

orem, which establishes a mathematical relationship between conditional probabilities. It enables us to calculate the probability of a random variable given certain observed outcomes or data. Bayes' Theorem is often expressed as:

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{z}) \cdot p(\mathbf{z})}{p(\mathbf{x})}. \tag{2.5}$$

2.4.1 Notes on latent-variable generative modeling

Generative modeling (Ruthotto & Haber, 2021; Salakhutdinov, 2015) aims to learn a probabilistic model of an observable random variable $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^D$, as illustrated in Figure 2.3. In this context, the goal is to estimate the parameters θ that minimize the discrepancy between the model distribution $p_\theta(\mathbf{x})$ and the empirical distribution $\hat{p}(\mathbf{x})$, where this discrepancy is often quantified using metrics such as the KL divergence:

$$\min_\theta \Big\{ D_{KL}(\hat{p}(\mathbf{x}) \parallel p_\theta(\mathbf{x})) = \mathbb{E}_{\hat{p}(\mathbf{x})}\Big[\log \hat{p}(\mathbf{x}) - \log p_\theta(\mathbf{x}) \Big] \Big\}. \tag{2.6}$$

This optimization problem is equivalent to the maximum log-marginal likelihood parameter estimation

$$\max_\theta \Big\{ \underbrace{\mathbb{E}_{\hat{p}(\mathbf{x})}\Big[\log p_\theta(\mathbf{x}) \Big]}_{L(\theta)} \Big\}. \tag{2.7}$$

From a mathematical perspective, we consider a dataset $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathcal{X}\}$ comprising N independent and identically distributed (i.i.d.) observations of $\mathbf{x}$. The empirical distribution of $\mathbf{x}$ is defined as $\hat{p}(\mathbf{x}) = \frac{1}{N} \sum_{\mathbf{x}_n \in \mathcal{D}} \delta(\mathbf{x} - \mathbf{x}_n)$, where δ denotes the Dirac delta function. The Dirac delta function is nonzero only at zero, taking the value 1. Generative models generally fall into two categories: **explicit** and **implicit**. Explicit models define a distribution's density function, characterizing data distribution and inferring sample likelihood (Kingma & Welling, 2014; Ngiam et al., 2011; J. Xie et al., 2016). We often refer to them as *latent variable models*. However, they often encounter computational challenges due to unknown normalization constants [4]. On the other hand, implicit models, like GANs (Goodfellow et al., 2014), directly produce diverse samples by transforming random noise into generated samples, but encounter challenges during training, such as instability and mode collapse.

4. The normalization constant is a constant factor used to scale a probability distribution function so that the total probability over all possible outcomes sums up to 1.

Latent variable In this framework, we assume that the observed data $\mathbf{x} \in \mathbb{R}^D$ is generated from an unobserved latent variable $\mathbf{z} \in \mathbb{R}^L$ through a probabilistic process (Bishop, 1998; Borsboom et al., 2003). Consider facial images $\mathbf{x}$ as an example, where substantial variability arises from factors like gender, eye color, hair color, pose, etc. However, these underlying factors are latent, meaning they are not directly accessible unless explicitly annotated. The objective is to create explicit models for these latent factors using the latent variables $\mathbf{z}$. Incorporating latent variables into parameterized models serves multiple purposes. It can simplify the probabilistic representation of a problem, making it more accessible for sampling the overall model $p_\theta(\mathbf{x}, \mathbf{z})$. Additionally, latent variables allow for modeling the impact of an external factor, whether observable or hidden, on the data $\mathbf{x}$. Moreover, they offer control over the generation of $\mathbf{x}$ by selecting a latent variable $\mathbf{z}$ and obtaining corresponding $\mathbf{x}$ samples from $p_\theta(\mathbf{x}|\mathbf{z})$.

Relationships between models Deep generative models are designed to capture the intricate connections between observed and latent variables. In latent variable models, the marginal likelihood $p_\theta(\mathbf{x})$ is defined by marginalizing the joint distribution $\mathbf{x}$ and $\mathbf{z}$ (or called the generative model) as follows:

$$\log p_\theta(\mathbf{x}) = \log \int p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \log \int p(\mathbf{z}) p_\theta(\mathbf{x}|\mathbf{z}) d\mathbf{z}, \tag{2.8}$$

where $p(\mathbf{z})$ is the prior over the latent vector and $p_\theta(\mathbf{x}|\mathbf{z})$ is the parametric conditional likelihood, which indicates how the observed data are generated from the latent vector. In latent-variable generative modeling, finding the posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$ (or called the inference model) is important because it represents the conditional distribution of latent variables $\mathbf{z}$ given the observations $\mathbf{x}$. Automatically discovering the underlying structure in the latent space is a fundamental aspect of model training.

When directly maximizing the logarithm of the marginal likelihood $L(\theta)$ (assuming all distributions are tractable), as depicted by Equation 2.7, the gradient of the log-marginal likelihood over the dataset $\mathcal{D}$ is obtained as (Y. Kim et al., 2018):

$$\nabla_\theta L(\theta) = \sum_{n=1}^{N} \mathbb{E}_{p_\theta(\mathbf{z}|\mathbf{x})} \Big[\nabla_\theta p(\mathbf{x}_n, \mathbf{z}) \Big]. \tag{2.9}$$

It is important to note that the aforementioned gradient involves an expectation over the posterior $p_\theta(\mathbf{z}|\mathbf{x}_n)$, illustrating how inference serves as a subroutine in the learning process.

With this gradient expression, we update the parameters using: $\theta^{new} = \theta^{old} + \eta \nabla_\theta L(\theta^{old})$, where η is the learning rate. In scenarios where the posterior inference becomes intractable, approximation methods like *variational inference* come into play, which will be discussed in Subsection 2.4.3. When learning the parameters of a latent variable model, variational inference optimizes a *lower bound* on the log-marginal likelihood. This lower bound is based on an approximate posterior distribution over latent variables.

Notations In the following, $\mathcal{N}(.; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, diag{.} is the operator that forms a diagonal matrix from a vector by putting the vector entries on the diagonal.

2.4.2 From PCA to variational autoencoder

Principal component analysis (PCA), Figure 2.4(a) PCA is interested in finding projections $\hat{\mathbf{x}}$ of data points $\mathbf{x}$ that are as similar to the original data points as possible (i.e., we have $\hat{\mathbf{x}} \approx \mathbf{x}$) but which have a significantly lower intrinsic dimensionality (Pearson, 1901). More concretely, we consider an i.i.d dataset $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^D\}$, $\forall n \in \{1, \cdots, N\}$, with the data covariance matrix $\mathbf{S} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu})^T (\mathbf{x}_n - \boldsymbol{\mu})$, where $\boldsymbol{\mu}$ is an estimation of the dataset mean.

We assume the existence of a low-dimensional compressed representation $\mathbf{z} \in \mathbb{R}^L$, where $L \ll D$, and establish a linear relationship as $\mathbf{x}_n = \mathbf{B}\mathbf{z}_n + \boldsymbol{\mu} \in \mathbb{R}^L$. Here, $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_L] \in \mathbb{R}^{D \times L}$ is the projection matrix, and its columns, known as principal components, are orthogonal (i.e., $\mathbf{b}_i^T \mathbf{b}_j = 0$ when $i \neq j$) and normalized (i.e., $\mathbf{b}_i^T \mathbf{b}_i = 1$).

PCA, illustrated in Figure 2.4(a), can be conceptualized as a pair of operations resembling an encoder-decoder setup: $\mathbf{z}_n = \mathbf{B}^T(\mathbf{x}_n - \boldsymbol{\mu})$ represents the encoder, while $\mathbf{x}_n = \mathbf{B}\mathbf{z}_n + \boldsymbol{\mu}$ corresponds to the decoder. The linear transformation by matrix $\mathbf{B}$ serves as the decoder, translating the lower-dimensional vector $\mathbf{z}_n \in \mathbb{R}^L$ back into the original data space $\mathbb{R}^D$. Conversely, $\mathbf{B}^T$ acts as the encoder, transforming the centred original data $(\mathbf{x}_n - \boldsymbol{\mu})$ into a lower-dimensional vector $\mathbf{z}_n$.

Theorem 2.4.1. *The principal components* $[\mathbf{b}_1, \ldots, \mathbf{b}_L]$ *of PCA are the eigenvectors of the data covariance matrix* $\mathbf{S}$*, ordered by decreasing eigenvalue.*

This theorem can be demonstrated using mathematical techniques, like maximizing the variance in the projected space (Murphy, 2012; Tharwat, 2016) or minimizing the

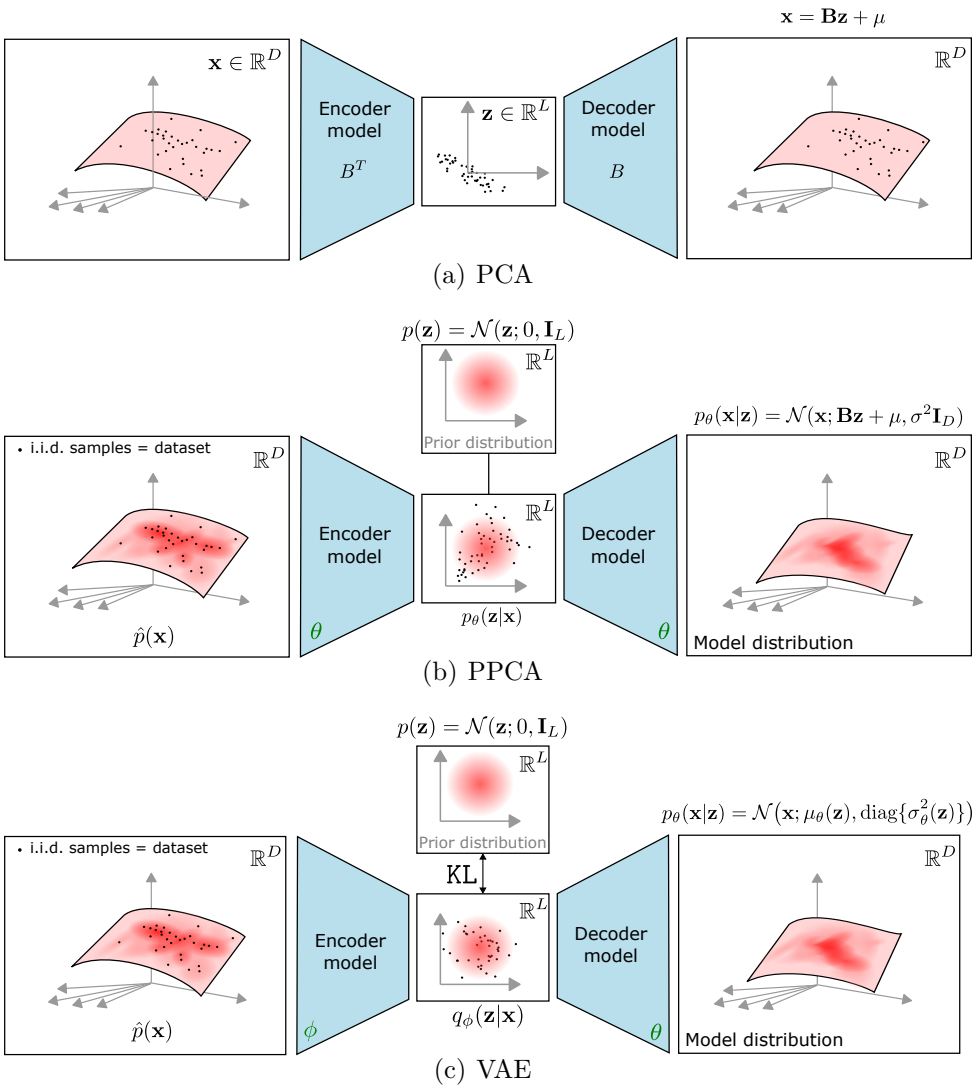


Figure 2.4 – Illustration showcasing the architectural evolution from PCA to PPCA and VAE. Each method represents a step towards richer probabilistic modeling and more expressive latent representations

average reconstruction error (Bishop, 2006b; Deisenroth et al., 2020).

Probabilistic principal component analysis (PPCA), Figure 2.4(b) is a probabilistic extension of PCA (Tipping & Bishop, 1999). Its goal is to bring probabilistic modeling to PCA, which can be advantageous in several contexts (e.g., missing data imputation (Qu et al., 2009)).

We will explicitly define the probabilistic model for linear dimensionality reduction. In

this context, we assume the presence of a continuous latent variable $\mathbf{z} \in \mathbb{R}^L$ with a standard normal prior distribution, denoted as $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}_L)$, where $\mathbf{I}_L$ is a $L \times L$ identity matrix. Additionally, we assume the presence of an affine relationship between these latent variables and the observed data $\mathbf{x}$, which can be expressed as $\mathbf{x} = \mathbf{B}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon} \in \mathbb{R}^D$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \sigma^2\mathbf{I}_D)$ represents the Gaussian observation noise. We have $\mathbf{B} \in \mathbb{R}^{D \times L}$ and $\boldsymbol{\mu} \in \mathbb{R}^D$, which describe the affine mappings from the latent to the observed variables. Consequently, PPCA connects the latent and observed variables through the following parametric ($\theta = \{\mathbf{B}, \boldsymbol{\mu}, \sigma\}$) conditional likelihood distribution:

$$p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}\Big(\mathbf{x}; \mathbf{B}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I}_D\Big), \tag{2.10}$$

where the likelihood distribution acts as a mapping from latent vectors $\mathbf{z}$ to observed data vectors $\mathbf{x}$ (i.e., a decoder). We note that the PCA is the specific case of PPCA when the variance of the noise tends to 0 (i.e., degenerate Gaussian $\sigma \to 0$). Using the theory of Gaussian models, we can easily formalize the parametric joint distribution $p_\theta(\mathbf{x}, \mathbf{z})$ as well as the parametric posterior distributions $p_\theta(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \mathbf{m}, \mathbf{C})$, where the mean vector $\mathbf{m} = \mathbf{B}^T\Big(\mathbf{B}\mathbf{B}^T + \sigma^2\mathbf{I}\Big)^{-1}(\mathbf{x} - \boldsymbol{\mu})$ and the posterior covariance $\mathbf{C} = \mathbf{I} - \mathbf{B}^T\Big(\mathbf{B}\mathbf{B}^T + \sigma^2\mathbf{I}_D\Big)^{-1}\mathbf{B}$. Note that the posterior covariance does not depend on the observed data $\mathbf{x}$, while the mean vector $\mathbf{m}$ has an affine relationship with $\mathbf{x}$. The posterior distribution maps from observed data to latent space (i.e., encoder). Posterior distributions also help in quantifying uncertainty (Deisenroth et al., 2020). For a new observation $\mathbf{x}_{new}$ and the corresponding latent variable $\mathbf{z}_{new}$, if the posterior distribution $p(\mathbf{z}_{new}|\mathbf{x}_{new})$ exhibits a high variance, it often indicates the possibility of encountering an outlier.

In PPCA, while the maximum marginal likelihood estimation is analytically tractable, practical parameter estimation in θ often requires iterative algorithms like the expectation-maximization (EM) algorithm. EM is an iterative method for learning latent variable models with tractable posterior inference. It maximizes a lower bound on the log marginal likelihood at each iteration and can be scaled to *large high-dimensional* datasets. EM is based on a two-step procedure: (i) an *expectation* step, estimating the expectation $\mathbb{E}_{p_\theta(\mathbf{z}|\mathbf{x})}[p_\theta(\mathbf{x}, \mathbf{z})]$ over the latent variables $\mathbf{z}$; (ii) a *maximization* step, that optimizes the parameters θ to maximize the first expectation.

Remark. PPCA is a specific instance within the family of factor analysis (FA) methods (Rummel, 1988). In PPCA, the Gaussian noise is isotropic, meaning that the covariance matrix takes the form $\boldsymbol{\sigma}^2\mathbf{I}$. In contrast, FA assumes any diagonal covariance matrix,

allowing for variability along different directions.

Variational autoencoder (VAE), Figure 2.4(c) The VAE was initially proposed by Kingma and Welling, 2014 and Rezende et al., 2014. A VAE decoder can be considered a generalization of the PPCA, with a nonlinear (instead of linear) relationship between $\mathbf{z}$ and the parameters θ.

The VAE decoder models $p_\theta(\mathbf{x}|\mathbf{z})$ (representing the probability of generating the observed data $\mathbf{x}$ given a specific vector of the latent variable $\mathbf{z}$), which is usually assumed to be Gaussian distribution:

$$p_\theta\left(\mathbf{x}|\mathbf{z}\right) = \mathcal{N}\Big(\mathbf{x}; \boldsymbol{\mu}_\theta(\mathbf{z}), \mathrm{diag}\{\boldsymbol{\sigma}_\theta(\mathbf{z})\}\Big), \tag{2.11}$$

where $\boldsymbol{\mu}_\theta$ and $\boldsymbol{\sigma}_\theta$ are nonlinear functions of $\mathbf{z}$ modeled by the decoder DNN, where θ denotes the weights of the decoder neural network. For the sake of computational efficiency, we typically opt for diagonal covariance matrices. This choice stems from the fact that the number of parameters within a covariance matrix grows quadratically with the dimensions of the variables.

On the other hand, the VAE encoder models $q_\phi(\mathbf{z}|\mathbf{x})$, which approximates the intractable exact posterior $p_\theta(\mathbf{z}|\mathbf{x})$ using *variational inference* (this part will be addressed in the following section). A common choice for the approximate posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$ is to use a Gaussian distribution:

$$q_\phi\left(\mathbf{z}|\mathbf{x}\right) = \mathcal{N}\Big(\mathbf{z}; \boldsymbol{\mu}_\phi(\mathbf{x}), \mathrm{diag}\{\boldsymbol{\sigma}_\phi(\mathbf{x})\}\Big), \tag{2.12}$$

where $\boldsymbol{\mu}_\phi$ and $\boldsymbol{\sigma}_\phi$ are nonlinear functions of $\mathbf{x}$ modeled by the encoder DNN, where ϕ denotes the weights of the encoder neural network.

2.4.3 Variational Inference

Optimizing the generative model involves maximizing the likelihood parameter estimation in Problem 2.7. However, directly addressing this optimization problem can be challenging, if not impossible, when the marginal likelihood becomes analytically intractable. The intractability issue is rooted in the integral within Equation 2.8, which cannot be calculated directly. In the context of the VAE, the non-linearity is a result of the complex connection between the latent and observed variables. The observed variables are

generated from the latent ones through a DNN, turning $p_\theta(\mathbf{x}|\mathbf{z})$ into a nonlinear function of $\mathbf{z}$.

A common strategy to alleviate this problem is to exploit the latent variable aspect of the model to maximize an intractable log-likelihood lower bound, as described in Neal and Hinton, 1998's work. This procedure essentially relies on the posterior distribution of latent variables or its approximation. Taking an arbitrary *variational distribution/family* $q(\mathbf{z})$, we can derive the following decomposition of the log-marginal likelihood:

$$\begin{aligned}
\log p_\theta(\mathbf{x}) &= \mathbb{E}_{q(\mathbf{z})}\Big[\log p_\theta(\mathbf{x})\Big] \\
&= \mathbb{E}_{q(\mathbf{z})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{p_\theta(\mathbf{z}|\mathbf{x})}\right] \\
&= \mathbb{E}_{q(\mathbf{z})}\left[\log \left\{\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})} \cdot \frac{q(\mathbf{z})}{p_\theta(\mathbf{z}|\mathbf{x})}\right\}\right] \\
&= \mathbb{E}_{q(\mathbf{z})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right] + \mathbb{E}_{q(\mathbf{z})}\left[\log \frac{q(\mathbf{z})}{p_\theta(\mathbf{z}|\mathbf{x})}\right] \\
&= \underbrace{\mathbb{E}_{q(\mathbf{z})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right]}_{\mathcal{L}(q,\theta)} + D_{KL}\Big(q(\mathbf{z}) \parallel p_\theta(\mathbf{z}|\mathbf{x})\Big),
\end{aligned} \tag{2.13}$$

Since $D_{KL}\Big(q(\mathbf{z}) \parallel p_\theta(\mathbf{z}|\mathbf{x})\Big)$ is always non-negative, the $\mathcal{L}(q,\theta)$ is a lower-bound on $\log p_\theta(x)$

$$\log p_\theta(\mathbf{x}) \geq \mathcal{L}(q,\theta), \tag{2.14}$$

where $\mathcal{L}(q,\theta)$ is commonly known as the evidence lower bound (ELBO) (Jordan et al., 1999). In Equation 2.14, equality is achieved, signifying that the ELBO precisely matches the log-marginal likelihood, only when the variational distribution $q(\mathbf{z})$ matches with the exact posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$ (i.e., $q(\mathbf{z}) = p_\theta(\mathbf{z}|\mathbf{x})$).

When employing the EM algorithm to optimize both $q(\mathbf{z})$ and θ for the maximum ELBO, which is called the *variational expectation maximization* (Neal & Hinton, 1998), the process unfolds in two steps. The expectation step, which usually performs exact posterior inference, is replaced with variational inference, which finds the best variational approximation to the true posterior. In contrast, the maximization step focuses on optimizing the ELBO with respect to θ (i.e., maximizes the expected complete data likelihood where the expectation is taken with respect to the variational posterior). Let us consider the case where the variational family is flexible enough to include the true posterior. The above reduces to the

classic EM algorithm, since in the first step $D_{KL}\Big(q(\mathbf{z}) \parallel p_\theta(\mathbf{z}|\mathbf{x})\Big)$ is minimized when $q(\mathbf{z})$ matches the true posterior $p_\theta(\mathbf{z}|\mathbf{x})$, which is why $q(\mathbf{z})$ will be denoted by the approximate posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$. The EM algorithm can optimize an implicit distribution to match the posterior without directly computing the marginal log-likelihood, resolving Problem 2.7. However, this method might remain computationally expensive (Y. Kim et al., 2018). An alternative strategy, such as amortized variational inference (Hoffman et al., 2013) and VAE, employs a trained neural network called an inference network. This network predicts variational parameters for a given input $\mathbf{x}$ by undergoing training through gradient ascent, enabling it to perform variational inference across all data points.

2.4.4 Training the variational autoencoder

In the VAE methodology (Kingma & Welling, 2014; Rezende et al., 2014), the ELBO represented in Equation 2.13, now denoted as $\mathcal{L}_{\text{VAE}}(\theta, \phi)$, is optimized using stochastic gradient-based techniques. This process entails the iterative refinement of both the generative and inference model parameters. Throughout VAE training, the encoder and decoder networks are interconnected, as illustrated in Figure 2.2, and the parameters θ and ϕ are jointly updated.

The ELBO can be further expressed as (the best-known formulation):

$$\mathcal{L}_{\text{VAE}}(\theta, \phi) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big[\ln p_\theta(\mathbf{x}|\mathbf{z})\Big]}_{\text{Reconstruction}} - \underbrace{D_{\text{KL}}\Big(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\Big)}_{\text{Regularisation}}. \tag{2.15}$$

The ELBO consists of two components: the first part represents the *accuracy of the encoding-decoding* process. If we opt for a Gaussian generative model $p_\theta(\mathbf{x}|\mathbf{z})$ with an identity covariance matrix, this term quantifies the negative mean-squared error between the original data and the decoder's output, with some additional constants.

The second component serves as a *regularization term*, compelling the approximate posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$ to closely match the prior distribution $p(\mathbf{z})$. In Gaussian VAE where $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; 0, \mathbf{I}_L)$ and $\mathrm{q}_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}\Big(\mathbf{z}; \boldsymbol{\mu}_\phi(\mathbf{x}), \mathrm{diag}\{\boldsymbol{\sigma}_\phi(\mathbf{x})\}\Big)$, the regularization term possesses a known analytical expression, dependent on $\boldsymbol{\mu}_\phi$ and $\boldsymbol{\sigma}_\phi$, as elaborated in further detail in the work by Kingma and Welling, 2014.

In practical VAE training, a training dataset $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathcal{X}\}$ is employed. Assuming the training vectors are i.i.d., the VAE training goal is to maximize the ELBO. This objective involves summing the individual ELBO scores computed for each training

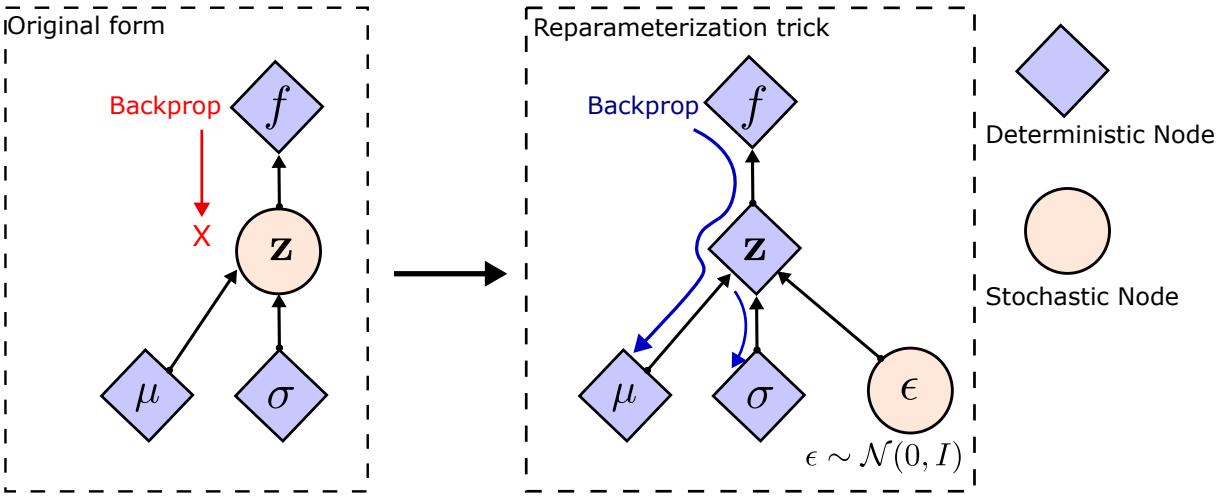


Figure 2.5 – The reparameterization trick involves reformulating the sampling of latent variables. When $\mathbf{z}$ is sampled stochastically (in the left) from a parameterized distribution, the gradients must flow through the stochastic node. In contrast, the reparameterization trick (in the right) allows a gradient path through a deterministic node, which makes it differentiable for gradient-based optimization, ensuring efficient training of VAEs.

vector. The ELBO can be expressed as:

$$\mathcal{L}_{\text{VAE}}(\theta, \phi) = \sum_{n=1}^{N} \mathbb{E}_{q_\phi(\mathbf{z}_n|\mathbf{x}_n)}\Big[\ln p_\theta(\mathbf{x}_n|\mathbf{z}_n)\Big] - \sum_{n=1}^{N} D_{\text{KL}}\Big(q_\phi(\mathbf{z}_n|\mathbf{x}_n) \parallel p(\mathbf{z}_n)\Big). \tag{2.16}$$

To efficiently find the optimal parameters, stochastic gradient descent (SGD) is commonly employed. The typical approach involves rearranging terms to transform the gradient of the expectation into an expectation of the gradient. This can then be estimated using Monte Carlo samples.

The gradient of the ELBO with respect to θ is given by

$$\begin{aligned}\nabla_\theta \mathbb{E}_{q_\phi}\Big[\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})\Big] &= \mathbb{E}_{q_\phi}\Big[\nabla_\theta \log p_\theta(\mathbf{x}, \mathbf{z})\Big] \\ &= \mathbb{E}_{q_\phi}\Big[\nabla_\theta \log p_\theta(\mathbf{x}|\mathbf{z})\Big],\end{aligned}$$

the first equality holds because the distribution for which we compute the expectation does not depend on θ, so we can push the gradient inside the expectation. The second equality holds because the prior $p(\mathbf{z})$ does not depend on θ. The expectation in the gradient above is typically estimated with Monte Carlo samples. The same principle cannot be applied to $\nabla_\phi \mathbb{E}_{q_\phi}\Big[\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})\Big]$ because ϕ is involved in the probability distribution

itself, preventing the straightforward placement of ∇_ϕ inside the expectation.

To address this, the reparameterization trick is employed as illustrated in Figure 2.5. The reparameterization trick introduces an auxiliary random variable $\boldsymbol{\epsilon}$ that is sampled from a simple and fixed distribution, typically a standard Gaussian $\mathcal{N}(\boldsymbol{\epsilon}; 0, \mathbf{I}_L)$. Doing so allows us to express the random variable $\mathbf{z}$ as a deterministic function of $\boldsymbol{\epsilon}$. The sampled $\boldsymbol{\epsilon}$ is then transformed using a differentiable function, typically involving multiplication by the standard deviation and addition of the mean of $q_\phi(\mathbf{z}|\mathbf{x})$. This transformation allows us to make the sampling process differentiable (because $\boldsymbol{\epsilon}$ is not a function of ϕ), enabling the VAE to be trained efficiently using SGD.

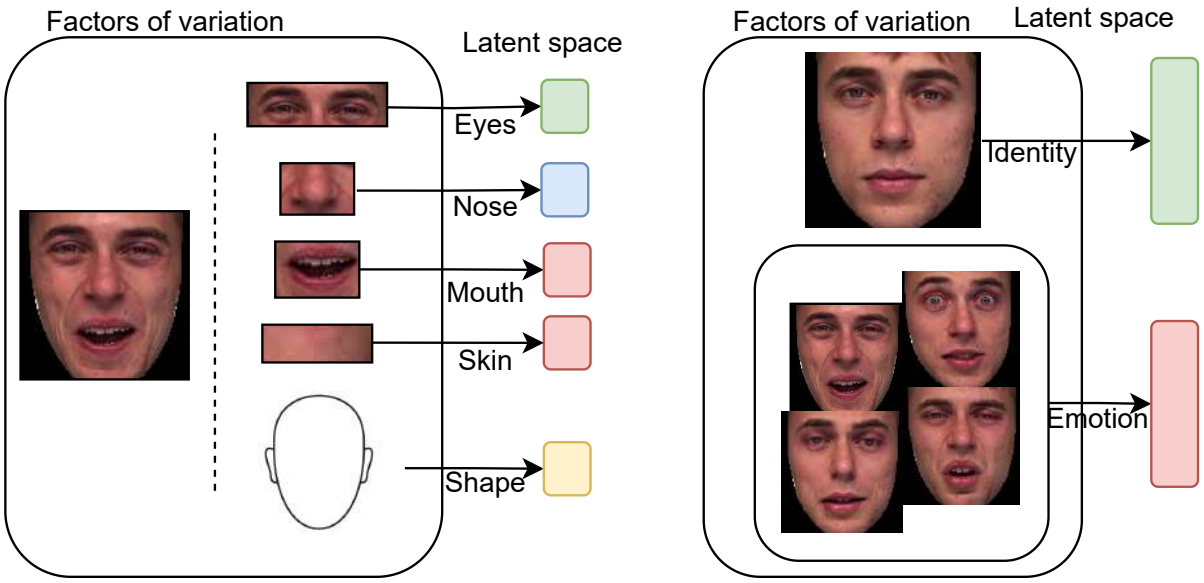


Figure 2.6 – The illustration of the comparison between dimension-wise (left) and vector-wise (right) RL.

2.4.5 Disentanglement in variational autoencoder

Two approaches to disentangled representation learning methods

Disentangled representation methods, as discussed by X. Wang et al., 2022, fall into two broad categories: *dimension-wise* and *vector-wise*, as illustrated in Figure 2.6.

Dimension-wise methods take a fine-grained approach, where individual dimensions (or small sets of dimensions) within the representation correspond to specific generative factors. These methods are typically evaluated using synthetic and simple datasets that

often involve multiple fine-grained latent factors, such as those found in the Dsprite dataset (Matthey et al., 2017).

Vector-wise methods, on the other hand, adopt a more coarse-grained perspective. Different vectors within the representation capture various semantic meanings. Vector-wise disentanglement methods find practical application in real-world scenarios like identity swapping and image classification. Real-world datasets and applications typically involve fewer coarse-grained factors, such as identity and pose.

In the upcoming section, we will delve into dimension-wise methods, with a specific focus on those based on VAE. VAEs serve as the foundational framework for many state-of-the-art disentanglement methods.

Application of disentangled representation

Applications of disentangled representations encompass a wide array of domains, including but not limited to controllable image generation (Zhu et al., 2018), image manipulation (Gabbay & Hoshen, 2019), and domain adaptation (Peng et al., 2019). Moreover, it is expected that better disentangled representations will significantly impact model interpretability (W.-N. Hsu et al., 2017a), abstract reasoning (Van Steenkiste et al., 2019), and the pursuit of fairness in machine learning (Creager et al., 2019).

Some methods of disentanglement in VAE

As noted previously, $\mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}_L)$ is selected as the prior distribution for the latent variable $\mathbf{z}$ in the context of VAEs. This choice is made to introduce independent constraints on the representations that the neural network learns, and it is considered a key factor contributing to VAE's potential for disentanglement, as noted in prior research (R. T. Chen et al., 2018; X. Liu et al., 2022). However, it has been observed that vanilla VAEs exhibit limited disentanglement capability, especially when dealing with complex datasets. To address this challenge, substantial efforts have been made to enhance disentanglement by introducing implicit or explicit inductive biases. These enhancements involve incorporating various regularization techniques, including but not limited to methods like β-VAE (Higgins et al., 2017b), AnnealedVAE (Burgess et al., 2018), FactorVAE (H. Kim & Mnih, 2018) and β-TCVAE (R. T. Chen et al., 2018).

β-VAE The β-VAE is a variant of the VAE framework that incorporates an adjustable hyperparameter, denoted as β, into the original VAE ELBO (Higgins et al., 2017b).

$$\mathcal{L}_{\beta\text{-VAE}}(\theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big[\log p_\theta(\mathbf{x}|\mathbf{z})\Big] - \beta D_{KL}\Big(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\Big). \tag{2.17}$$

Carefully selecting values for β, typically $\beta > 1$, results in more disentangled latent representations, denoted as $\mathbf{z}$. When $\beta = 1$, the β-VAE aligns with the original VAE framework. This introduces additional constraints on the latent bottleneck $\mathbf{z}$ and exerts extra pressure to maintain its factorized structure while being sufficient for data reconstruction (Higgins et al., 2017a). Higher β values, aimed at promoting disentanglement, often entail a trade-off between the fidelity of β-VAE reconstructions and the disentangled nature of its latent variable $\mathbf{z}$. This trade-off is explained in the FactorVAE paragraph below.

AnnealedVAE introduced a modified version of β-VAE, which dynamically adjusts the capacity of the latent encoding during training (Burgess et al., 2018). With low encoding capacity, the model initially focuses on capturing the most salient features to enhance reconstruction quality. As training progresses and capacity increases, the model gradually incorporates additional semantic factors into the latent representation while preserving the disentanglement of previously learned factors. The loss function for AnnealedVAE is defined as:

$$\mathcal{L}_{\text{AnnealdVAE}}(\theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big[\log p_\theta(\mathbf{x}|\mathbf{z})\Big] - \gamma|D_{KL}\Big(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\Big) - C|, \tag{2.18}$$

where γ and C are hyperparameters. During training, C is annealed from zero to some value which is large enough to produce good reconstruction.

FactorVAE H. Kim and Mnih, 2018 introduced a variation of VAE called FactorVAE, which addresses disentanglement by modifying the decomposition of the ELBO. They decompose the expectation of the KL term in Equation 2.15 using a formulation from Hoffman and Johnson, 2016:

$$\mathbb{E}_{\hat{p}(\mathbf{x})}\Big[D_{KL}(q(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}))\Big] = I(\mathbf{x}; \mathbf{z}) + D_{KL}\Big(q(\mathbf{z}) \parallel p(\mathbf{z})\Big), \tag{2.19}$$

where $I(\mathbf{x}; \mathbf{z})$ represents the MI between $\mathbf{x}$ and $\mathbf{z}$, $\hat{p}(\mathbf{x})$ is the empirical distribution of $\mathbf{x}$ and $q(\mathbf{z}) = \mathbb{E}_{\hat{p}(\mathbf{x})}\Big[q(\mathbf{z}|\mathbf{x})\Big]$ is the the aggregated posterior distribution over all data (Makhzani et al., 2015). In β-VAE, both terms in the equation are jointly impacted: (i) $D_{KL}[q(\mathbf{z}) \parallel$

$p(\mathbf{z})]$ encourages $q(\mathbf{z})$ to match the factorized prior $p(\mathbf{z})$, promoting disentanglement; (ii) Additionally, $I(\mathbf{x};\mathbf{z})$ reduces the information about the data $\mathbf{x}$ stored in the latent vector $\mathbf{z}$, which may result in less accurate reconstructions for high values of β in β-VAE model.

H. Kim and Mnih, 2018 argue that penalizing the MI between $\mathbf{x}$ and $\mathbf{z}$ might not be necessary or desirable for improved disentanglement. Instead, they introduce an additional term to the VAE objective (Equation 2.31) that penalizes the dependence of variables within the latent space:

$$\mathcal{L}_{\text{FactorVAE}}(\theta,\phi) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \beta D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\right)}_{\beta\text{-VAE}} - \gamma D_{KL}\Big(q(\mathbf{z}) \parallel \prod_{j=1}^{L} q(\mathbf{z}_j)\Big), \tag{2.20}$$

where $\mathbf{z}_j$ denotes the j-th dimension of the latent variable. The second term is the total correlation (TC) (a popular measure of dependence for multiple random variables), which evaluates the degree of dimension-wise independence on $\mathbf{z}$. The exact values of TC are difficult to calculate without the availability of the closed-form distributions (Bai et al., 2023). H. Kim and Mnih, 2018 propose an approach using "the density ratio trick" (Sugiyama et al., 2012). This method involves training a binary classifier (discriminator) to determine the probability $d(\mathbf{z})$ that its input was sampled from $q(\mathbf{z})$ rather than from $\prod_{j=1}^{L} q(\mathbf{z}_j)$:

$$TC(\mathbf{z}) = D_{KL}(q(\mathbf{z}) \parallel \prod_{j=1}^{L} q(\mathbf{z}_j)) = \mathbb{E}_{q(\mathbf{z})}\left[\log \frac{q(\mathbf{z})}{\prod_{j=1}^{L} q(\mathbf{z}_j)}\right] \approx \mathbb{E}_{q(\mathbf{z})}\left[\log \frac{d(\mathbf{z})}{1-d(\mathbf{z})}\right]. \tag{2.21}$$

For a more detailed explanation of how density ratio estimation $\left(q(\mathbf{z})/\prod_{j=1}^{L} q(\mathbf{z}_j)\right)$ can be reduced to probabilistic classification, please refer to Tiao, 2018. The training process involves the joint training of the discriminator and the VAE.

β-TCVAE serves as an enhancement of the β-VAE. In their work, R. T. Chen et al., 2018 are inspired by an alternative breakdown of the second term presented in Equation 2.15, leading to the following partition (Hoffman & Johnson, 2016):

$$\mathbb{E}_{\hat{p}(\mathbf{x})}\Big[D_{KL}\Big(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\Big)\Big] = \\ \underbrace{D_{KL}\Big(q(\mathbf{z},\mathbf{x}) \parallel q(\mathbf{z})p(\mathbf{x})\Big)}_{\text{(i)}} + \underbrace{D_{KL}\Big(q(\mathbf{z}) \parallel \prod_{j} q(\mathbf{z}_j)\Big)}_{\text{(ii)}} + \underbrace{\sum_{j} D_{KL}\Big(q(\mathbf{z}_j) \parallel p(\mathbf{z}_j)\Big)}_{\text{(iii)}}. \tag{2.22}$$

The first term (i) is the index-code MI, which denotes the mutual information between the data $\mathbf{x}$ and the latent $\mathbf{z}$ variable. The second term (ii) is TC. The last term (iii) is the dimension-wise KL, which primarily encourages the latent dimensions to better match their corresponding priors. According to R. T. Chen et al., 2018, the total correlation term in the ELBO is the most important term in this decomposition for learning disentangled representations. To verify this claim, the authors propose a simple yet general framework for training using "minibatch-weighted sampling" to stochastically estimate the decomposition terms in Equation 2.22.

Note on inductive bias As mentioned in Section 2.3.2 concerning the properties of disentangled representations, achieving disentanglement in an unsupervised manner faces the challenge of identifiability (unsupervised learning of disentangled representations is fundamentally impossible (Locatello et al., 2019)). Using domain knowledge, often called *inductive bias*, becomes pivotal to addressing this challenge. Inductive bias leverages disentanglement priors that introduce a structured framework into learned representations, aligning them more closely with the underlying data generation process. It is worth noting that previous representation learning methods have already harnessed the inductive biases inherent in techniques such as CNNs (LeCun, Bengio, et al., 1995) and RNNs (Graves et al., 2013) or/and the bias of the training data itself. In Chapter 3, we will delve into weakly supervised VAE-based approaches for achieving disentanglement.

However, all the model-based VAE discussed above are primarily designed for static (non-sequential) and unimodal (involving a single input source) data. In the following subsection, we will delve into various models designed specifically to handle sequential data. Additionally, in Subsection 2.4.7, we will explore approaches that tackle the challenges presented by multimodal data in VAE methods. These two subsections provide essential background for understanding our contribution in Chapter 4, where we merge the processing of dynamical and multimodal data within a *unified* disentangled VAE framework, specifically applied to audiovisual speech data.

2.4.6 Dynamical variational autoencoder

The original VAE framework is not designed to handle sequential data and lacks the ability to model temporal dependencies. Each data vector and its corresponding

latent vector are treated independently. However, in recent years, several research papers have proposed extensions of VAEs to process sequential data by incorporating recurrent neural networks (RNNss) (Bayer & Osendorfer, 2014; Chung et al., 2015; Fabius & Van Amersfoort, 2014) or transformers (Jiang et al., 2020), called dynamical variational autoencoders (DVAE) (Girin et al., 2021b). These extensions aim to capture the temporal dependencies within a sequence of data vectors and their corresponding latent vectors. We first define a DVAE in terms of a generative model and then present the general lines of inference and training in the DVAE framework

Generative model of DVAEs DVAEs deal with sequences of observed vectors $\mathbf{x}_{1:T} = \{\mathbf{x}_t \in \mathbb{R}^D\}_{t=1}^T$ and their corresponding latent vectors $\mathbf{z}_{1:T} = \{\mathbf{z}_t \in \mathbb{R}^L\}_{t=1}^T$. The definition of a DVAE generative model requires specifying the joint distribution of both the observed and latent sequential data, denoted by $p_\theta(\mathbf{x}_{1:T}, \mathbf{z}_{1:T})$. A DVAE is thus defined by the following joint probability density function:

$$p_\theta(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = p_\theta(\mathbf{x}_{1:T}|\mathbf{z}_{1:T})p_\theta(\mathbf{z}_{1:T}). \tag{2.23}$$

However, this form lacks insight into the generative process, so we find it more informative to reformulate Equation 2.24 using the chain rule as follows:

$$p_\theta(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = \prod_{t=1}^{T} p_\theta(\mathbf{x}_t|\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t})p_\theta(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{x}_{1:t-1}). \tag{2.24}$$

This specific reformulation, one of the multiple plausible alternatives, follows a *causal* structure: $\mathbf{z}_t$ is generated based on the past latent vector $\mathbf{z}_{1:t-1}$ and the past observed data $\mathbf{x}_{1:t-1}$. The generation of $\mathbf{x}_t$ is based on the past data $\mathbf{x}_{1:t-1}$ and the latent sequence $\mathbf{z}_{1:t}$. In practice, these probabilistic distributions are parameterized using DNNs, often using recurrent neural networks, and are characterized by a set of parameters denoted by θ, where $\mathbf{z}_t$ (or $\mathbf{x}_t$) is the output of RNNs that take $\mathbf{z}_{1:t-1}$ and $\mathbf{x}_{1:t-1}$ (or $\mathbf{z}_{1:t}$ and $\mathbf{x}_{1:t-1}$) as inputs.

Remark. A key observation is that the choice of ordering within the chain rule allows different arrangements of random vectors, which significantly impacts practical implementations, ultimately yielding different sampling procedures.

Inference model of DVAEs The posterior distribution for the state sequence $\mathbf{z}_{1:T}$ is denoted as $p_\theta(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$. Similar to the standard VAE, this posterior distribution becomes intractable due to non-linearities in the generative model. To address this challenge, we introduce an inference model, denoted by $q_\phi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$, which is an approximation to the intractable posterior distribution $p_\theta(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$. Using the chain rule, we can reshape the inference model in the following general form:

$$q_\phi(\mathbf{z}_{1:T}|\mathbf{x}_{1:T}) = \prod_{t=1}^{T} q_\phi(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{x}_{1:T}), \tag{2.25}$$

where the parametric probabilistic distributions $q_\phi(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{x}_{1:T})$ is parameterized by RNNs and rely on a set of parameters denoted by ϕ, which takes the past latent sequence $\mathbf{z}_{1:t-1}$ and the entire $\mathbf{x}_{1:T}$ as inputs.

Training of DVAEs As for the standard VAE, training a DVAE is based on maximizing the ELBO (Equation 2.15). In the case of DVAEs, ELBO is extended to data sequences, using Equations 2.24 and 2.25, as follows (Girin et al., 2021b):

$$\begin{aligned}\mathcal{L}_{\text{DVAE}}(\theta, \phi) = &\sum_{t=1}^{T} \mathbb{E}_{q_\phi(\mathbf{z}_{1:t}|\mathbf{x}_{1:T})} \left[\log p_\theta(\mathbf{x}_t|\mathbf{z}_{1:t-1}, \mathbf{z}_{1:t})\right] \\ &- \sum_{t=1}^{T} \mathbb{E}_{q_\phi(\mathbf{z}_{1:t-1}|\mathbf{x}_{1:T})} \left[D_{KL}(q_\phi(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{x}_{1:T}) \parallel p_\theta(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{x}_{1:t-1}))\right]\end{aligned} \tag{2.26}$$

Similar to the VAE, the ELBO of the DVAE comprises a reconstruction-accuracy term and a regularization term. However, in contrast to the standard VAE, where the regularization term typically has a known analytical form for common distributions, in the DVAE, both the reconstruction accuracy and regularization terms necessitate the computation of Monte Carlo estimates, specifically empirical averages.

There are several DVAE models that differ in the structure and formulation of the generative model and the inference model (Alias Partg Goyal et al., 2017; Bayer & Osendorfer, 2014; Chung et al., 2015; Fabius & Van Amersfoort, 2014; Fraccaro et al., 2016). These DVAE models have two points in common when modeling sequential data: (i) unsupervised training is preserved, and (ii) the structure of the VAE is maintained;

this means that the inference and generative models are jointly learned by maximizing the ELBO. Of particular interest to the present chapter is the disentangled sequential autoencoder (DSAE) (Y. Li & Mandt, 2018), which separates dynamical from static latent information.

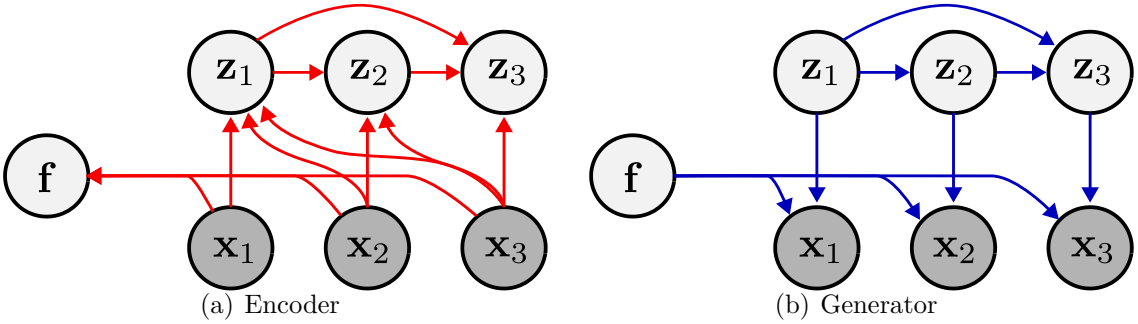


Figure 2.7 – A graphical model visualisation of the generator and the encoder of DSAE.

Disentangled Sequential Autoencoders (DSAE) Introduced by Y. Li and Mandt, 2018, DSAE extends the sequence of latent variables, $\mathbf{z}_{1:T}$, by incorporating an additional latent vector referred to as $\mathbf{f}$. This extension is designed to encode the sequence-level characteristics of the data. Specifically, $\mathbf{z}_t$ captures *time-dependent* features, such as the dynamical behavior of a face in a video clip. In contrast, $\mathbf{f}$ is dedicated to representing all other relevant information, which are inherently *time-independent*, including the intrinsic characteristics of the face within the same video clip (e.g., identity, gender).

If we classify this method based on the two categories introduced in Section 2.4.5, DSAE falls into the category of vector-wise methods. This categorization stems from the fact that disentanglement within DSAE occurs at the level of vector structures, with $\mathbf{f}$ addressing static aspects and $\mathbf{z}_t$ capturing dynamical features.

Y. Li and Mandt, 2018 characterize the generative model $p_\theta\Big(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}, \mathbf{f}\Big)$ as represented in Equation 2.27, which is also represented through the Bayesian network in Figure 2.7(b).

$$p_\theta\Big(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}, \mathbf{f}\Big) = p(\mathbf{f}) \prod_t^T p_\theta\Big(\mathbf{z}_t|\mathbf{z}_{1:t-1}\Big) p_\theta\Big(\mathbf{x}_t|\mathbf{z}_t, \mathbf{f}\Big), \tag{2.27}$$

Equation 2.27 indicates that, at time index t, the observed data vector $\mathbf{x}_t$ is generated from the static latent variable $\mathbf{f}$ and the dynamical latent variable at time index t ($\mathbf{z}_t$). All latent variables are assumed independent, and the prior for the dynamical variable has the autoregressive structure $p_\theta\Big(\mathbf{z}_t|\mathbf{z}_{1:t-1}\Big)$.

Each conditional distribution that appears in a product over the time indices in Equation 2.27 is modeled as a Gaussian with a diagonal covariance, and its parameters are provided by deep neural networks (decoders) that take as input the variables after the conditioning bars. Standard feed-forward fully connected neural networks are used to parameterize the conditional distributions $p_\theta\Big(\mathbf{x}_t|\mathbf{z}_t, \mathbf{f}\Big)$. The autoregressive structure of the priors over the latent dynamical variable $p_\theta(\mathbf{z}_t|\mathbf{z}_{1:t-1})$ requires the use of RNNs. Finally, the prior over the static latent variable $p(\mathbf{f})$ is a Gaussian with zero mean and identity covariance matrix.

Using the Bayesian network of the model, the chain rule of probabilities, and D-separation (Geiger et al., 1990), it is possible to analyze how the observed and latent variables depend on each other in the exact posterior, and define an inference model with the same dependencies. We can factorize the inference model as follows:

$$q_\phi\Big(\mathbf{z}_{1:T}, \mathbf{f}|\mathbf{x}_{1:T}\Big) = q_\phi\Big(\mathbf{f}|\mathbf{x}_{1:T}\Big) \prod_t^T q_\phi\Big(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{x}_{t:T}, \mathbf{f}\Big). \tag{2.28}$$

The probabilistic graphical model of DSAE during inference is represented in Figure 2.7(a), corresponding to the factorization in Equation 2.28. It can be interpreted as follows: First, we infer the static latent variable $\mathbf{f}$ from the observed data sequence $\mathbf{x}_{1:T}$, which corresponds to the computation of $q_\phi\Big(\mathbf{f}|\mathbf{x}_{1:T}\Big)$. Next, we infer the dynamical latent variable $\mathbf{z}_t$ from the previously inferred variable $\mathbf{f}$ and the observed data variable $\mathbf{x}_{t:T}$, which corresponds to the computation of $q_\phi\Big(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{x}_{t:T}, \mathbf{f}\Big)$.

In this inference model, each conditional distribution is modeled as a Gaussian with a diagonal covariance, and its parameters (mean vector and variance coefficients) are provided by deep neural networks (encoders) that take as input the variables after the conditioning bars.

As in standard DVAEs, learning the DSAE generative and inference model parameters consists in maximizing the ELBO:

$$\begin{aligned}\mathcal{L}_{\text{DSAE}}(\theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{f}|\mathbf{x}_{1:T})}\Bigg[\sum_{t=1}^T \mathbb{E}_{q_\phi(\mathbf{z}_t|\mathbf{f}, \mathbf{x}_{1:T})}\left[\log p_\theta(\mathbf{x}_t|\mathbf{z}_t, \mathbf{f})\right] \\ - \sum_{t=1}^T \mathbb{E}_{q_\phi(\mathbf{z}_{1:t-1}|\mathbf{f}, \mathbf{x}_{1:T})}\Big[D_{KL}\Big(q_\phi(\mathbf{z}_t|\mathbf{f}, \mathbf{x}_{t:T}) \parallel p_\theta(\mathbf{z}_t|\mathbf{z}_{1:t-1})\Big)\Big]\Bigg] - D_{KL}\Big(q_\phi(\mathbf{f}|\mathbf{x}_{1:T}) \parallel p(\mathbf{f})\Big)\end{aligned} \tag{2.29}$$

The authors present compelling results across two sequential modalities in the exper-

imental section. First, with cartoon video clips, the model demonstrates its ability to transform the content of a given sequence using the variable $\mathbf{f}$, effectively performing content permutations while preserving the dynamical information. Secondly, with audio modality, DSAE showcases its capacity to convert a male speaker's voice into that of a female and vice versa.

2.4.7 Multimodal variational autoencoder

VAEs have gained substantial attention for modeling multimodal data due to their advantages over other generative models, particularly GANs (Goodfellow et al., 2014). VAEs, equipped with both encoder and decoder models, offer a more stable and efficient training process compared to GANs, rendering them well-suited for multimodal generative modeling (Suzuki & Matsuo, 2022). Various techniques have been developed to learn a unified latent space for multiple diverse input data within the VAE framework, referred to as multimodal variational autoencoder (MVAE). In the subsequent sections, we introduce the foundational aspects of an MVAE, encompassing its generative model, and outline the key elements of inference and training within this framework.

Generative model of MVAEs MVAEs deal with a set of vectors describing N modalities $\mathbf{x}^{(1:N)} = \{\mathbf{x}^{(i)} \in \mathbb{R}^D\}_{i=1}^N$ and their corresponding latent vectors $\mathbf{z}^{(1:N)} = \{\mathbf{z}^{(i)} \in \mathbb{R}^L\}_{i=1}^N$. The definition of a MVAE generative model requires to specify the joint distribution of both the observed and latent variable, denoted as $p_\theta(\mathbf{x}^{(1:N)}, \mathbf{z}^{(1:N)})$. A MVAE is thus defined by the following joint probability density function using chain rule:

$$p_\theta(\mathbf{x}^{(1:N)}, \mathbf{z}^{(1:N)}) = p_\theta(\mathbf{x}^{(1:N)}|\mathbf{z}^{(1:N)})p_\theta(\mathbf{z}^{(1:N)}). \tag{2.30}$$

These probabilistic distributions are parameterized by DNNs, where their parameters are called θ.

Inference model of MVAEs The posterior distribution for the latent variables $\mathbf{z}^{(1:N)}$ is denoted as $p_\theta(\mathbf{z}^{(1:N)}|\mathbf{x}^{(1:N)})$. Similar to the standard VAE, this posterior distribution becomes intractable due to the presence of non-linearities in the generative model. To address this challenge, we introduce an inference model, represented

as $q_\phi(\mathbf{z}^{(1:N)}|\mathbf{x}^{(1:N)})$, which serves as an approximation to the intractable posterior distribution $p_\theta(\mathbf{z}^{(1:N)}|\mathbf{x}^{(1:N)})$. The infernce model is parameterized by DNNs and relies on a set of parameters represented as ϕ.

Training of MVAEs As for the standard VAE, training a MVAE is based on maximizing the ELBO (Equation 2.15). The multimodal ELBO, initially introduced by M. Wu and Goodman, 2018, represents the primary objective function optimized by all MVAE models:

$$\begin{aligned}\mathcal{L}(\theta,\phi) = &\mathbb{E}_{q_\phi(\mathbf{z}^{(1:N)}|\mathbf{x}^{(1:N)})}\left[\log p_\theta(\mathbf{x}^{(1:N)}|\mathbf{z}^{(1:N)})\right] \\ &- \mathbb{E}_{q_\phi(\mathbf{z}^{(1:N)}|\mathbf{x}^{(1:N)})}\left[D_{KL}(q_\phi(\mathbf{z}^{(1:N)}|\mathbf{x}^{(1:N)}) \parallel p(\mathbf{z}^{(1:N)}))\right].\end{aligned} \tag{2.31}$$

According to Shi et al., 2019, the multimodal generative model should satisfy four criteria:

— **Latent Factorization**: The model's latent space should implicitly be divided into subspaces that capture shared and modality-specific information. This factorization is crucial for downstream tasks because well-separated representations are more versatile and adaptable to various applications;

— **Coherent Joint Generation**: When generating data in different modalities from the same latent variable, there should be *coherence* regarding the shared aspects represented in the latent space. For example, if the latent representation signifies "happiness", the generated speech audio should correspondingly express happiness, and the generated facial expression should portray a happy emotion;

— **Coherent Cross Generation**: Imagine a scenario where the model generates facial expressions based on audio input, such as spoken words carrying an emotion. If the audio input describes "anger," the generated facial expressions should be coherent with anger-related cues, such as a furrowed brow or clenched jaw;

— **Synergy** [5]: Observing data in multiple modalities should enhance the quality of the generative model for each modality. In other words, combining multimodal observations should lead to improved data generation in each modality compared with considering each modality in isolation. For instance, images and textual

5. In **liang2023quantifying**, synergy is defined as the emergence of novel information that was not initially present in either modality.

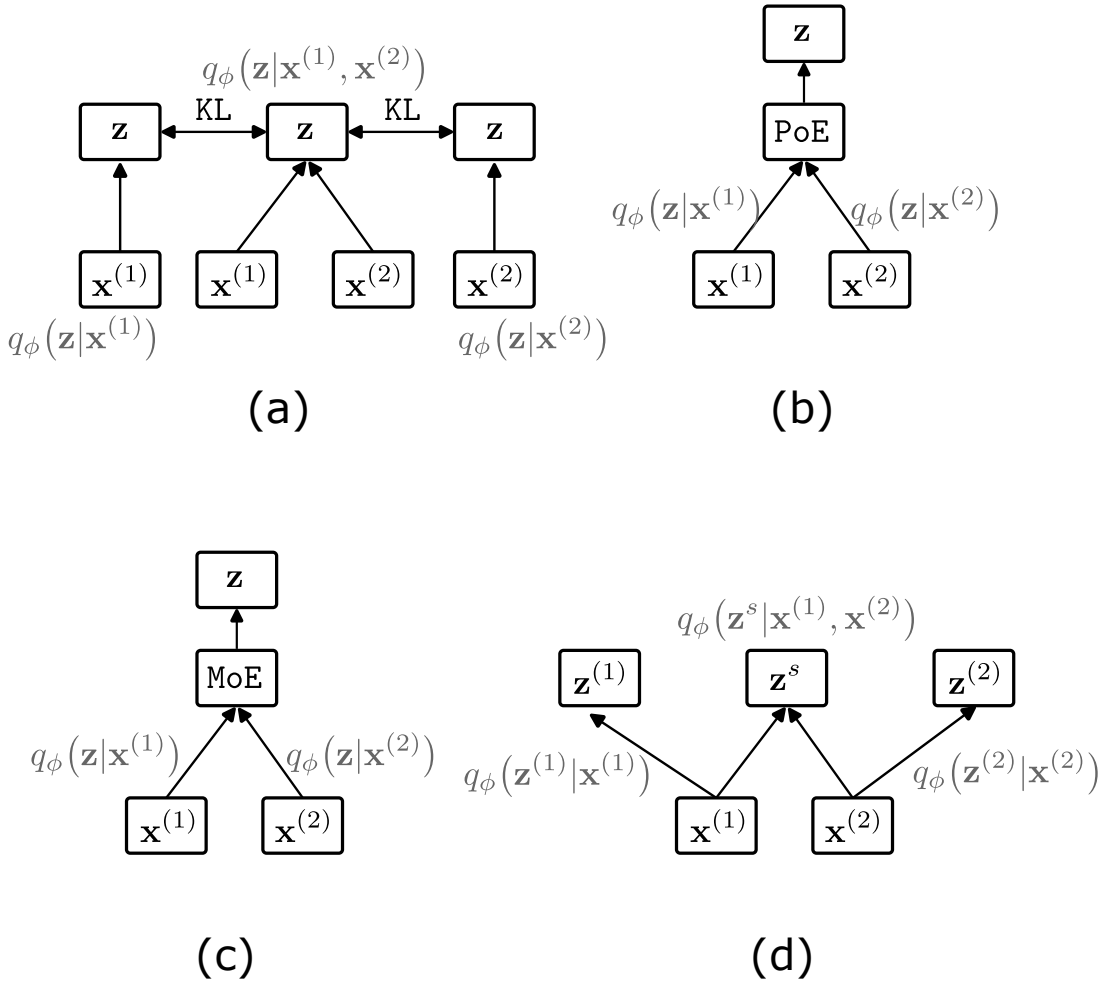


Figure 2.8 – A graphical model visualization of the encoder of (a) JointMVAE (Suzuki et al., 2016), (b) PoE-VAE (M. Wu & Goodman, 2018), (c) MoE-VAE (Shi et al., 2019), (d) PMVAE (W.-N. Hsu & Glass, 2018).

descriptions should result in more detailed and accurate image generation (and description generation) than considering only one modality.

Several MVAE models have been developed to satisfy these criteria. In the following discussion, as depicted in Figure 2.8, we introduce four of the most prominent ones (for a more extensive list, please refer to the comprehensive study by Suzuki and Matsuo, 2022):

Joint multimodal VAE (JointMVAE) JointMVAE leverages a joint representation shared among all modalities (Suzuki et al., 2016). Let us define $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ two modalities that are conditioned independently on the same latent vector $\mathbf{z}$; i.e., $p_\theta(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}|\mathbf{z}) = p_{\theta_1}(\mathbf{x}^{(1)}|\mathbf{z})p_{\theta_2}(\mathbf{x}^{(2)}|\mathbf{z})$, where θ_1, θ_2 represent the model parameters. Con-

sidering an approximate posterior distribution as $q_\phi(\mathbf{z}|\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$, we can estimate a lower bound of the log-likelihood $\log p(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$ as follows:

$$\begin{aligned}\mathcal{L}(\theta, \phi) = &\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{(1)},\mathbf{x}^{(2)})}\left(p_{\theta_1}(\mathbf{x}^{(1)}|\mathbf{z})\right) + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{(1)},\mathbf{x}^{(2)})}\left(p_{\theta_2}(\mathbf{x}^{(2)}|\mathbf{z})\right) \\ &- D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \parallel p(\mathbf{z})\right)\end{aligned} \tag{2.32}$$

When informative modality inputs (e.g., images) are missing from the input of a neural network-based inference distribution, the inferred representation is significantly corrupted (Suzuki et al., 2016). To accurately generate missing modalities from the available ones (cross generation), the authors introduce two additional networks for each modality, these additional encoders ($q_{\phi_1}(\mathbf{z}|\mathbf{x}^{(1)})$, $q_{\phi_2}(\mathbf{z}|\mathbf{x}^{(2)})$) are trained to match $q_\phi(\mathbf{z}|\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$. Therefore, the final ELBO of JointMVAE becomes:

$$\begin{aligned}&\mathcal{L}_{\text{JointMVAE}}(\theta, \phi) = \mathcal{L}(\theta, \phi) \\ &- \alpha\left[D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \parallel q_{\phi_1}(\mathbf{z}|\mathbf{x}^{(1)})\right) + D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \parallel q_{\phi_2}(\mathbf{z}|\mathbf{x}^{(2)})\right)\right].\end{aligned} \tag{2.33}$$

The authors establish that this objective is a lower bound for the variation of information $\mathbb{E}_{\hat{p}(\mathbf{x})}\left[p_\theta(\mathbf{x}_1|\mathbf{x}_2) + p_\theta(\mathbf{x}_2|\mathbf{x}_1)\right]$. In other words, the JointMVAE is optimized to promote cross-modal generation.

Product-of-experts VAE (PoE-VAE) The JointMVAE employs an ELBO objective (Equation 2.33) that incorporates two supplementary divergence terms to minimize the disparity between unimodal and multimodal posterior distributions. Unfortunately, the JointMVAE necessitates training a novel inference network for each multimodal subset. This implies the inclusion of 2^N additional encoders, denoted as $q_\phi(\mathbf{z}|\bar{\mathbf{x}})$, for every subset of modalities $\bar{\mathbf{x}} \subset \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}\}$. This is *intractable* in the general setting.

To tackle this challenge, M. Wu and Goodman, 2018 propose to model the joint posterior as a product of experts (PoE) over the marginal posteriors (Hinton, 2002), where they decompose the posterior distribution as $q_\theta(\mathbf{z}|\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}) \approx p(\mathbf{z}) \prod_{i=1}^{N} q_\theta(\mathbf{z}|\mathbf{x}^{(i)})$. When no modalities are observed, this posterior matches with the prior. As the number of modalities increases, the precision of this posterior distribution also increases, which is a result of the product property. In other words, this posterior becomes sharper with more observed modalities.

At the training stage, The ELBO is sub-sampled during each gradient step. This sub-sampling includes three specific scenarios: (i) optimizing the ELBO using the product

of all N Gaussians (in the case where all the modalities are available), (ii) optimizing all ELBO terms using a single modality, and (iii) optimizing k ELBO terms from k randomly chosen subsets denoted as $\bar{\mathbf{x}}_k$. This technique is called *sub-sampled training paradigm*.

Mixture-of-experts VAE The PoE factorization does not appear to be practically suited for multimodal learning, likely due to the precision miscalibration of experts (Shi et al., 2019). When PoE is applied, each expert wields significant influence over the joint distribution. If the inference for a particular modality is very sharp, the joint inference will be heavily dominated by it; therefore, the optimization of unimodal inference with low precision might be greatly degraded. Consequently, Shi et al., 2019 propose factorizing the joint variational posterior as a combination of unimodal posteriors, using a mixture of experts (MoE) as $q_\phi(\mathbf{z}|\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}) = \sum_{i=1}^{N} \alpha_i q_\phi(\mathbf{z}|\mathbf{x}^{(i)})$ where α_i is constrained to $\sum_i \alpha_i = 1$, and in many cases $\alpha_i = 1/N$. MoE does not face the issue of potentially overconfident experts. However, a drawback of the MoE approach is that combining the expertise of multiple components does not yield *a sharper* distribution than individual experts. Consequently, increasing the number of experts does not enhance the informativeness of the shared representation, unlike the PoE approach. This limitation hinders the ability to perform proper *aggregated inference*. To mitigate the trade-off between PoE and MoE, a generalization of PoE and MoE, called a mixture of products of experts (MoPoE), is introduced by Sutter et al., 2020.

Note on the limitations of MVAE A recent study by Daunhawer et al. (Daunhawer et al., 2021) has shed light on the limitations of methods like PoE-VAE, MoE-VAE, and MoPoE-VAE. While promising for multimodal learning, these approaches exhibit a notable generative quality gap compared to their unimodal VAE counterparts, which operate in an entirely unsupervised manner. To explain this disparity, the authors uncover a fundamental limitation that applies to a wide range of mixture-based multimodal VAEs. They establish that the sub-sampling of modalities imposes an undesirable upper bound on the multimodal ELBO, effectively constraining the generative capabilities of these models. Their empirical investigations find that none of the existing approaches can fully satisfy all the desired criteria for an effective multimodal generative model, especially when applied to complex datasets.

Partitioned multimodal VAE (PMVAE) The MVAE methods described above have been mainly focused on the scenario where the objective is to extract the shared explanatory factors while discarding the rest. W.-N. Hsu and Glass, 2018 investigate the task of discovering explanatory factors from multimodal sensory data, such as parallel images and speech recordings, resembling what humans perceive during learning. W.-N. Hsu and Glass, 2018 introduce PMVAE, a model to learn representations for not only the shared explanatory factors $\mathbf{s}$, but also the modality-dependent factors $\mathbf{z}^{(m)}$, and to encode them in different latent variables for *disentanglement* and interoperability.

The generative model for PMVAE is formulated as:

$$p_\theta(\mathbf{x}^{(1:N)}, \mathbf{z}^{(1:N)}, \mathbf{s}) = p_\theta(\mathbf{s}) \prod_{i=1}^{N} p_\theta(\mathbf{z}^{(i)}) p_\theta(\mathbf{x}^{(i)}|\mathbf{z}^{(i)}, \mathbf{s}) \tag{2.34}$$

Specifically, we assume the prior distributions over $\mathbf{s}$ and $\mathbf{z}^{(i)}, i \in [1, N]$ to be centered isotropic Gaussian with no trainable parameters. The conditional distribution of each modality ($p_\theta(\mathbf{x}^{(i)}|\mathbf{z}^{(i)}, \mathbf{s})$) is assumed to be a diagonal Gaussian, whose mean and variance are parameterized by DNN that take the corresponding latent variables as input.

Using the Bayesian network of the model, the chain rule of probabilities, and D-separation, it is possible to analyze how the observed and latent variables depend on each other in the exact posterior:

$$q_\phi(\mathbf{z}^{(1:N)}, \mathbf{s}|\mathbf{x}^{(1:N)}) = q_\phi(\mathbf{s}|\mathbf{x}^{(1:N)}) \prod_{i=1}^{N} q_\phi(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}, \mathbf{s}). \tag{2.35}$$

A variational lower bound on the log-likelihood is given as follows:

$$\begin{aligned} \mathcal{L}_{\text{PMVAE}}(\theta, \phi) = & \sum_{i=1}^{N} \left[\mathbb{E}_{q_\phi(\mathbf{z}^{(i)}, \mathbf{s}|\mathbf{x}^{(1:N)})} \left(p_\theta(\mathbf{x}^{(i)}|\mathbf{z}^{(i)}, \mathbf{s}) \right) - D_{KL} \left(q_\phi(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}, \mathbf{s} \parallel p_\theta(\mathbf{z}^{(i)}) \right) \right] \\ & - D_{KL} \left(q_\phi(\mathbf{s}|\mathbf{x}^{(1:N)}) \parallel p(\mathbf{s}) \right). \end{aligned} \tag{2.36}$$

In the original paper, the authors introduce two additional regularizations to Equation 2.36. The first, called *multimodal-unimodal coherence*, addresses scenarios where only a single modality is available. The second loss, named *cross-modality semantic contrastiveness*, encourages similarity when inferences of the latent semantic variable $\mathbf{s}$ are made by different modalities from the same sample and dissimilarity when inferences are drawn from different samples. While the latter encourages shared explanatory factors to be

captured by the latent semantic variable, the former ensures that non-shared factors are not inadvertently encoded within it.

2.4.8 Discrete variation autoencoder

Until now, our focus has been on VAEs with continuous latent spaces, represented as $\mathbf{z}$. Let us now shift our attention to models where the latent space is discrete, represented as $\mathbf{z}_q$, with $\dim(\mathbf{z}_q) = K$. Nevertheless, why should we be intrigued by learning discrete latent codes? When we examine latent variable models, one key objective is to find a more concise data representation, where $\mathbf{z}_q$ serves as a "compact" space that encapsulates essential information about the observations in $\mathbf{x}$. This concept of compression is often framed in terms of dimensionality, with VAEs typically featuring a latent space of significantly lower dimensionality than that of $\mathbf{x}$. However, while valid, this perspective needs to be revised in information theory. After all, storing an unlimited amount of information (as per Shannon's theory) in a single latent variable is theoretically possible. On the other hand, with a finite $\mathbf{z}_q$ space, we can establish a clear and precise understanding of compression, as the quantity of information represented by $\mathbf{z}_q$ is rigorously bounded by $\log_2(K)$ (Thickstun, 2020).

Since the ELBO does not impose any restrictions on the continuity of $\mathbf{z}$, we can employ this variational objective to optimize a discrete VAE. If $\mathbf{z}$ is discrete, it is natural to consider a uniform prior $p(\mathbf{z})$. The divergence term in Equation 4.11 is just the entropy:

$$D_{\mathrm{KL}}\Big(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\Big) = \log(K) - H\Big(q_\phi(\mathbf{z}|\mathbf{x})\Big) \tag{2.37}$$

Let us revisit the ELBO we previously derived:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big[\ln p_\theta(\mathbf{x}|\mathbf{z})\Big] + H(q_\phi(\mathbf{z}|\mathbf{x})) - \log(K) \tag{2.38}$$

Nonetheless, as explained in Section 2.4.4, obtaining gradients of Equation 2.38 concerning ϕ directly is unfeasible. In contrast to continuous latent variables, we cannot resolve this issue using the reparameterization trick. Several methods have been proposed to solve this problem (Jang et al., 2016; Mnih & Gregor, 2014; Razavi et al., 2019; Van den Oord et al., 2017).

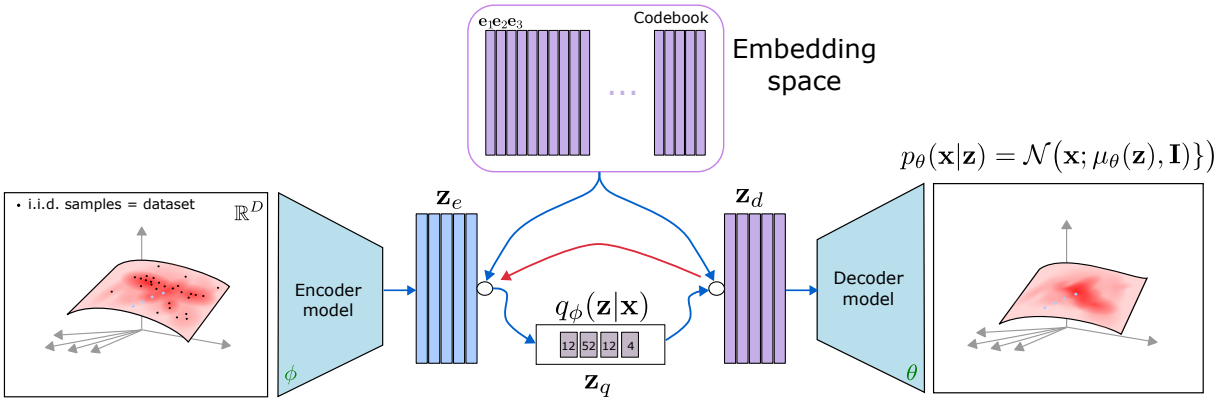


Figure 2.9 – A figure describing the VQ-VAE.

Vector quantization through the straight-through estimator Van den Oord et al., 2017 proposes a vector quantized variational autoencoder (VQ-VAE). The VQ-VAE employs discrete latent variables and is inspired by vector quantization for training. The posterior and prior distributions take the form of categorical distributions, and samples drawn from these distributions serve as indices for an embedding codebook. These embeddings, denoted by $\mathbf{e}$, serve as inputs to the decoder network. Because the quantization operation is non-differentiable, the authors applied the straight-through estimator (Bengio, Léonard, & Courville, 2013) to calculate the gradient, which involves directly copying gradients from the decoder input $\mathbf{z}_d$ to the encoder output $\mathbf{z}_e$. This means that the gradient from the first layer of the decoder is directly passed to the last layer of the encoder, skipping the codebook altogether. The total training objective of VQ-VAE becomes:

$$\mathcal{L}_{\text{VQ-VAE}}(\theta) = \underbrace{\log\left(p_\theta(\mathbf{x}|\mathbf{z}_d)\right)}_{(1)} + \underbrace{\|\mathtt{sg}(\mathbf{z}_e) - \mathbf{e}\|^2}_{(2)} + \underbrace{\beta\|\mathbf{z}_e - \mathtt{sg}(\mathbf{e}))\|^2}_{(3)} \tag{2.39}$$

where $\mathtt{sg}(.)$ represents the stop gradient operator. The first term (1) in Equation 2.39 corresponds to the reconstruction term (e.g., mean square error). This term is responsible for optimizing both the decoder and the encoder, using the straight-through estimator for gradient flow as illustrated in the red arrow in Figure 2.9. The second term (2) corresponds to the codebook loss. Since the embeddings ($\mathbf{e}$) do not receive gradients from the reconstruction loss, the authors employ vector quantization, serving as a straightforward dictionary learning algorithm. It helps in the learning process of the embedding space by adjusting each codebook vector to be nearer to the vector it replaces. Finally, the third term (3) is the "commitment loss", which addresses concerns related to the dimensionless nature

of the embedding space. Without this loss, the embedding space could potentially grow without bounds if the embeddings $\mathbf{e}_i$ do not train as rapidly as the encoder parameters.

By coupling these quantized representations with an autoregressive prior (e.g., PixelCNN (Van den Oord et al., 2016)), the model demonstrates the capability to produce high-quality images, videos, and speech. Additionally, it excels in tasks such as speaker conversion and the unsupervised learning of phonemes, offering compelling evidence for the effectiveness of the acquired representations.

Vector quantization through other approaches Various alternative methods have been introduced for training discrete VAE. The NVIL estimator employs a single-sample objective for optimizing the variational lower bound and incorporates several variance-reduction techniques to enhance training efficiency (Mnih & Gregor, 2014). In contrast, VIMCO optimizes a "multi-sample objective", accelerating convergence by leveraging multiple samples from the inference network (Mnih & Rezende, 2016). Jang et al., 2016 presented the *Gumbel-softmax trick*, which uses a continuous distribution with a temperature constant that can be gradually reduced during training to approach a discrete distribution in the limit. This approach can be viewed as an extension of the reparameterization trick tailored for categorical distributions.

2.5 Masked autoencoder

All the VAE-based techniques mentioned earlier, along with their subsequent variations and enhancements, operate within the framework of Bayesian probabilistic models and are trained using unsupervised methods. However, in the following section, we introduce a "masked autoencoder" model that abstains from the Bayesian probabilistic framework and employs a self-supervised approach instead. While the previous generative models centered on probabilistic principles, focusing on modeling the generative structure, inference, and training, this new approach primarily revolves around the paradigm of mask-based learning governed by a single loss, which has shown interesting results. Nevertheless, before tackling this approach, let us look at self-supervised methods.

2.5.1 A rapid tour of self-supervised learning methods

As already discussed in Subsection 2.2.5, SSL is a promising approach in machine learning that enables learning from vast amounts of unlabeled data. Unlike supervised

learning, which relies on labeled data, SSL defines pretext tasks based on unlabeled inputs to generate descriptive representations. Pretext tasks are essential for SSL. They create supervisory signals within unlabeled data, serving as proxies for the target tasks, and enabling meaningful representation learning. These tasks involve predicting data aspects, like spatial, temporal, or semantic properties, teaching the model to extract valuable features for downstream tasks. Examples of pretext tasks include information restoration, using temporal relationships, learning spatial context, grouping similar images together, etc.

SSL methods can be broadly categorized into two main categories: *contrastive learning* and *generative learning*. Each of these categories will be further elaborated in the following paragraphs.

Contrastive learning Contrastive learning is a SSL method that revolves around the multiview assumption property (see Subsection 2.2.5). Many modern SSL methods use contrastive learning to create feature representations invariant to simple transforms (X. Liu et al., 2021). The idea of contrastive learning is to encourage a model to represent two augmented versions of an input similarly. It encourages the model to bring similar samples close while separating dissimilar ones in the learned feature space. Positive pairs are augmented versions of the same data (e.g., an image with rotations and another with changes in brightness or audio clips of the same spoken sentence with varying noise or pitch) or across different modalities (e.g., an image with its corresponding audio clip or text description), while negative pairs match augmented samples with distinct dataset entries. Key loss functions like InfoNCE (Oord et al., 2018) or NT-Xent (T. Chen et al., 2020) guide the model to maximize similarity for positives and minimize it for negatives, promoting shared information capture and discriminative learning.

Contrastive learning encompasses several successful methods. **SimCLR** is a simple framework for contrastive learning of visual representations, which trains visual representations by promoting similarity between two augmented image views. These views are created through various transformations, like resizing, cropping, and random blurring. After encoding each view, SimCLR employs a projector, often implemented as a Multi-Layer Perceptron (MLP) followed by a Rectified Linear Unit (ReLU) activation. This mapping transforms the initial embeddings into a space where a contrastive loss is applied to encourage similarity between the views. For downstream tasks, extracting the representation before applying the projector has been demonstrated to enhance performance (T. Chen

et al., 2020). **BYOL** (Bootstrap Your Own Latent) introduces self-distillation to prevent collapse[6], using two networks and a predictor. The "online" or student network maps one network's outputs to the other's, while the other is the "target" or teacher network. Both networks receive different views of the same image through transformations, such as resizing, cropping, color adjustments, and brightness changes (Grill et al., 2020). Many other methods belong to this self-distillation family like SimSiam (L.-W. Chen & Rudnicky, 2021), MoCo (He et al., 2020), and DINO (Caron et al., 2021).

Generative learning Contrastive SSL has gained prominence as an effective alternative to supervised training. However, its effectiveness is now challenged by a generative SSL called masked modeling (C. Zhang et al., 2022).

Early attempts at masked autoencoding can be traced back to denoising autoencoders, which aimed to learn higher-level representations by filling in missing portions of images (Vincent et al., 2008). Variations of this approach, such as feature learning through inpainting and masked channel prediction, have shown promise, particularly in tasks like dense semantic segmentation (Larsson et al., 2016; R. Zhang et al., 2016).

The success of masked prediction in language models like GPT (proposed by OpenAI) and BERT (Devlin et al., 2019) sparked interest in applying it to image modeling. Models like iGPT (M. Chen et al., 2020) and iBERT (Dosovitskiy et al., 2020) demonstrated potential. However, their practicality was limited by high computational requirements and inferior performance compared to contrastive methods based on convolutional neural networks (CNNs) (T. Chen et al., 2020).

A significant breakthrough came with BEiT (Bao et al., 2021), which adopted autoencoder-based masked prediction in a novel way. BEiT leverages a discrete variational autoencoder to train an image tokenizer. BEiT outperformed the contrastive method DINO (Caron et al., 2021), highlighting the effectiveness of the masked modeling approach.

BEiT relies on a pretrained discrete variational autoencoder, making it a *non-end-to-end* solution. In contrast the masked autoencoder (MAE) (He et al., 2022) and SimMIM (Z. Xie et al., 2022) explore end-to-end training of masked autoencoders. MAE employs a transformer decoder, while SimMIM opts for a simpler single-layer decoder. MAE sets a new standard for self-supervised pretraining, particularly evident in its performance on the ImageNet-1K dataset (Deng et al., 2009). It outshines robust competitors like BEiT (Bao

6. i.e., the model converges to a state where it produces the same output or representation for all input data, regardless of the variations in the data

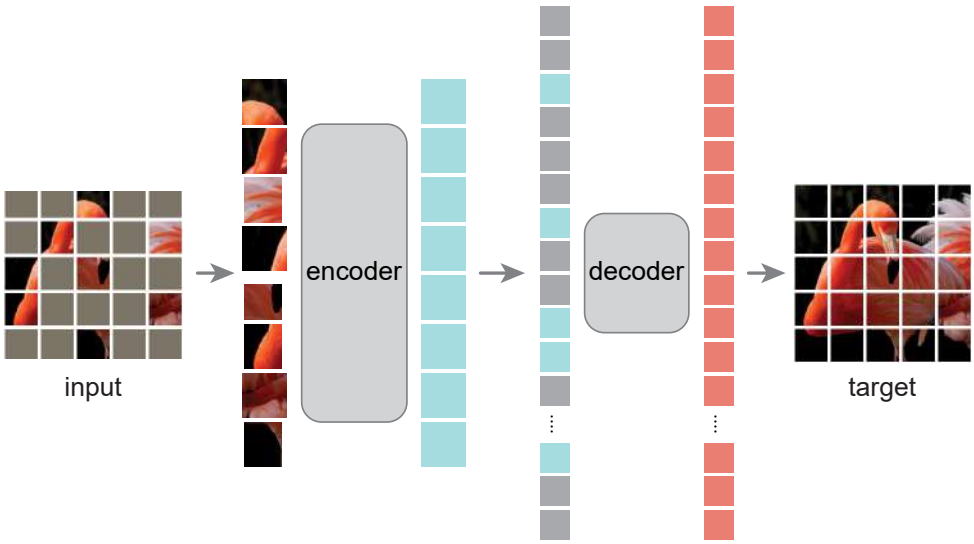


Figure 2.10 – (Image is taken from MAE article (He et al., 2022)) In the pre-training phase, a substantial portion of image patches, typically around 75%, is randomly masked out. The encoder operates solely on the smaller subset of visible patches (i.e., 25% of the image). Subsequently, mask tokens are introduced after the encoding step, and the complete set of encoded patches, along with these mask tokens, is passed through a compact decoder. This decoder's role is to reconstruct the original image at the pixel level.

et al., 2021) by a significant margin, all the while adopting a simpler and more efficient approach.

2.5.2 Understanding the masked autoencoder

MAE essentially functions as a denoising autoencoder (He et al., 2022), employing a straightforward approach where it randomly masks patches of the input image and then reconstructs the missing pixels, as illustrated in Figure 2.10. The MAE model is underpinned by two key architectural choices:

1. Asymmetric Encoder-Decoder Design: The encoder exclusively processes the visible patches, while the lightweight decoder is responsible for reconstructing the target image.
2. High Masking Ratio: An input image with a substantial masking ratio, often as high as 75%, is employed, resulting in a nontrivial and meaningful self-supervisory task.

Patchifying and masking In line with the methodology pioneered by the Vision Transformer (ViT) [16], MAE systematically divides an image into equally sized, non-

overlapping patches. Subsequently, MAE employs a patch sampling mechanism, selectively choosing a subset of these patches while masking (i.e., concealing) the remainder. This patch sampling strategy, termed "random sampling," is straightforward and effective.

Why and how does patchifying contribute to MAE? S. Cao et al., 2022 prove that the random patch selecting of MAE preserves the information of the original image while reducing the computing costs under common assumptions on the low-rank nature of images.

How does masking influence the learned representation? A recent study demonstrates the influence of the masking ratio in MAE (L. Kong et al., 2023). They show how key hyperparameters in MAE (the masking ratio and the patch size) determine which true latent variables to be recovered, therefore influencing the level of semantic information in the representation; masking too much or too little does not recover high-level representations from low-level features.

Does MAE benefit from other corruptions? As MAE functions as a denoising autoencoder, Tian et al., 2022 examines whether alternative image degradation techniques, aside from masking, are effective for visual pretraining. They investigated five methods: zoom-in, zoom-out, distortion, blurring, and de-colorizing. Their findings indicate that these methods outperform no pretraining, highlighting a unified denoising perspective behind MAE's success. Notably, blurring and de-colorizing are less effective than other methods involving spatial transformations, as they alter the image style from the pretext task to the downstream task. Among these techniques, zoom-in excels and complements masking to boost performance.

MAE encoder and decoder The MAE encoder is based on the ViT architecture [16]. However, it distinguishes itself by operating exclusively on visible, non-masked patches. Like a standard ViT, the encoder embeds these patches through a linear projection augmented by positional embeddings. Subsequently, it processes this set of embeddings using a sequence of Transformer blocks (Vaswani et al., 2017). However, a pivotal deviation sets the encoder apart. It exclusively functions on a smaller subset, typically encompassing around 25%, of the entire patch collection. This approach yields several benefits. Most notably, it allows us to train substantially large encoders while harnessing only a fraction of the computational resources and memory.

The MAE decoder takes as input the complete set of tokens, comprising (i) encoded visible patches and (ii) mask tokens (depicted in Figure 2.10). Each mask token represents a

shared and learned vector that signifies the presence of a missing patch awaiting prediction. MAE applies positional embeddings to all tokens within this comprehensive set, ensuring mask tokens contain information about their spatial location within the image. The decoder proceeds with additional Transformer blocks to perform its tasks.

Does MAE solely rely on adjacent neighbor patches to reconstruct each masked patch? According to S. Cao et al., 2022, MAE employs a global interpolation of latent representations for masked patches, determined by an inter-patch topology learned through its attention mechanism.

Loss function Regarding the loss function, inspired by BERT (Devlin et al., 2019), the MAE calculates the Mean Squared Error (MSE) between the pixel values of the reconstructed image and those of the original image. This loss is computed exclusively for the masked patches, focusing the learning process on the areas requiring reconstruction.

Remark. Using MSE losses for reconstruction can however result in a blurred image. As He et al. suggest (He et al., 2022), improving the quality of MAE predictions can potentially lead to better representations for downstream tasks.

Does MAE require extensive data? The conventional belief is that transfer learning benefits from pretraining on a considerably larger dataset than the target data. However, El-Nouby et al., 2021 challenges this idea by exploring if self-supervised pretraining on a smaller dataset can provide similar advantages. Their study is particularly noteworthy as it utilizes a ViT-based masked autoencoder, which, compared to CNN, generally demands more samples (Dosovitskiy et al., 2020). Interestingly, El-Nouby et al., 2021 demonstrates that pretraining a masked autoencoder BEiT (Bao et al., 2021) on just 1% of the ImageNet dataset (Deng et al., 2009) can achieve transfer performance comparable to full ImageNet pretraining. In contrast, previous methods like DINO (Caron et al., 2021) are more sensitive to data size and type. More recently, Z. Xie et al., 2023 conducted an extensive study on data scaling, ranging from 10% of ImageNet to full ImageNet-22K, with masked autoencoder models varying from 49 million to 1 billion parameters. This study suggests that masked image modeling requires larger datasets.

2.6 Conclusion of the chapter

This chapter delved into the representation learning methods, focusing on unsupervised and self-supervised techniques. As discussed in Section 2.3.2, the ultimate goal is to obtain learned representations that exhibit disentanglement (Higgins et al., 2018), meaning they should be inherently structured to capture independent and interpretable high-level data characteristics. A recent study by Van Steenkiste et al., 2019 highlights the advantages of solving complex downstream tasks using disentangled representations, showcasing improved sample efficiency (i.e., with limited training data), robustness, and generalization. This has spurred numerous studies to leverage these representations in the context of emotion recognition (Eskimez et al., 2018; Ong et al., 2022; H.-C. Yang & Lee, 2019; K. Yang et al., 2023). These motivations served as the guiding principles throughout my thesis work, leading to the following contributions:

- In the following chapter, we present a weakly supervised approach to tackle the issue discussed in Section 2.3.2 (identifiability). This method is implemented in the field of audio speech to disentangle pitch from formants within the source-filter model;
- In Chapter 4, we introduce a multimodal dynamical autoencoder designed to learn structured audiovisual representations across distinct latent spaces, including those related to joint vs. specific modality and static vs. dynamic information. One application of this model is audiovisual emotion recognition.
- In Chapter 5, we introduce a vector quantized MAE for audiovisual speech representation learning, applied to emotion recognition. Unlike the original MAE, which works with raw data (e.g., pixels in images), our approach takes compact and discrete representations obtained from two distinct vector quantized variational autoencoders as input.

Chapter 3

Larning and controlling the source-filter representation of speech with a variational autoencoder

Contents

Summary

This chapter introduces the first contribution of this thesis, which focuses on learning and controlling factors of variation in the VAE. The focus of this chapter shifts specifically to audio speech representation, leveraging the generative capabilities of the VAE model. The following key findings have been established. By examining the VAE's latent space, we have identified a link between the standard source-filter model of speech production and the learned representation. Specifically, we noticed that the fundamental and formant frequencies are encoded in *pseudo-orthogonal* latent subspaces. Leveraging this association, we introduce an approach for *generating* and *controlling* speech signals, managed through interpretable trajectories of f_0 and formant frequencies.

3.1 Introduction

In the previous chapter, we discussed how high-dimensional data, like natural images or speech signals, exhibit regularity, indicating the existence of a lower-dimensional latent representation from which the observed data is generated. As mentioned previously, representation learning aims to uncover this latent representation of complex data, and deep latent-variable generative models have emerged as promising unsupervised approaches (R. T. Chen et al., 2018; Goodfellow et al., 2014; Higgins et al., 2017a; H. Kim & Mnih, 2018; Kingma & Welling, 2014; Le Moine et al., 2021; Rezende et al., 2014). The VAE (Kingma & Welling, 2014; Rezende et al., 2014), equipped with both a generative and inference model, not only enables data generation but also facilitates analysis and transformation. Additionally, the VAE, as a learned probability density function (pdf), can serve as a powerful prior for solving inverse problems like compressed sensing (Bora et al., 2017), speech enhancement (Bando et al., 2018; Leglaive et al., 2018), and source separation (Jayaram & Thickstun, 2020; Kameoka et al., 2019). Understanding the learned latent representation in a VAE and controlling the underlying factors of variation in the data are key challenges in building more expressive and interpretable generative models and probabilistic priors.

A series of previous works on representation learning with deep generative models, in particular VAEs, have focused on images (R. T. Chen et al., 2018; Higgins et al., 2017a; H. Kim & Mnih, 2018; Locatello, Bauer, Lucic, et al., 2020; Locatello et al., 2019). Yet, it is not always easy to define the ground-truth latent factors of variation involved in generating natural images. Speech data exhibits a direct relationship between latent variation factors and speech production's anatomical mechanisms. The source-filter model, proposed by Fant (1970), explains that speech signals are generated through the interaction of a source signal with a linear filter. In voiced speech, the vibration of the vocal folds produces a quasi-periodic glottal sound wave, where the fundamental frequency (referred to as "pitch") plays a crucial role in speech prosody. Unvoiced speech, on the other hand, involves a noise source generated by turbulent airflow or acoustic impulses. The vocal tract modifies the source signal as a linear filter. The vocal tract's cavities create resonances known as "formants", characterized by their frequency, amplitude, and bandwidth. When individuals manipulate speech articulators such as the tongue, lips, and jaw, they modify the shape of their vocal tract, resulting in changes to the acoustic filter, associated resonances, and the resulting speech sounds. The power spectra and the spectral envelopes of two French vowels are displayed in Figure 3.1. The spectral envelopes show that the formant frequencies differ for the two vowels. In this example, however, the harmonic structure of the spectra shows that the fundamental frequency is the same for the two vowels. Formant frequencies are important distinctive features of vowels. In a first approximation, they can be related to the opening of the mouth, the front/rear position of the tongue, and the rounding of the lips for the first, second, and third formant, respectively. In the context of voiced phonemes, humans possess the ability to manipulate the formants independently from the pitch, indicating the capacity to adjust the filter without affecting the source (Fant, 1970), and independently from one another (MacDonald et al., 2011). According to the source-filter model, speech signals are primarily characterized by a few continuous latent factors of variation, representing the source (with fundamental frequency f_0 playing a central role) and the filter (predominantly described by the formants). The independence between the source and filter characteristics makes speech signals an intriguing domain for disentangled representation learning methods, particularly with deep generative latent-variable models such as the VAE.

In this chapter, we analyze and control the latent space of a VAE from the perspective of the source-filter model of speech production, which can be beneficial for various applications in speech analysis, transformation, and synthesis. Figure 3.2 shows an overview of the

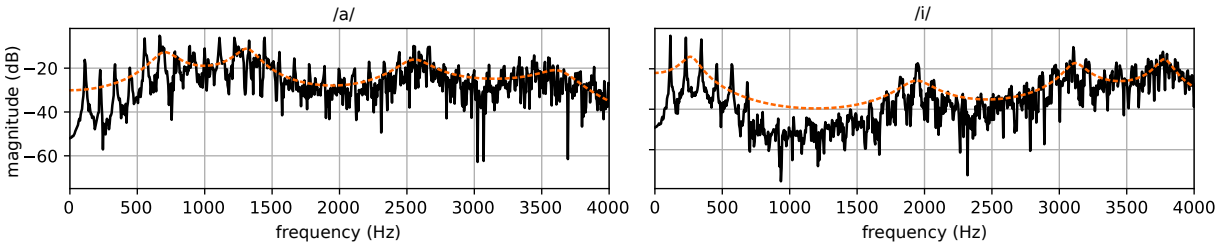


Figure 3.1 – Power spectrum (solid black line) and spectral envelop (orange dashed line) for two vowels uttered by a male speaker.

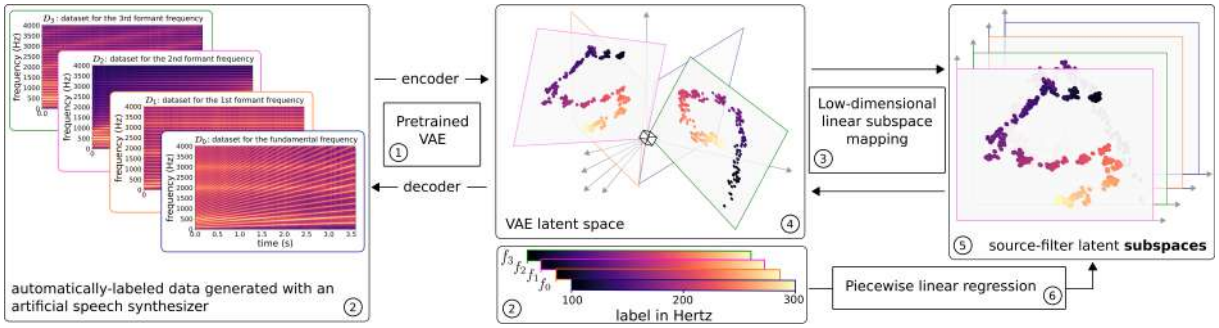


Figure 3.2 – Overview of the proposed method. First, ① a VAE is trained in an unsupervised manner by maximizing a lower bound of the data log-marginal likelihood (see Section 3.3.1) on a large dataset of unlabeled natural speech signals (not shown on this figure for clarity). Given the pretrained VAE and given ② a few seconds of automatically-labeled speech generated with an artificial speech synthesizer, we then propose ③ a linear subspace identification method to put in evidence that ④ the VAE latent space is structured into ⑤ orthogonal subspaces that encode f_0 and the formant frequencies, thus complying with the source-filter model of speech production. The subspaces are identified by minimizing the L2 norm of the reconstruction error obtained after passing the artificially-generated speech trajectories through the VAE encoder and projecting on the subspaces (see Section 3.3.2). Finally, we propose ⑥ a piecewise linear regression model to learn how to move into the source-filter latent subspaces, so as to perform speech manipulations in a disentangled manner. This model is also learned using the automatically-labeled artificial speech trajectories, by minimizing the L2 norm of the difference between the output of the regression model and the data coordinates in the previously-learned latent subspaces (see Section 3.3.4). No supervision is used to constrain the structure of the VAE latent space during its training. Supervision is only used after the training of the VAE, to identify the disentangled latent subspaces encoding the f_0 and formant frequencies, and to learn how to move into these subspaces to perform speech manipulations.

proposed approach. We first train a VAE on a dataset of about 25 hours of unlabeled speech signals. Then, using only a few seconds of automatically labeled speech signals generated with an artificial speech synthesizer, we propose a method to identify and

independently control the source-filter continuous latent factors of speech variation within the latent space of the VAE. Our contributions are the following: (i) We identify the source-filter model in the VAE latent space by showing experimentally that f_0 and the frequency of the first three formants, f_1, f_2, and f_3, are encoded in different subspaces. We put in evidence the orthogonality of the learned subspaces, which not only shows that the representation learned by the VAE complies with the source-filter model of speech production, but also suggests that we can perform speech transformations in a disentangled manner (i.e., modifying one of the factors would not affect the others) by moving into the learned subspaces. (ii) For each factor f_i, $i \in \{0, 1, 2, 3\}$, we propose to learn a piecewise linear regression model from the factor value in the synthetic speech dataset to the coordinates in the corresponding latent subspace. This method allows us to precisely and independently control the source-filter factors of speech variation within the learned subspaces, as confirmed experimentally on both artificial and natural signals. Without requiring additional information such as text or human-labeled data, the proposed approach leads to a deep generative model of speech spectrograms that is conditioned on f_0 and the formant frequencies. (iii) Finally, to illustrate the interest of the learned representation for downstream tasks, we propose an f_0 estimation method that exploits the projection of a speech signal onto the learned latent subspace associated with f_0. Experiments show that this approach competes with state-of-the-art methods in terms of precision and robustness to noise.

To the best of our knowledge, this is the first study *showing the link between the classical source-filter model of speech production and the representation learned in the latent space of a VAE*. Exploiting this link, we propose a principled method to generate and transform speech signals controlled with interpretable trajectories of f_0 and the formant frequencies. Regarding this latter application, our objective is not to compete with traditional signal processing methods (discussed in the next subsection), which remain state-of-the-art to the best of our knowledge. The present chapter's interest is to advance the understanding of deep generative modeling of speech signals while comparing fairly with traditional signal-model-based systems specifically designed for a given task. Moreover, advancing on the interpretability and control of the VAE latent space is expected to be beneficial for downstream tasks, for instance, to develop pitch-informed extensions of VAE-based speech enhancement methods such as those of Bando et al., 2018; Bie et al., 2022; Leglaive et al., 2018, 2020.

3.2 Related work

Time-scale, pitch-scale, and timbre modification of speech signals is a highly covered research problem originally addressed with signal processing methods. Three main groups of approaches exist (Laroche, 2002): time-domain methods such as the pitch-synchronous overlap and add (PSOLA) algorithm (Moulines & Charpentier, 1990), methods that work in the short-time Fourier transform (STFT) domain such as the phase vocoder (Flanagan & Golden, 1966; Laroche & Dolson, 1999), and parametric approaches based for instance on linear predictive coding (LPC) (Makhoul, 1975; Markel & Gray, 1976), sinusoidal modeling (George & Smith, 1997; McAulay & Quatieri, 1986), or sinusoidal plus noise modeling (Laroche et al., 1993; Serra & Smith, 1990). Other signal-processing-based approaches to real-time speech manipulations include the STRAIGHT (Banno et al., 2007; Kawahara, 2006) and WORLD (Morise et al., 2016) vocoders, which exploit a decomposition of the speech signal into f_0, spectral envelope, and aperiodicity.

Deep learning has recently emerged as a powerful approach to speech signal manipulation. A few methods have investigated combining traditional signal processing models with deep learning (Choi et al., 2021; Juvela et al., 2019; Lee et al., 2019; Valin & Skoglund, 2019; X. Wang et al., 2019). LPCNet is a successful neural vocoder inspired by the source-filter model (Valin & Skoglund, 2019). It was recently extended to pitch shifting and time stretching of speech signals by (Morrison et al., 2021). Yet, the authors showed that time-domain PSOLA (TD-PSOLA) (Moulines & Charpentier, 1990) remains a very strong baseline that is difficult to outperform with deep learning methods.

Regarding the use of deep generative models (in particular VAEs) for speech modeling and transformation, the studies of Akuzawa et al., 2018; Blaauw and Bonada, 2016; C.-C. Hsu et al., 2016; W.-N. Hsu et al., 2017a, 2017b are pioneering. Of particular interest to the present chapter is the work of (W.-N. Hsu et al., 2017a). The authors proposed using VAEs to modify the speaker identity and the phonemic content of speech signals by translations in the latent space of a VAE. Yet, this method requires knowing predefined values of the latent representations associated with both the source and target speech attributes to be modified. The method's performance thus depends on the quality of the estimation of the source attribute (e.g., f_0), which has to be obtained from the input speech signal to be transformed. This differs from the proposed method, which relies on projection onto the latent subspace associated with a given attribute and only requires the target value for this attribute. Moreover, W.-N. Hsu et al., 2017a did not address the

control of continuous factors of speech variation in the VAE latent space, contrary to the present work.

For deep latent representation learning methods, the challenge is to relate the learned representation to interpretable speech attributes. In Qian et al., 2020 and Webber et al., 2020, this interpretability is enforced by the design of the model. Qian et al., 2020 proposed to use three independent encoder networks to decompose a speech signal into f_0, timbre, and rhythm latent representations. Webber et al., 2020 focused on controlling source-filter parameters in speech signals, where the ability to control a given parameter (e.g., f_0) is enforced explicitly using labeled data and adversarial learning. In this approach, each parameter to be controlled requires dedicated training of the model. Moreover, these methods are speaker-dependent, as speech generation in Qian et al., 2020 is conditioned on the speaker identity, and Webber et al., 2020 used a single-speaker training dataset. This contrasts with the proposed method which is speaker-independent, and in which the source-filter representation is shown to emerge as orthogonal subspaces of the latent space of a single unsupervised VAE model.

In the machine learning and computer vision communities, variants of the VAE have recently led to considerable progress in disentangled representation learning (R. T. Chen et al., 2018; Higgins et al., 2017a; H. Kim & Mnih, 2018). From experimental analyses on image data, these methods suggest that a vanilla VAE cannot learn a disentangled representation. Moreover, Locatello et al., 2019 and Locatello, Bauer, Lucic, et al., 2020 recently showed both theoretically and from a large-scale experimental study that the unsupervised learning of disentangled representations is impossible without inductive biases (i.e., implicit or explicit assumptions by which a machine learning algorithm is able to generalize) on both the models and the data. Weakly-supervised (Hosoya, 2018; Locatello, Poole, et al., 2020; Shu et al., 2020) and semi-supervised (Locatello, Tschannen, et al., 2020; Sorrenson et al., 2020) methods have thus been proposed to learn disentangled representations. For example, the semi-supervised approach of Locatello, Tschannen, et al., 2020 exploits a small amount of labeled data to enforce the disentanglement of the representation at training time. This differs from the proposed approach where, after training a VAE on unlabeled natural speech signals, a few examples of artificially generated labeled speech data are used to identify the disentangled structure of the VAE latent representation in terms of source-filter factors of speech variation. This allows us to experimentally show that learning a disentangled source-filter representation of speech using a simple VAE is possible, complying with the definition of disentanglement proposed

in Higgins et al., 2018.

Several methods have been recently proposed to control continuous factors of variation in deep generative models (Goetschalckx et al., 2019; Härkönen et al., 2020; Jahanian et al., 2019; Plumerault et al., 2020), focusing essentially on generative adversarial networks (Goodfellow et al., 2014). They consist in identifying and then moving onto semantically meaningful directions in the latent space of the model. The present work is inspired by Plumerault et al., 2020, which assumes that a factor of variation can be predicted from the projection of the latent vector along a specific axis, learned from artificially generated trajectories. The proposed method is however more generic, thanks to the learning of latent subspaces associated with the latent factors and to the introduction of a general formalism based on the use of "biased" aggregated posteriors. Moreover, the previous works on controlling deep generative models only allow for moving "blindly" onto semantically meaningful directions in the latent space. In the present study, we are able to generate data conditioned on a specific target value for a given factor of variation (e.g., a given formant frequency value). Finally, previous works focused on image data. To the best of our knowledge, the present chapter proposes the first approach to identify and control source-filter factors of speech variation in a VAE.

The rest of this chapter is organized as follows: Section 3.3 presents the proposed method for analyzing and controlling source-filter factors of speech variation in a VAE. The method is evaluated experimentally and compared with traditional signal processing algorithms and with the approach of W.-N. Hsu et al., 2017a in Section 3.4. We finally conclude in Section 3.5.

3.3 Analyzing and controlling source-filter factors of speech variation in a VAE

In this section, we introduce the Itakura-Saito VAE model as our foundation (for a comprehensive understanding of VAEs, refer to Section 2.4, page 41). This model is trained on a large dataset of unlabeled natural speech signals. We then present our proposed method, which includes the following components: (i) The identification of latent subspaces that encode the source-filter factors of speech variation using a small amount of labeled speech signals generated artificially. (ii) An assessment of the disentanglement of the learned representation through a simple measurement strategy. (iii) A method to control the continuous factors of variation within the learned subspaces and generate corresponding

speech signals. (iv) A straightforward approach to estimate the f_0 contour of a speech signal by leveraging its projection onto the associated latent subspace.

3.3.1 Itakura-Saito variational autoencoder

As mentioned in Section 2.4, Page 41, generative modeling involves learning a probabilistic model of an observable random variable $\mathbf{x}$, which belongs to a subset $\mathcal{X}$ in $\mathbb{R}^D$. We consider a dataset $\mathcal{D} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ consisting of N independent and identically distributed (i.i.d.) observations of $\mathbf{x}$. Here, N represents the cardinality of $\mathcal{D}$. The empirical distribution of $\mathbf{x}$ is given by $\hat{p}(\mathbf{x}) = \frac{1}{N}\sum_{\mathbf{x}_n \in \mathcal{D}} \delta(\mathbf{x} - \mathbf{x}_n)$, where δ denotes the Dirac delta function. The Dirac delta function takes the value 1 only at 0 and is zero elsewhere.

VAE (Kingma & Welling, 2014; Rezende et al., 2014) is designed to approximate the empirical distribution $\hat{p}(\mathbf{x})$ with a parametric probability density function $p_\theta(\mathbf{x})$, where θ represents the model parameters. In the case of high-dimensional data like natural images or speech signals, the D dimensions of $\mathbf{x}$ exhibit regularity, indicating that they are not independent of each other. This suggests the presence of a lower-dimensional latent variable $\mathbf{z} \in \mathbb{R}^L$, where $L \ll D$, from which the observed data are generated. The VAE models the distribution by integrating over the joint distribution of the latent and observed variables, resulting in $p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$.

In this work, the observed data vector $\mathbf{x} \in \mathcal{X} = \mathbb{R}_+^D$ denotes the power spectrum of a short frame of speech signal (i.e., a column of STFT power spectrogram). Its entries are thus non-negative and its dimension D equals the number of frequency bins. We use the Itakura-Saito VAE (IS-VAE) (Bando et al., 2018; Girin et al., 2019b; Leglaive et al., 2018) defined by

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}), \tag{3.1}$$

$$p_\theta(\mathbf{x}|\mathbf{z}) = \prod_{d=1}^{D} \mathrm{Exp}\Big([\mathbf{x}]_d; [\mathbf{v}_\theta(\mathbf{z})]_d^{-1}\Big), \tag{3.2}$$

where $\mathcal{N}$ and Exp denote the densities of the multivariate Gaussian and univariate exponential distributions, respectively, and $[\mathbf{v}]_d$ denotes the d-th entry of $\mathbf{v}$. The inverse scale parameters of $p_\theta(\mathbf{x}|\mathbf{z})$ are provided by a neural network called the decoder, parametrized by θ and taking $\mathbf{z}$ as input.

The marginal likelihood $p_\theta(\mathbf{x})$ and the posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$ are intractable due to the nonlinearities of the decoder, so it is necessary to introduce an inference model

$q_\phi(\mathbf{z}|\mathbf{x}) \approx p_\theta(\mathbf{z}|\mathbf{x})$, which is defined by

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mathbf{z}; \boldsymbol{\mu}_\phi(\mathbf{x}), \mathrm{diag}\{\mathbf{v}_\phi(\mathbf{x})\}\right), \tag{3.3}$$

where the mean and variance parameters are provided by a neural network called the encoder network, parametrized by ϕ and taking $\mathbf{x}$ as input. Then, the VAE training consists in maximizing a lower-bound of $\ln p_\theta(\mathbf{x})$, called the evidence lower-bound (ELBO) and defined by

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{\hat{p}(\mathbf{x})}\Big[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big[p_\theta(\mathbf{x}|\mathbf{z})\Big] - D_{\mathrm{KL}}\Big(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\Big)\Big], \tag{3.4}$$

During training, the generative and inference model parameters θ and ϕ are jointly estimated by maximizing the ELBO, using (variants of) stochastic gradient descent with the so-called reparameterization trick (Kingma & Welling, 2014; Rezende et al., 2014).

3.3.2 Learning source-filter latent subspaces

In addition to the pre-trained IS-VAE speech spectrogram model introduced in the previous subsection, we also assume the availability of an artificial speech synthesizer allowing for an accurate and independent control of f_0 and the formant frequencies. We use Soundgen (Anikin, 2019), a parametric synthesizer based on the source-filter model of speech production. For a given speech sound, the voiced component of the source signal is generated by a sum of sine waves, the noise component by a filtered white noise, and both components are then summed and passed through a linear filter simulating the effect of the human vocal tract. Importantly, this synthesizer allows us to easily generate artificial speech data labeled with f_0 and formant frequency values.

Formally, let f_i denote the speech factor of variation (in Hz) corresponding to the fundamental frequency, for $i = 0$, and to the formant frequencies, for $i \in \{1, 2, 3\}$. Let $\mathcal{D}_i$ denote a dataset of artificially-generated speech vectors (more precisely short-term power spectra) synthesized by varying only f_i, all other factors $\{f_j, j \neq i\}$ being arbitrarily fixed. All examples in $\mathcal{D}_i$ are labeled with the index and the value of the factor of variation. It would be relatively difficult to build such a dataset from existing corpora of unlabeled natural speech. In contrast, it is a very easy task using an artificial speech synthesizer such as Soundgen (Anikin, 2019), which precisely takes f_0 and the formant parameters as input, and outputs waveforms from which we extract power spectra.

Let $\hat{p}^{(i)}(\mathbf{x})$ denote the empirical distribution associated with $\mathcal{D}_i$, defined similarly as $\hat{p}(\mathbf{x})$. We also introduce the following marginal distribution over the latent vectors:

$$\hat{q}_\phi^{(i)}(\mathbf{z}) = \int q_\phi(\mathbf{z}|\mathbf{x})\hat{p}^{(i)}(\mathbf{x})d\mathbf{x} = \frac{1}{\#\mathcal{D}_i}\sum_{\mathbf{x}_n\in\mathcal{D}_i} q_\phi(\mathbf{z}|\mathbf{x}_n). \tag{3.5}$$

In the literature, this quantity is referred to as the aggregated posterior (Makhzani et al., 2015), and its introduction can be found in Section 2.4.5 (page 51). However, $q_\phi(\mathbf{z}|\mathbf{x})$ is usually aggregated over the empirical distribution $\hat{p}(\mathbf{x})$ such that the aggregated posterior is expected to match with the prior $p(\mathbf{z})$ (R. T. Chen et al., 2018; Dai & Wipf, 2018). In contrast, in 3.5 we aggregate over the "biased" data distribution $\hat{p}^{(i)}(\mathbf{x})$, where we know only one latent factor varies. This defines the explicit inductive bias (Locatello et al., 2019) that we exploit to learn the latent source-filter representation of speech in the VAE.

In the following of the chapter, without loss of generality, we assume that, for each data vector in $\mathcal{D}_i$, the associated latent vector $\mathbf{z}$ has been centered by subtracting the mean vector

$$\boldsymbol{\mu}_\phi(\mathcal{D}_i) = \mathbb{E}_{\hat{q}_\phi^{(i)}(\mathbf{z})}[\mathbf{z}] = \frac{1}{\#\mathcal{D}_i}\sum_{\mathbf{x}_n\in\mathcal{D}_i} \boldsymbol{\mu}_\phi(\mathbf{x}_n). \tag{3.6}$$

Because only one factor varies in $\mathcal{D}_i$, we expect latent vectors drawn from the "biased" aggregated posterior in 3.5 to live on a low-dimensional manifold embedded in the original latent space $\mathbb{R}^L$. We assume this manifold to be a subspace characterized by its semi-orthogonal basis matrix $\mathbf{U}_i \in \mathbb{R}^{L\times M_i}$, $1 \leq M_i < L$. This matrix is computed by solving the following optimization problem:

$$\min_{\mathbf{U}\in\mathbb{R}^{L\times M_i}} \mathbb{E}_{\hat{q}_\phi^{(i)}(\mathbf{z})}\left[\left\|\mathbf{z} - \mathbf{U}\mathbf{U}^\top\mathbf{z}\right\|_2^2\right], \quad s.t.\ \mathbf{U}^\top\mathbf{U} = \mathbf{I}. \tag{3.7}$$

The space spanned by the columns of $\mathbf{U}_i$ is a subspace of the original latent space $\mathbb{R}^L$ in which the latent vectors associated with the variation of the factor f_i in $\mathcal{D}_i$ are expected to live.

Using Equation 3.5, the fact that $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$, and the inference model (centred version) $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_\phi(\mathbf{x}) - \boldsymbol{\mu}_\phi(\mathcal{D}_i), \mathrm{diag}\{\mathbf{v}_\phi(\mathbf{x})\})$, the cost function in the optimization

problem 3.7 can be rewritten as follows:

$$\begin{aligned}
&\mathbb{E}_{\hat{q}_\phi^{(i)}(\mathbf{z})}\left[\left\|\mathbf{z}-\mathbf{U}\mathbf{U}^\top\mathbf{z}\right\|_2^2\right] \\
&= \frac{1}{\#\mathcal{D}_i}\sum_{\mathbf{x}_n\in\mathcal{D}_i}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_n)}\left[\left\|\mathbf{z}-\mathbf{U}\mathbf{U}^\top\mathbf{z}\right\|_2^2\right] \\
&= \mathrm{tr}\left\{(\mathbf{I}-\mathbf{U}\mathbf{U}^\top)\frac{1}{\#\mathcal{D}_i}\sum_{\mathbf{x}_n\in\mathcal{D}_i}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}_n)}[\mathbf{z}\mathbf{z}^\top]\right\} \\
&= \mathrm{tr}\left\{(\mathbf{I}-\mathbf{U}\mathbf{U}^\top)\mathbf{S}_\phi(\mathcal{D}_i)\right\},
\end{aligned} \tag{3.8}$$

where $\mathbf{S}_\phi(\mathcal{D}_i)$ is defined as follows:

$$\mathbf{S}_\phi(\mathcal{D}_i) = \frac{1}{\#\mathcal{D}_i}\sum_{\mathbf{x}_n\in\mathcal{D}_i}\left[\boldsymbol{\mu}_\phi(\mathbf{x}_n)\boldsymbol{\mu}_\phi(\mathbf{x}_n)^\top + \mathrm{diag}\{\mathbf{v}_\phi(\mathbf{x}_n)\}\right] - \boldsymbol{\mu}_\phi(\mathcal{D}_i)\boldsymbol{\mu}_\phi(\mathcal{D}_i)^\top. \tag{3.9}$$

From the equality 3.8, we see that the optimization problem 3.7 is equivalent to

$$\max_{\mathbf{U}\in\mathbb{R}^{L\times M_i}} \mathrm{tr}\left\{\mathbf{U}^\top\mathbf{S}_\phi(\mathcal{D}_i)\mathbf{U}\right\}, \qquad s.t.\ \mathbf{U}^\top\mathbf{U}=\mathbf{I}. \tag{3.10}$$

Very similarly to PCA (Pearson, 1901), the solution to the problem 3.7 is given by the M_i dominant eigenvectors of $\mathbf{S}_\phi(\mathcal{D}_i)$ (i.e., associated to the M_i largest eigenvalues) (Bishop, 2006b, Section 12.1). The dimension M_i of the subspace can be chosen such as to retain a certain percentage of the data variance in the latent space. Note that the only source of supervision used here is the knowledge that only the factor f_i varies in the dataset $\mathcal{D}_i$.

3.3.3 Disentanglement analysis of the latent representation

As defined by Higgins et al., 2018, a representation is disentangled if it is possible to learn orthogonal latent subspaces associated with each factor of variation, whether they are single- or multi-dimensional. The approach presented in the previous subsection exactly follows this definition and offers a natural and straightforward way to objectively measure if the unsupervised VAE managed to learn a disentangled representation of the factors of variation under consideration. First, by simply looking at the eigenvalues associated with the columns of $\mathbf{U}_i \in \mathbb{R}^{L\times M_i}$, we can measure the amount of variance that is retained by the projection $\mathbf{U}_i\mathbf{U}_i^\top$. If a small number of components M_i represents most of the variance, it indicates that only a few intrinsic dimensions of the latent space are dedicated to the factor of variation f_i and varying this factor can be done by affine transformations.

Second, if for two different factors of variation f_i and f_j, with $i \neq j$, the columns of $\mathbf{U}_i$ are orthogonal to those of $\mathbf{U}_j$, this indicates that the two factors are encoded in orthogonal subspaces and therefore disentangled. It should however be verified experimentally that applying transformations by moving onto the subspace associated with f_i generalizes to values of $\{f_j, j \neq i\}$ different than the ones used in $\mathcal{D}_i$.

3.3.4 Controlling the source-filter factors of variation

So far, for each factor f_i, we have defined a methodology to learn a latent subspace $\mathbf{U}_i \in \mathbb{R}^{L \times M_i}$ that encodes its variations in the dataset $\mathcal{D}_i$, containing a few examples of speech data generated by an artificial synthesizer. Making now use of the values of the factor f_i for the data in $\mathcal{D}_i$, we learn a regression model $\mathbf{g}_{\eta_i} : \mathbb{R}_+ \mapsto \mathbb{R}^{M_i}$ from f_i, whose value is denoted by $y \in \mathbb{R}_+$, to the data coordinates in the latent subspace defined by $\mathbf{U}_i$. The parameters η_i are thus defined as the solution of the following optimization problem:

$$\min_{\eta} \left\{ \mathbb{E}_{\hat{q}_\phi^{(i)}(\mathbf{z},y)} \left[\left\| \mathbf{g}_\eta(y) - \mathbf{U}_i^\top \mathbf{z} \right\|_2^2 \right] \stackrel{c}{=} \frac{1}{\#\mathcal{D}_i} \sum_{(\mathbf{x}_n, y_n) \in \mathcal{D}_i} \left\| \mathbf{g}_\eta(y_n) - \mathbf{U}_i^\top \Big(\boldsymbol{\mu}_\phi(\mathbf{x}_n) - \boldsymbol{\mu}_\phi(\mathcal{D}_i) \Big) \right\|_2^2 \right\}, \tag{3.11}$$

where $\hat{q}_\phi^{(i)}(\mathbf{z}, y) = \int q_\phi(\mathbf{z}|\mathbf{x}) \hat{p}^{(i)}(\mathbf{x}, y) d\mathbf{x}$, $\hat{p}^{(i)}(\mathbf{x}, y)$ is the empirical distribution associated with $\mathcal{D}_i$, considering now both the speech data vector $\mathbf{x}$ and the value y of f_i, and $\stackrel{c}{=}$ denotes equality up to an additive constant w.r.t. η. This approach can be seen as a probabilistic extension of principal component regression (Hotelling, 1957; Kendall, 1957). For simplicity and because it revealed efficient for this task, $\mathbf{g}_{\eta_i}$ is chosen as a piece-wise linear regression model learned independently for each output coordinate $m \in \{1, ..., M_i\}$. This choice is supported by the fact that the semi-orthogonal matrix $\mathbf{U}_i$ decorrelates the data (Bengio, Courville, & Vincent, 2013). Solving the optimization problem 3.11 then consists in solving a linear system of equations. In this work, we used the Python library of Jekel and Venter, 2019. It is important to remind that even if the regression model is supervised, the labeled dataset $\mathcal{D}_i$ is very small with only a few seconds of speech signals (see experimental setup details in Appendix A.1), it is synthetic, and the values of f_i are automatically obtained during the generation of the data with an artificial speech synthesizer, so no manual annotation effort is required.

We can now transform a speech spectrogram by analyzing it with the VAE encoder, then linearly moving in the learned subspaces using the above regression model, and finally

resynthesizing it with the VAE decoder. Given a source latent vector $\mathbf{z}$ and a target value y for the factor f_i, we apply the following affine transformation:

$$\tilde{\mathbf{z}} = \mathbf{z} - \mathbf{U}_i\mathbf{U}_i^\top\mathbf{z} + \mathbf{U}_i\mathbf{g}_{\eta_i}(y). \tag{3.12}$$

This transformation consists in (i) subtracting the projection of $\mathbf{z}$ onto the subspace associated with the factor of variation f_i; and (ii) adding the target component provided by the regression model $\mathbf{g}_{\eta_i}$ mapped from the learned subspace to the original latent space by the matrix $\mathbf{U}_i$. This operation allows us to move only in the latent subspace associated with the factor f_i. If this subspace is orthogonal to the latent subspaces associated with the other factors $\{f_j, j \neq i\}$, the latter should remain the same between $\mathbf{z}$ and $\tilde{\mathbf{z}}$, only f_i should be modified. This process can be straightforwardly generalized to multiple factors, by subtracting and adding terms corresponding to each one of them. Contrary to W.-N. Hsu et al., 2017a, the operation in 3.12 does not require the knowledge of the factor f_i value associated with the source vector $\mathbf{z}$, it only requires the value y of the factor f_i corresponding to the target vector $\tilde{\mathbf{z}}$ (this value y being used as input to the regression model.)

Finally, as the prior $p(\mathbf{z})$ and inference model $q_\phi(\mathbf{z}|\mathbf{x})$ are Gaussian (see Equations 3.2 and 3.3), the transformation in 3.12 has the following probabilistic formulation (using $\mathbf{U}_i^\top\mathbf{U}_i = \mathbf{I}$):

$$p(\tilde{\mathbf{z}}; f_i = y) = \mathcal{N}\Big(\tilde{\mathbf{z}}; \mathbf{U}_i\mathbf{g}_{\eta_i}(y), \mathbf{M}_i\Big) \tag{3.13}$$

$$q_\phi(\tilde{\mathbf{z}}|\mathbf{x}; f_i = y) = \mathcal{N}\Big(\tilde{\mathbf{z}}; \mathbf{U}_i\mathbf{g}_{\eta_i}(y) + \mathbf{M}_i\boldsymbol{\mu}_\phi(\mathbf{x}), \mathbf{M}_i\text{diag}\{\mathbf{v}_\phi(\mathbf{x})\}\Big), \tag{3.14}$$

where $\mathbf{M}_i = \mathbf{I} - \mathbf{U}_i\mathbf{U}_i^\top$. The prior in 3.13 is now conditioned on the factor f_i and can be used to generate speech data given input trajectories of f_0 and formant frequencies. As we assumed centered latent data, the mean vector $\boldsymbol{\mu}_\phi(\mathcal{D}_i)$ defined in Equation 3.6 must be added to $\tilde{\mathbf{z}}$ before mapping this vector through the generative model $p_\theta(\mathbf{x}|\mathbf{z})$.

3.3.5 Estimating the fundamental frequency using the learned latent representation

To illustrate the interest of the learned representation on an analysis task, we propose to estimate the f_0 contour of a speech signal using its projection onto the corresponding

latent subspace characterized by the estimated matrix $\mathbf{U}_0 \in \mathbb{R}^{L \times M_0}$ (cf. Section 3.3.2). As we focus on the analysis of f_0, in this subsection we assume that the latent vectors are centered by subtracting the mean vector $\boldsymbol{\mu}_\phi(\mathcal{D}_0)$ defined in Equation 3.6: $\mathbf{z} \leftarrow \mathbf{z} - \boldsymbol{\mu}_\phi(\mathcal{D}_0)$. Let $\mathbf{p} = \mathbf{U}_0^\top \mathbf{z} \in \mathbb{R}^{M_0}$ denote the projection of $\mathbf{z}$ onto the f_0 latent subspace[1]. Because $\mathbf{p}$ results from a linear transformation of $\mathbf{z}$ and the approximate posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$ defined in Equation 3.3 is Gaussian, we have:

$$q_\phi(\mathbf{p}|\mathbf{x}) = \mathcal{N}\left(\mathbf{p}; \mathbf{U}_0^\top \boldsymbol{\mu}_\phi(\mathbf{x}), \mathbf{U}_0^\top \mathrm{diag}\{\mathbf{v}_\phi(\mathbf{x})\}\mathbf{U}_0\right). \tag{3.15}$$

As will be confirmed experimentally in Section 3.4, the subspace generated by $\mathbf{U}_0$ encodes the fundamental frequency information, and the formant frequencies are encoded in other orthogonal subspaces. Therefore, the projection of $\mathbf{z}$ onto $\mathbf{U}_0$ is expected to provide invariance to a change of the formant frequencies, which is an appealing feature for estimating the fundamental frequency. The method we propose is simple but effective, as will be shown experimentally. For an input speech power spectrum $\mathbf{x}^{\text{test}}$ assumed to be voiced, the estimated fundamental frequency is given by the value y of f_0 associated with $\mathbf{x}^\star \in \mathcal{D}_0$ defined by

$$\mathbf{x}^\star = \mathrm{argmin}_{\mathbf{x} \in \mathcal{D}_0} D_{\mathrm{KL}}\left(q_\phi(\mathbf{p}|\mathbf{x}^{\text{test}}) \parallel q_\phi(\mathbf{p}|\mathbf{x})\right). \tag{3.16}$$

Using the KL divergence allows us to base the estimation of f_0 on the full distribution of the projection, i.e. taking not only the mean of the projection in Equation 3.15 into account but also the covariance. The proposed method requires computing the KL divergence between two multivariate Gaussians, which admits a closed-form expression. We can thus simply compute the KL divergence in Equation 3.16 numerically for all the examples $\mathbf{x}$ in the synthetic labeled dataset $\mathcal{D}_0$ and return the value y of f_0 associated with the minimum.

The f_0 estimation is done independently for each frame of the power spectrogram of an input speech signal. The resulting "raw" estimated f_0 trajectory is then smoothed by applying a median filter with a window size of 5 frames. Above, we assumed $\mathbf{x}^{\text{test}}$ was a voiced speech spectrum. The voiced/unvoiced detection can be made automatically by setting a threshold on the minimum value of the above KL divergence, i.e. $D_{\mathrm{KL}}\left(q_\phi(\mathbf{p}|\mathbf{x}^{\text{test}}) \parallel q_\phi(\mathbf{p}|\mathbf{x}^\star)\right)$.

1. The term projection used to refer to $\mathbf{p} \in \mathbb{R}^{M_0}$ is a misuse of language. Strictly speaking, the projection of $\mathbf{z} \in \mathbb{R}^L$ onto the subspace characterized by $\mathbf{U}_0 \in \mathbb{R}^{L \times M_0}$ is given by $\mathbf{U}_0\mathbf{U}_0^\top \mathbf{z} \in \mathbb{R}^L$.

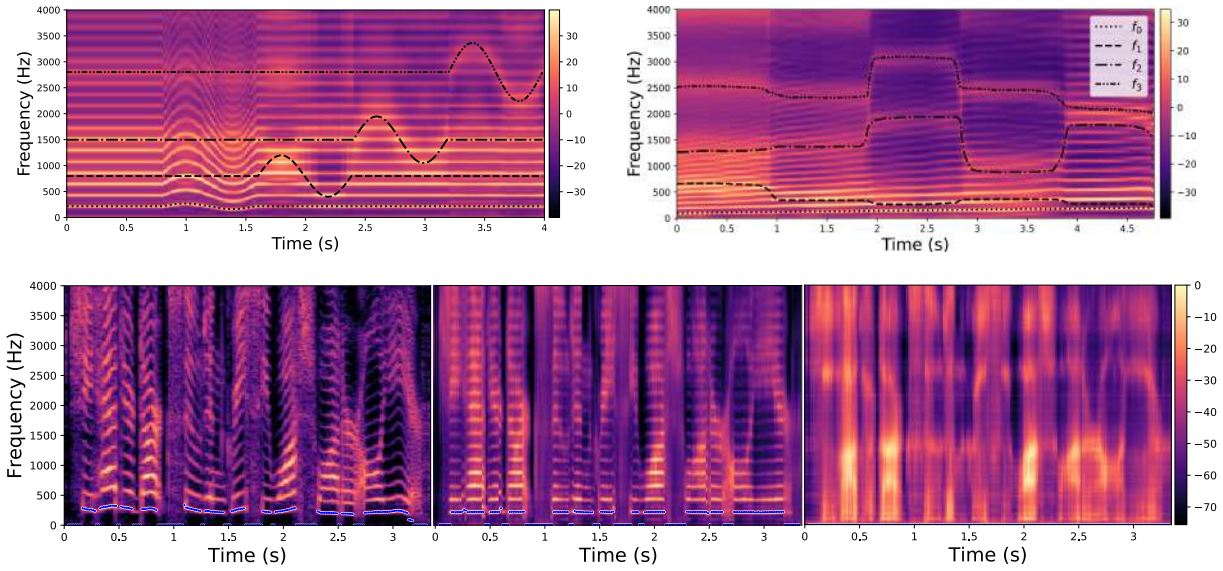


Figure 3.3 – Examples of spectrograms modified and generated with the proposed method. The color bar indicates the power in dB. Top left: f_0 and formant transformations of a vowel /a/ uttered by a female speaker. Top right: Spectrogram generated from input trajectories of f_0 and formant frequencies. The target values of the factors f_i are indicated by the black lines. Bottom left: Original spectrogram of a speech signal uttered by a female speaker; Bottom middle: Transformed spectrogram with f_0 (blue line) set constant over time; Bottom right: Transformed spectrogram where the original voiced speech signal (bottom left) is converted into a whispered speech signal (i.e., the pitch is removed).

3.4 Experiments

This section presents qualitative and quantitative experimental results obtained with the proposed VAE-based method for controlling f_0 and the formant frequencies of speech signals. The VAE is trained on about 25 hours of multi-speaker speech data from the Wall Street Journal (WSJ0) dataset (Garofalo et al., 1993). The data space dimension is 513 and the latent space dimension is set to 16. This dimension was chosen based on previous work showing it is optimal for the modeling of speech power spectra in the context of speech enhancement (Leglaive et al., 2018, 2019a; Sekiguchi et al., 2019). For a given factor of variation, the corresponding latent subspace is learned (see Section 3.3.2) using short trajectories of speech power spectra (corresponding to a few seconds of speech) generated with Soundgen (Anikin, 2019), all other factors being arbitrarily fixed. When solving the optimization problem 3.7, the latent subspace

dimension M_i of each factor of variation is chosen such that 80% of the data variance is retained. This leads $M_0 = 4$, $M_1 = 1$ and $M_2 = M_3 = 3$. The regression models used to control the speech factors of variation in the latent space (see Section 3.3.4) are learned on the same trajectories, but using the values of the Soundgen input control parameters (i.e., f_0 and formant frequencies values). More details on the experimental set-up can be found in A.1. Given a generated or transformed spectrogram, we use Waveglow (Prenger et al., 2019) to reconstruct the time-domain signal.

3.4.1 Qualitative results

In Figure 3.3, we illustrate the ability of the proposed method to modify f_0 and the formant frequencies in an accurate and independent manner. The top-left spectrogram contains five segments of equal length. The first segment corresponds to the original spectrogram of the steady vowel /a/ uttered by a female speaker. In the following segments, we vary successively each individual factor f_i, for $i = 0$ to 3, as indicated by the black lines in the figure. Variations of f_0 modify the harmonic structure of the signal while keeping the formant structure unaltered. Variations of f_i, $i \in \{1, 2, 3\}$, modify the formant frequencies, as indicated by the color map, while keeping f_0 unaltered.

The top-right spectrogram in Figure 3.3 was generated by using the conditional prior in 3.13 (generalized to conditioning on multiple factors). We can see that the characteristics of the generated speech spectrogram match well with the input trajectories represented by the lines in the figure.

In the second row of Figure 3.3, from left to right we show the original spectrogram of a speech signal uttered by a female speaker (left), the transformed spectrogram where f_0 is set constant over time (middle), and the transformed spectrogram where the pitch has been removed (i.e., the original voiced speech signal is converted into a whispered speech signal) (right). This last spectrogram is obtained by subtracting to $\mathbf{z}$ its projection onto the latent subspace corresponding to f_0 (i.e., by considering only the two first terms in the right-hand side of 3.12). This results in a spectrogram where the harmonic component is neutralized, while preserving the original formant structure. This is remarkable considering that the VAE was not trained on whispered speech signals, and it further confirms that the proposed method dissociates the source and the filter contributions in the VAE latent space.

Audio examples and additional examples of generated and transformed speech spec-

trograms can be found online [2] or in Appendix A.3. In Subsection 3.4.2, we provide plots of trajectories in the learned latent subspaces, illustrating that, according to each factor, the proximity of two speech spectra is preserved in the corresponding latent subspace. A graphical user interface to control audio speech has also been developed (see Appendix D.1).

3.4.2 Visualization of the learned latent subspaces

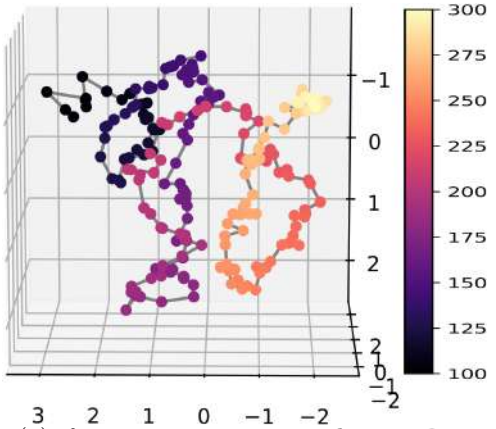
Table 3.1 – Cumulative variance (in %) retained by the projection $\mathbf{U}_i\mathbf{U}_i^\top$, $\mathbf{U}_i \in \mathbb{R}^{L\times M_i}$, as a function of the number of components M_i. We keep as much components as needed to retain at least 80 % of the data variance, as indicated by the underlined numbers.

	f_0	f_1	f_2	f_3
$M_i = 1$ component	33	<u>81</u>	39	48
$M_i = 2$ components	59	87	70	73
$M_i = 3$ components	78	90	<u>88</u>	<u>83</u>
$M_i = 4$ components	<u>90</u>	92	93	87

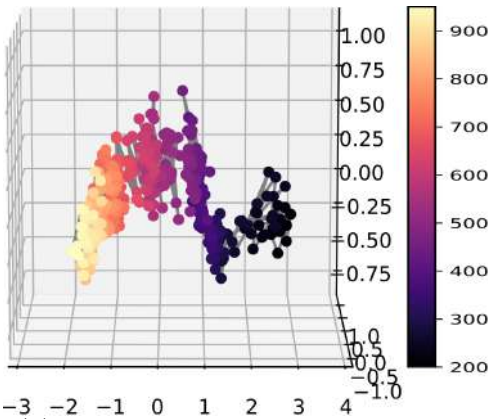
For $i = 0, 1, 2$ and 3, Figures 3.4(a), 3.4(b), 3.4(c) and 3.4(d) are respectively obtained by projecting the latent mean vectors $\boldsymbol{\mu}_\phi(\mathbf{x}) \in \mathbb{R}^L$, for all data vectors $\mathbf{x} \in \mathcal{D}_i$, within the latent subspace characterized by $\mathbf{U}_i \in \mathbb{R}^{L\times M_i}$ (i.e., we perform dimensionality reduction). In the reported experiments, the latent subspace dimension M_i for each factor of variation was chosen such that 80% of the data variance was retained in the latent space. As indicated in Table 3.1, this resulted in $M_0 = 4$, $M_1 = 1$ and $M_2 = M_3 = 3$. In this section, for visualization purposes, we set $M_i = 3$ for all $i \in \{0, 1, 2, 3\}$. However, we can see that the f_1 trajectory (Figure 3.4(b)) is mainly concentrated along a single axis, as indicated by the amount of variance retained by this axis 81% (see Table 3.1). Regarding f_0 (Figure 3.4(a)), setting $M_0 = 3$ retained 78% of the variance of $\mathcal{D}_0$ in the latent space. A recent study explored our method to investigate and give intuitions about the question (Jacquelin et al., 2023): *why the variation of such one-dimensional factor of variation is often explained by multiple latent dimensions?* For example, the authors showed that the first dimension of f_0 correlates with gender, while the second dimension corresponds to variations in f_0 specific to males, and the third dimension corresponds to variations in f_0 specific to females.

From Figure 3.4, we see that two data vectors $\mathbf{x}$ and $\mathbf{x}'$ corresponding to two close values of a given factor have projections of $\boldsymbol{\mu}_\phi(\mathbf{x})$ and $\boldsymbol{\mu}_\phi(\mathbf{x}')$ that are also close in the learned latent subspaces. This can be seen from the color bars which indicate the values of

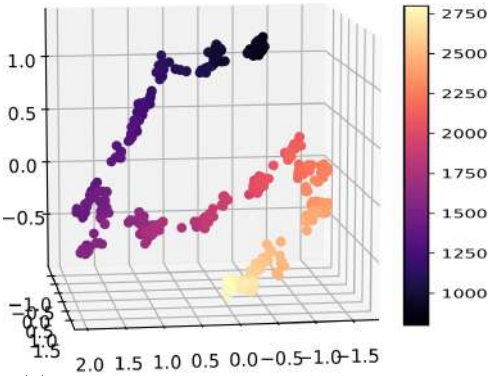
2. https://samsad35.github.io/site-sfvae/



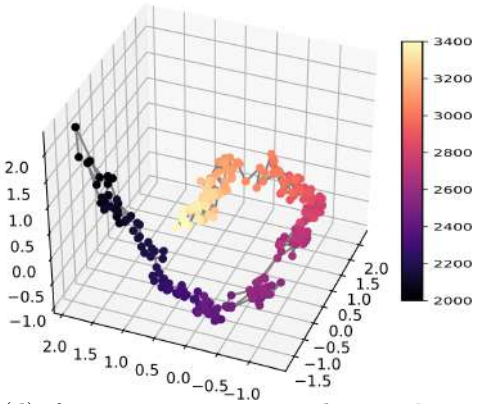
(a) f_0 trajectory corresponding to $\mathcal{D}_0$.



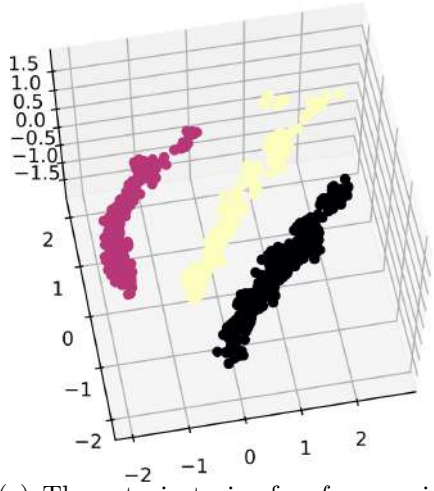
(b) f_1 trajectory corresponding to $\mathcal{D}_1$.



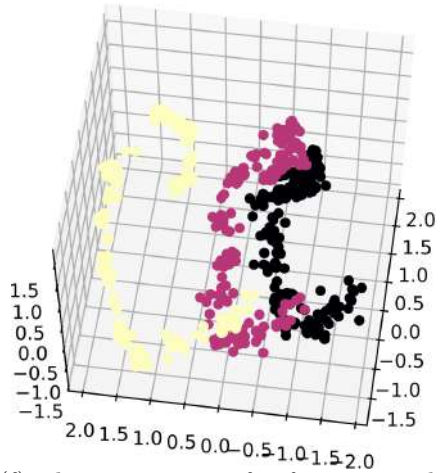
(c) f_2 trajectory corresponding to $\mathcal{D}_2$.



(d) f_3 trajectory corresponding to $\mathcal{D}_3$.



(e) Three trajectories for f_1, associated with different values of f_2.



(f) Three trajectories for f_2, associated with different values of f_1.

Figure 3.4 – Visualization of trajectories in the learned latent subspaces.

the factors of variation. The learned representation thus preserves the notion of proximity in terms of f_0 and formant frequencies.

In Figure 3.4(e), we project three different datasets $\mathcal{D}_1$, defined for three different values of f_2. Similarly, in Figure 3.4(f) we show the trajectories associated with the projection of three datasets $\mathcal{D}_2$, defined for three different values of f_1. We notice that as expected, the trajectories are very similar and mostly differ by a translation.

3.4.3 Quantitative results

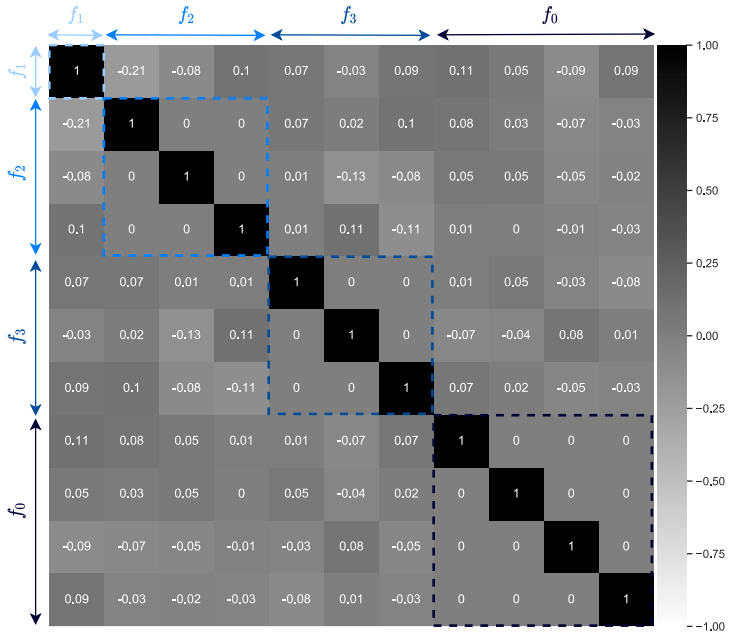


Figure 3.5 – Correlation matrix of the learned latent subspaces basis vectors.

Orthogonality of the latent subspaces

In this experiment, we quantitatively evaluate the proposed method in terms of disentanglement of the learned source-filter latent representation. Following the discussion in Section 3.3.3, we compute the dot product between all pairs of unit vectors in the matrices $\mathbf{U}_i \in \mathbb{R}^{L \times M_i}$, $i \in \{0, 1, 2, 3\}$. Figure 3.5 shows that the resulting correlation matrix is mainly diagonal. Except for a correlation value of -0.21 across f_1 and the first component of f_2, all other values are below 0.13 (in absolute value), confirming the orthogonality of the learned subspaces and thus the disentanglement of the learned

source-filter representation of speech. The methodology presented in Section 3.3.2 was applied to try to identify orthogonal source-filter subspaces directly in the space of speech power spectra and Mel-frequency cepstral coefficients (MFCCs). The results (see A.2) show that an organization into orthogonal source-filter subspaces does not exist when working on such raw signal representations. Indeed, there exist strong correlations between the learned subspaces. This confirms that the disentanglement in terms of source-filter factors is achieved during the unsupervised learning of the VAE model. We remind that in the proposed method, the synthetic labeled data are only used to identify the disentangled subspaces of the learned representation, after the VAE unsupervised training.

Pitch and formant transformations

In this experiment, we quantitatively evaluate the performance of the proposed method regarding the modification of f_0 and the formant frequencies in speech signals (see Section 3.3.4).

Experimental set-up We use a corpus of 12 English vowels uttered by 50 male and 50 female speakers (Hillenbrand et al., 1995), which is labeled with the value of f_0 and the formant frequencies. We also use the TIMIT dataset (Garofolo et al., 1993), a corpus of phonemically and lexically transcribed speech of American English speakers of different sexes and dialects. We used the test corpus containing 1680 utterances. Because we are interested in studying the interaction between modifications of f_0 and the formant frequencies, we only evaluate the method on the voiced phonemes (40 phonemes over a total of 52), which are identified using the annotations. We transform each test signal in the English vowels and TIMIT datasets by varying one single factor f_i at a time, for $i \in \{0, 1, 2, 3\}$, according to the ranges and step sizes given in Table 3.2. For instance, when performing transformations of f_0, for each test signal in the English vowels dataset, we vary the target f_0 value between 100 and 300 Hz linearly, with a step size of 1 Hz, thus resulting in 200 transformations.

Metrics For the modification of each factor f_i, we measure the performance regarding three aspects: First, in terms of *accuracy* by comparing the target value for the factor (see 3.12) and its estimation computed from the modified output speech signal. Second, in terms of *disentanglement*, by comparing the values of f_j for $j \neq i$, before and after modification of the factor f_i. Third, in terms of speech *naturalness* of the transformed

Factor	Min (in Hz)	Max (in Hz)	Step size (in Hz)		Relative variation (%)
			English vowels	TIMIT	
f_0	100	300	1	10	±50
f_1	300	900	10	50	±50
f_2	1100	2700	20	100	±42
f_3	2200	3200	20	50	±18

Table 3.2 – Variation range (min and max values) and step size used for the transformation of each test signal in the English vowels and TIMIT datasets, for each factor of variation f_i, $i \in \{0, 1, 2, 3\}$. The last column indicates by how much a factor varies relative to the center value of its variation range. Its entries are computed as ± (max − min)/(max + min)×100%.

signal.

Accuracy and disentanglement are measured in terms of relative magnitude error (in percent, the lower the better). For a given factor f_i, it is defined by $\delta f_i = 100\% \times |\hat{y} - y|/y$ where y is the target value of f_i and $\hat{y}$ its estimation from the output transformed signal. Let us take the example of a modification of f_0: δf_0 measures the accuracy of the transformation on f_0 while δf_1, δf_2 and δf_3 are used to assess if the other factors of variation f_1, f_2 and f_3 remained unchanged after modifying f_0. We use CREPE (J. W. Kim et al., 2018) to estimate f_0 and Parselmouth (Jadoul et al., 2018), which is based on PRAAT (Boersma & Weenink, 2021a), to estimate the formant frequencies. Regarding speech naturalness, we use the NISQA objective measure (Mittag & Möller, 2020). This metric (the higher the better) was developed in the context of speech transformation algorithms and it was shown to highly correlate with subjective mean opinion scores (MOS) (i.e., human ratings). As a reference, the NISQA score on the original dataset of English vowels (i.e., without any processing) is equal to 2.60 ± 0.53.

TIMIT is phonemically richer than the English vowels dataset, however, it is not labeled with f_0 and formant frequency values. Therefore, we do not have the ground truth values which makes the evaluation in terms of disentanglement more difficult than with the English vowels labeled dataset. Instead of the ground truth, we use the formant frequencies and f_0 values computed on the original speech utterances (i.e., before transformation). This makes the evaluation on TIMIT less reliable than on the English vowels dataset, but it allows us to test the methods on a larger variety of phonemes.

		English vowels dataset					TIMIT dataset				
Factor	Method	NISQA (↑)	δf_0 (%, ↓)	δf_1 (%, ↓)	δf_2 (%, ↓)	δf_3 (%, ↓)	NISQA (↑)	δf_0 (%, ↓)	δf_1 (%, ↓)	δf_2 (%, ↓)	δf_3 (%, ↓)
f_0	TD-PSOLA	2.32 ± 0.55	3.8 ± 2.5	6.3 ± 2.8	3.7 ± 0.9	2.1 ± 0.5	2.36 ± 0.50	2.4 ± 1.9	7.9 ± 0.6	4.5 ± 0.3	3.9 ± 0.2
	WORLD	2.49 ± 0.60	4.5 ± 0.6	3.7 ± 1.8	2.3 ± 0.7	1.2 ± 0.2	2.45 ± 0.47	0.3 ± 0.1	7.1 ± 1.2	6.2 ± 0.4	4.2 ± 0.2
	VAE baseline	1.94 ± 0.43	6.21 ±2.8	10.4 ± 2.4	6.2 ± 0.9	4.5 ± 0.2	1.59 ± 0.43	16.1 ±6.3	17.0 ± 3.0	12.1± 0.2	10.9 ± 1.3
	Proposed	2.08 ± 0.48	0.8 ± 0.2	7.2 ± 1.3	3.6 ± 1.2	3.8 ± 0.3	2.28 ± 0.57	0.8 ± 0.6	9.1 ± 1.1	8.3 ± 0.9	6.0 ± 1.8
f_1	VAE baseline	1.84 ± 0.5	11.3 ± 4.2	15.1 ± 3.5	6.0 ± 1.2	4.2 ± 0.4	1.42 ± 0.34	10.1 ± 2.8	16.4 ± 1.4	12.4 ± 0.9	11.2 ± 2.6
	Proposed	1.85 ± 0.4	6.0 ± 1.6	8.4 ± 3.2	5.7 ± 0.4	4.4 ± 0.3	1.66 ± 0.31	7.1 ± 3.6	9.2 ± 0.8	9.0 ± 1.3	7.8 ± 1.1
f_2	VAE baseline	2.01 ± 0.4	19.5 ± 3.2	10.7 ± 0.5	10.9 ± 1.9	5.8 ± 0.6	1.46 ± 0.30	19.3 ± 5.0	16.4 ± 0.8	20.3 ± 6.3	11.5 ± 0.5
	Proposed	2.03 ± 0.43	8.5 ± 1.1	8.7 ± 1.1	6.2 ± 1.5	5.8 ± 0.2	1.49 ± 0.30	9.1 ± 2.2	8.3 ± 1.3	4.3 ± 1.3	8.1 ± 0.2
f_3	VAE baseline	1.82 ± 0.14	27.0 ± 1.5	13.0 ± 1.3	12.0 ± 1.8	7.3 ± 1.5	1.40 ± 0.48	20.4 ± 1.0	17.4 ± 0.2	14.4 ± 0.2	11.7 ± 2.3
	Proposed	1.94 ± 0.48	8.3 ± 1.0	8.6 ± 0.7	4.9 ± 0.9	2.0 ± 0.4	1.48 ± 0.42	8.5 ± 1.9	8.7 ± 0.9	5.7 ± 2.1	2.5 ± 1.8

Table 3.3 – Performance (mean and standard deviation) for the transformation of f_0 and the formant frequencies (f_1, f_2 and f_3) on the English vowel and TIMIT datasets.

Methods We compare the proposed approach with several methods from the literature: (i) TD-PSOLA (Moulines & Charpentier, 1990) performs f_0 modification through the decomposition of the signal into pitch-synchronized overlapping frames. (ii) WORLD (Morise et al., 2016) is a vocoder also used for f_0 modification. It decomposes the speech signal into three components characterizing f_0, the aperiodicity, and the spectral envelope. (iii) The method proposed by W.-N. Hsu et al., 2017a (here referred to as "VAE baseline") consists in applying translations directly in the latent space of the VAE. Unlike the proposed approach, this method requires predefined latent attribute representations $\boldsymbol{\mu}_{\text{src}}$ and $\boldsymbol{\mu}_{\text{trgt}}$ associated with the source and target values of the factor to be modified, respectively. In particular, computing $\boldsymbol{\mu}_{\text{src}}$ requires analyzing the input speech signal, for instance, to estimate f_0, which is not the case for the proposed method. The source and target latent attribute representations are then used to perform the translation $\tilde{\mathbf{z}} = \mathbf{z} - \boldsymbol{\mu}_{\text{src}} + \boldsymbol{\mu}_{\text{trgt}}$, where $\mathbf{z}$ and $\tilde{\mathbf{z}}$ are respectively the original and modified latent vectors. To ensure a fair comparison, we build dictionaries of predefined latent attribute representations using the same artificially-generated speech data used in the proposed method's training stage. All the methods we compare with require a pre-processing of the input speech signal to compute the input trajectory of the factor to be modified, which is not the case of the proposed method.

Discussion Experimental results (mean and standard deviation) are shown in Table 3.3. Compared to the VAE baseline, the proposed method obtains better performance in terms of accuracy, disentanglement, and naturalness, for both test datasets. These results confirm the effectiveness of performing the transformations in the learned latent subspaces and not

directly in the latent space, as well as the advantage of using regression models instead of predefined latent attribute representations. To analyze the disentanglement results, when performing the transformation for a given factor f_i, one must compare the movement of other factors $\{f_j, j \neq i\}$ relative to their fixed targets (as given by the metrics $\{\delta f_j, j \neq i\}$) to the movement of the factor f_i that is varied relative to the center value of its variation range (as given in the last column of Table 3.2). For instance, when f_0 is varied in the experiments on the English vowels dataset, Table 3.3 shows that the proposed method makes the other factors move from their fixed targets by 7.2%, 3.6%, and 3.8% for f_1, f_2, and f_3 respectively. These values are much smaller than the relative variation of the factor f_0 itself, as indicated in the last column of Table 3.2 is equal to 50%. We can thus conclude that modifying f_0 has little effect on the other factors. Similar conclusions can be drawn by analyzing the disentanglement results for the variation of other factors, confirming the disentanglement of the learned representation. Regarding f_0 transformation, WORLD obtains the best performance in terms of disentanglement, which is because the source and filter contributions are decoupled in the architecture of the vocoder. In terms of naturalness, WORLD and then TD-PSOLA obtain the best performance. This may be explained by the fact that these methods operate directly in the time domain, therefore they do not suffer from phase reconstruction artifacts, unlike the proposed and VAE baseline methods. Naturalness is indeed greatly affected by phase reconstruction artifacts, even from an unaltered speech spectrogram (i.e., without transformation). Phase reconstruction in a multi-speaker setting is still an open problem in speech processing. We want to emphasize that the objective of this study is not to compete with traditional signal processing methods such as TD-PSOLA and WORLD. It is rather to advance on the understanding of deep generative modeling of speech signals and to compare honestly with highly-specialized traditional systems. TD-PSOLA and WORLD exploit signal models specifically designed for the task at hand, which for instance prevents them to be used for modifying formant frequencies. In contrast, the proposed method is fully based on learning and the same methodology applies to modifying f_0 or the formant frequencies.

Robustness with respect to the VAE training dataset

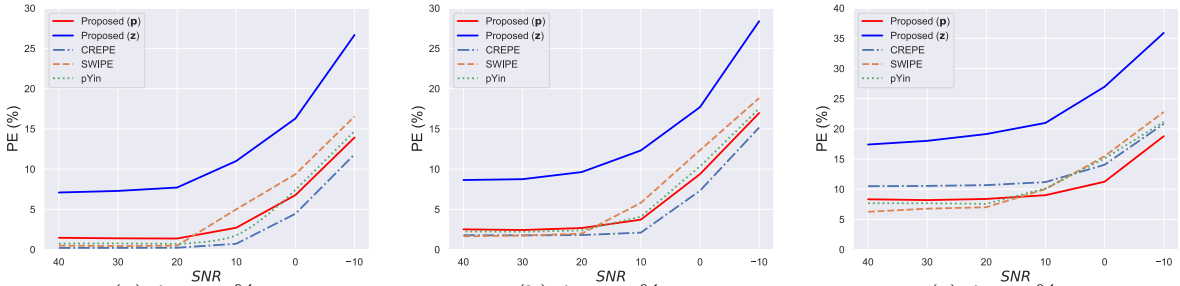
In this Section, we investigate the robustness of the proposed method with respect to different datasets used to train the VAE model. We considered three training datasets in addition to the WSJ0 dataset used in the previous experiments: (i) the SIWIS French speech synthesis dataset (Honnet et al., 2017), which contains more than ten hours of

Dataset	NISQA (↑)	δf_0 (%, ↓)	δf_1 (%, ↓)	δf_2 (%, ↓)	δf_3 (%, ↓)
WSJ	2.08 ± 0.48	0.8 ± 0.2	7.2 ± 1.3	3.6 ± 1.2	3.8 ± 0.3
SIWIS	1.93 ± 0.43	1.2 ± 0.5	10.0 ± 4.2	8.3 ± 1.1	14.0 ± 0.2
TESS	1.98 ± 0.50	2.7 ± 2.3	9.3 ± 3.5	9.0 ± 0.8	7.0 ± 0.2
LJspeech	1.96 ± 0.40	1.2 ± 0.6	9.3 ± 1.2	5.6 ± 0.6	4.6 ± 0.1

Table 3.4 – Performance (mean and standard deviation) of f_0 transformation with the proposed method, on the English vowels test dataset, using different training datasets for the unsupervised VAE model.

French speech recordings; (ii) the Toronto emotional speech (TESS) dataset (Dupuis & Pichora-Fuller, 2010), which contains 2,800 utterances spoken by two actresses using different emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral); and (iii) the LJspeech dataset (Ito & Johnson, 2017), which contains 13,100 short audio clips of a single speaker reading passages from 7 non-fiction books. The artificially-generated speech dataset used for learning the latent subspaces and the regression models remain the same.

Table 3.4 presents the results for the modification of f_0 only, applied to the English vowels dataset. It can be seen in Table 3.4 that the performance remains quite stable with different VAE training datasets. WSJ0 is the largest dataset and therefore leads to the best performance. Interestingly, the results obtained with the SIWIS dataset of French speech signals remain satisfactory, even if there is a mismatch between the training (French) and testing (English) datasets.



(a) $\lambda = 20\%$ (b) $\lambda = 10\%$ (c) $\lambda = 1\%$

Figure 3.6 – Results of the f_0 tracking experiment: Pitch error (PE, in %) as a function of the SNR (in dB), for different values of the threshold λ (in %). "Proposed($\mathbf{p}$)" and "Proposed($\mathbf{z}$)" denote the proposed approach for f_0 estimation using the projection of $\mathbf{z}$ into the learned subspace of the pitch and using $\mathbf{z}$ directly without the projection, respectively.

Robust f_0 estimation

Experimental set-up For the experiments on fundamental frequency estimation, we used the Pitch Tracking Database from Graz University of Technology (PTDB-TUG) (Pirker et al., 2011). This dataset provides microphone and laryngograph signals from 20 native English speakers. The ground truth f_0 is extracted from the laryngograph signals. To evaluate the robustness of the f_0 estimation methods, we corrupted each clean speech test signal by adding real-world cafeteria noise extracted from the DEMAND dataset (Thiemann et al., 2013), where the signal-to-noise ratio (SNR) is varied from -10 to 40 dB. In this experiment, for both the proposed and reference methods, we only considered the problem of estimating f_0 given the ground-truth voicing labels, i.e., we did not consider the problem of detecting which frames are voiced.

Metric Following Rabiner et al., 1976, the performance is measured in terms of pitch error (PE) defined as the proportion of frames considered as voiced by both the estimation algorithm and the ground truth for which the relative f_0 error is higher than a certain threshold λ in %:

$$PE = \frac{N_{est}}{N_v} \times 100\%, \tag{3.17}$$

where N_v is the number of voiced frames and N_{est} is the number of the frames for which $|\hat{y} - y|/y \geq \lambda$ with $\hat{y}$ and y the estimated and ground-truth f_0 values, respectively. The parameter λ can be interpreted as the tolerance for which the f_0 prediction is considered as correct.

Methods We compared the proposed approach described in Section 3.3.5 with several state-of-the-art methods from the literature: (i) pYin (Mauch & Dixon, 2014) improvement of the Yin-based autocorrelation algorithm in the time domain using probabilistic threshold distribution. We used the one implemented in the Librosa package (McFee et al., 2015); (ii) SWIPE (Camacho & Harris, 2008) is a method for pitch detection based on the autocorrelation of the speech signal in the frequency domain. We used the one implemented in the pysptk toolkit (Yamamoto et al., 2019); (iii) CREPE (J. W. Kim et al., 2018) is a supervised method based on CNN already used in Section 3.4.3. We used the one implemented in the torchcrepe github of Morrison, 2020. To confirm the interest of estimating the f_0 in the corresponding latent subspace and not directly in the VAE latent space, we also apply the proposed method described in 3.3.5 by directly working with

$q_\phi(\mathbf{z}|\mathbf{x})$ instead of $q_\phi(\mathbf{p}|\mathbf{x})$ to compute the KL divergence in (3.16). This aims to show that projecting the speech signal in the latent subspace encoding f_0 indeed helps discarding information that is not related to this factor and thus improves the estimation accuracy.

Discussion The results are shown in Figure 3.6. Figures 3.6(a), 3.6(b), and 3.6(c) display the PE metric (%) as a function of the SNR for λ equal to 20%, 10%, and 1%, respectively. The proposed pitch estimation method based on the projection of the latent variable $\mathbf{z}$ on the f_0 subspace outperforms the same approach using the latent variable $\mathbf{z}$ directly. This result confirms the fact that the f_0 subspace is less sensitive to formant variations (invariance property), and projecting $\mathbf{z}$ into this subspace globally preserves well the pitch information, which makes the detection more robust to variations that are independent of f_0. For large λ values (i.e., high precision tolerance), the proposed method gives results that are competitive with the state-of-the-art methods for moderate noise levels (high SNRs) and that outperform SWIPE for a high level of noise (low SNRs). For small λ values (i.e., low precision tolerance), the proposed approach shows the best trade-off between robustness to noise and precision. Indeed, the proposed method is about 2% PE below CREPE, consistently over the whole SNR range. It also outperforms SWIPE and pYin below 14dB SNR.

3.5 Conclusion of the chapter

The source-filter model of speech production is a fundamental concept in speech processing. In this work, using only a few seconds of artificially generated labeled speech data, we showed that the fundamental frequency and formant frequencies are encoded in orthogonal latent subspaces of an unsupervised VAE. We proposed to exploit this disentangled source-filter latent representation for the transformation and analysis of speech spectrograms. Using a regression model trained on the artificially-generated labeled speech data, we proposed a method to control f_0 and the formant frequencies in speech spectrograms by applying affine transformations in the learned latent subspaces. We also proposed to exploit the projection of the speech signal in the latent subspace associated with f_0 to robustly estimate the latter from speech signals corrupted by noise. Even if the identification of the latent source-filter subspaces, the learning of the regression models, and the f_0 estimation method were designed using a

very limited set of artificially-generated signals, we showed experimentally that the speech transformations and analysis are effective on natural signals. To the best of our knowledge, this is the first approach that, with a single methodology, is able to extract, identify and control the source and filter low-level speech attributes within a VAE latent space. This is an important step towards a better understanding of deep latent-variable generative modeling of speech signals.

Chapter 4

A MULTIMODAL DYNAMICAL VARIATIONAL AUTOENCODER FOR AUDIOVISUAL SPEECH REPRESENTATION LEARNING

Contents

Summary

In the previous chapter, we explored the learning and control of factors of variation in audio speech, focusing on aspects such as pitch and formants. While these factors are essential, they primarily address low-level characteristics and do not fully capture the sequential nature of the data. To address this limitation, we introduce a novel approach that considers both the data's multimodality and dynamical aspects. We seek to learn a multimodal dynamical VAE (MDVAE) that disentangles the audiovisual speech latent factors. We aim to learn a hierarchical latent space that separates static from dynamical information and modality-common from modality-specific information in *unsupervised learning*. In addition, we propose a two-stage training method that enhances reconstruction and generation quality, as well as improves training efficiency. The proposed approach is validated through extensive experiments to demonstrate the effectiveness of the proposed model: involving resynthesis, transformation-synthesis, image denoising, and emotion recognition.

4.1 Introduction

The world around us is represented by a multitude of different modalities (Lazarus, 1976). A single event can be observed from different perspectives, and combining these different views can provide a complete understanding of what is happening. For instance, speech in human interactions is a multimodal process where the audio and visual modalities carry complementary verbal, as well as non-verbal, information. By capturing the correlations between different modalities, we can reduce uncertainty and better understand a phenomenon (Bengio, Courville, & Vincent, 2013). Combining complementary sources of information from heterogeneous modalities is a challenging task, for which machine and deep learning techniques have shown their efficiency. In particular, the flexibility and versatility of deep neural networks allow them to efficiently learn from heterogeneous data to solve a given task (Baltrušaitis et al., 2018; Ramachandram & Taylor, 2017).

The rapid development of artificial intelligence technology and hardware acceleration has led to a shift towards multimodal processing (Ramachandram & Taylor, 2017), which aims to enhance machine perception by integrating various data types. With the explosion of

digital content and communication, audiovisual speech processing has become increasingly important for a range of applications, such as speech recognition (Afouras et al., 2018; Hori et al., 2019; Petridis et al., 2018), speaker identification (Roth et al., 2020; Vallet et al., 2012), speech enhancement in noise (Sadeghi et al., 2020), and emotion recognition (Augusma et al., 2023; F. Noroozi et al., 2017; Schoneveld et al., 2021; C.-H. Wu et al., 2014). However, in tasks such as emotion recognition, the limited availability of labeled data remains a significant challenge. As a result, researchers are investigating unsupervised or weakly supervised methods to learn effective audiovisual speech representations. This is extremely promising in problem settings involving a large amount of unlabeled data but limited labeled data.

Deep generative models (Goodfellow et al., 2014; Kingma & Welling, 2014; Rezende et al., 2014) have recently become very successful for unsupervised learning of latent representations from high-dimensional and structured data such as images, audio, and text. Learning meaningful representations is essential not only for synthesizing data but also for data analysis and transformation. For a learned representation to be effective, as discussed in Section 2.3, it should capture high-level characteristics invariant to small and local changes in the input data and be as disentangled as possible for explainability. Furthermore, hierarchical and disentangled generative models have demonstrated their efficacy in solving downstream learning tasks (Bengio, Courville, & Vincent, 2013; Van Steenkiste et al., 2019). Variants of generative models have recently led to considerable progress in disentangled representation learning, particularly with the VAE (Kingma & Welling, 2014; Rezende et al., 2014), for a comprehensive overview of the VAEs, please consult Section 2.4. Early methods for disentanglement using VAEs focused on modifying the evidence lower bound objective function (R. T. Chen et al., 2018; Higgins et al., 2017a; H. Kim & Mnih, 2018). Since unsupervised disentanglement in a generative model is impossible without incorporating inductive biases on both models and data (Locatello et al., 2019), new approaches are oriented towards weakly-supervised (Locatello, Poole, et al., 2020; Sadok, Leglaive, Girin, et al., 2023a) or semi-supervised learning (Klys et al., 2018). Because of their flexibility in modeling complex data, VAEs have been extended to various data types, including multimodal or sequential data.

While many extensions of the VAE have been proposed to handle either sequential (refer to Section 2.4.6, Page 55) or multimodal data (refer to Section 2.4.7, Page 60), none have been able to process both types of data simultaneously. This chapter presents a novel approach for modeling multimodal and sequential data in a single framework, specifically

applied to audiovisual speech data. We propose the first unsupervised generative model of multimodal and sequential data, to learn a hierarchical latent space that separates static from dynamical information and modality-common from modality-specific information. The proposed model, called Multimodal Dynamical VAE (MDVAE), is trained on an expressive audiovisual speech database and evaluated on three tasks: the transformation of audiovisual speech data, audiovisual facial image denoising, and audiovisual speech emotion recognition.

4.2 Multimodal dynamical VAE

This section presents the design and architecture of MDVAE. Initially, we motivate the structure of the MDVAE latent space from the audiovisual speech generative modeling perspective. Subsequently, we formalize the MDVAE generative and inference models. Finally, we introduce a two-stage training approach for unsupervised learning of the MDVAE model.

Table 4.1 – Summary of the notations.

Variable notation	Definition
T, t	Sequence length and time/frame index
$\mathbf{x}^{(a)} \in \mathbb{R}^{d_a \times T}$	Observed audio data sequence
$\mathbf{x}^{(v)} \in \mathbb{R}^{d_v \times T}$	Observed visual data sequence
$\mathbf{w} \in \mathbb{R}^{w}$	Latent static audiovisual vector
$\mathbf{z}^{(av)} \in \mathbb{R}^{l_{av} \times T}$	Latent dynamical audiovisual vectors
$\mathbf{z}^{(a)} \in \mathbb{R}^{l_a \times T}$	Latent dynamical audio vectors
$\mathbf{z}^{(v)} \in \mathbb{R}^{l_v \times T}$	Latent dynamical visual vectors
$\mathbf{z} = \{\mathbf{z}^{(a)}, \mathbf{z}^{(v)}, \mathbf{z}^{(av)}, \mathbf{w}\}$	Set of all latent variables
$\mathbf{x} = \{\mathbf{x}^{(a)}, \mathbf{x}^{(v)}\}$	Set of all observations

4.2.1 Motivation and notations

We aim to model emotional audiovisual speech at the utterance level, where a single speaker speaks and expresses a single emotion. Let $\{\mathbf{x}^{(a)}, \mathbf{x}^{(v)}\}$ denote the observed audiovisual speech data, where $\mathbf{x}^{(a)} \in \mathbb{R}^{d_a \times T}$ is a sequence of audio features

of dimension d_a, $\mathbf{x}^{(v)} \in \mathbb{R}^{d_v \times T}$ is a sequence of observed visual features of dimension d_v, and T is the sequence length. For the audio speech, features are extracted from the power spectrogram of the signal, and for the visual speech, features are extracted from the pre-processed face images. The feature extraction process will be further discussed below.

To motivate the structure of the generative model in MDVAE, let us reason about the latent factors involved in generating an emotional audiovisual speech sequence. First, the speaker's identity and global emotional state correspond to static and audiovisual latent factors. Indeed, these do not evolve with time at the utterance level, and they are shared between the two modalities as defined from both vocal and visual attributes (e.g., the average pitch and timbre of the voice and the visual appearance). Second, we have dynamical latent factors that are shared between the two modalities, so audiovisual factors that vary with time. This typically corresponds to the phonemic information carried by the movements of the speech articulators that are visible in the visual modality, namely the jaw and lips. Finally, we have dynamical latent factors that are specific to each modality. Visual-only dynamical factors include, for instance, facial movements that are not related to the mouth region and the head pose. Audio-only dynamical factors include the pitch variations, induced by the vibration of the vocal folds, and the phonemic information carried by the tongue movements, which is another important speech articulator not visible in the visual modality.

This analysis of the latent factors involved in the generative process of emotional audiovisual speech suggests structuring the latent space of the MDVAE model by introducing the following latent variables: $\mathbf{w} \in \mathbb{R}^w$ is a static latent variable assumed to encode audiovisual information that does not evolve with time; $\mathbf{z}^{(av)} \in \mathbb{R}^{l_{av} \times T}$ is a dynamical (i.e., sequential) latent variable assumed to encode audiovisual information that evolves with time; $\mathbf{z}^{(a)} \in \mathbb{R}^{l_a \times T}$ is a dynamical latent variable assumed to encode audio-only information; $\mathbf{z}^{(v)} \in \mathbb{R}^{l_v \times T}$ is a dynamical latent variable assumed to encode visual-only information. A time/frame index $t \in \{1, 2, ..., T\}$ is added in the subscript of dynamical variables to denote one particular frame within a sequence (i.e., $\mathbf{x}_t^{(a)}$, $\mathbf{x}_t^{(v)}$, $\mathbf{z}_t^{(a)}$, $\mathbf{z}_t^{(v)}$, $\mathbf{z}_t^{(av)}$). The above notations are summarized in Table 4.1.

In summary, the MDVAE model is a generative model of audiovisual speech data $\{\mathbf{x}^{(a)}, \mathbf{x}^{(v)}\}$ that involves four different latent variables $\{\mathbf{w}, \mathbf{z}^{(av)}, \mathbf{z}^{(a)}, \mathbf{z}^{(v)}\}$. In the latent space of MDVAE, we can dissociate the latent factors that are static ($\mathbf{w}$) from those that are dynamic ($\mathbf{z}^{(av)}, \mathbf{z}^{(v)}, \mathbf{z}^{(a)}$), and we can dissociate the latent factors that are

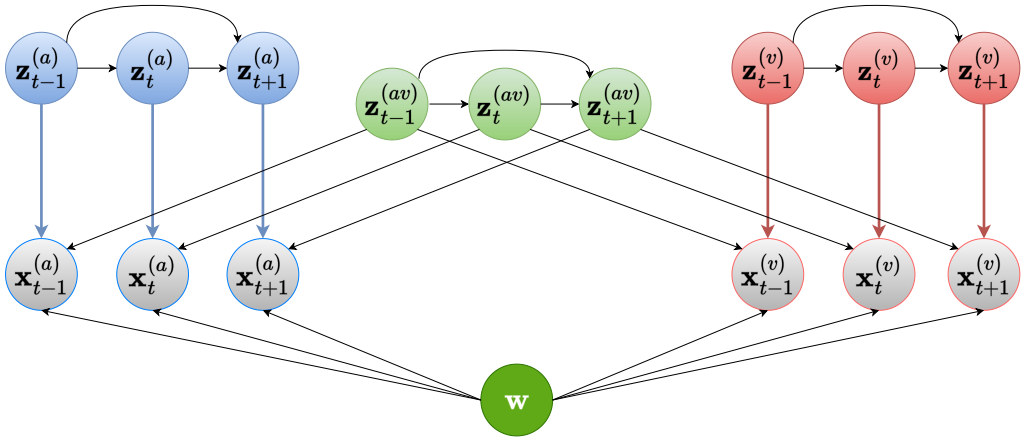


Figure 4.1 – MDVAE generative probabilistic graphical model.

shared between the modalities ($\mathbf{w}, \mathbf{z}^{(av)}$) from those that are specific to each modality ($\mathbf{z}^{(a)}, \mathbf{z}^{(v)}$). A study by Gao and Shinkareva, 2021 recently showed that the human brain distinguishes between modality-common and modality-specific information for affective processing in a multimodal context. In the MDVAE model, we also introduce temporal modeling on top of this dichotomy regarding modality-common vs modality-specific information. Our objective is to learn a multimodal and dynamical VAE that can disentangle the above-mentioned latent factors in an unsupervised manner to analyze and transform emotional audiovisual speech data. The next subsections detail the generative and inference models of MDVAE and its two-stage training.

4.2.2 Generative model

The generative model of MDVAE is represented as a Bayesian network in Figure 4.1, which also corresponds to the following factorization of the joint distribution of the observed and latent variables:

$$p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta\left(\mathbf{x}^{(a)}|\mathbf{w}, \mathbf{z}^{(av)}, \mathbf{z}^{(a)}\right) p_\theta\left(\mathbf{x}^{(v)}|\mathbf{w}, \mathbf{z}^{(av)}, \mathbf{z}^{(v)}\right) \times p\left(\mathbf{w}\right) p_\theta\left(\mathbf{z}^{(av)}\right) p_\theta\left(\mathbf{z}^{(a)}\right) p_\theta\left(\mathbf{z}^{(v)}\right), \tag{4.1}$$

where $\mathbf{x} = \{\mathbf{x}^{(a)}, \mathbf{x}^{(v)}\}$, $\mathbf{z} = \{\mathbf{z}^{(a)}, \mathbf{z}^{(v)}, \mathbf{z}^{(av)}, \mathbf{w}\}$, and

$$p_\theta\left(\mathbf{x}^{(a)}|\mathbf{w}, \mathbf{z}^{(av)}, \mathbf{z}^{(a)}\right) = \prod_{t=1}^{T} p_\theta\left(\mathbf{x}_t^{(a)}|\mathbf{w}, \mathbf{z}_t^{(av)}, \mathbf{z}_t^{(a)}\right); \tag{4.2}$$

$$p_\theta\left(\mathbf{x}^{(v)}|\mathbf{w},\mathbf{z}^{(av)},\mathbf{z}^{(v)}\right)=\prod_{t=1}^{T}p_\theta\left(\mathbf{x}_t^{(v)}|\mathbf{w},\mathbf{z}_t^{(av)},\mathbf{z}_t^{(v)}\right);\tag{4.3}$$

$$p(\mathbf{z}^{(av)})=\prod_{t=1}^{T}p_\theta\left(\mathbf{z}_t^{(av)}|\mathbf{z}_{1:t-1}^{(av)}\right);\tag{4.4}$$

$$p(\mathbf{z}^{(a)})=\prod_{t=1}^{T}p_\theta\left(\mathbf{z}_t^{(a)}|\mathbf{z}_{1:t-1}^{(a)}\right);\tag{4.5}$$

$$p(\mathbf{z}^{(v)})=\prod_{t=1}^{T}p_\theta\left(\mathbf{z}_t^{(v)}|\mathbf{z}_{1:t-1}^{(v)}\right).\tag{4.6}$$

Equation 4.2 (resp. 4.3) indicates that, at time index t, the observed audio (resp. visual) speech vector $\mathbf{x}_t^{(a)}$ (resp. $\mathbf{x}_t^{(v)}$) is generated from the audiovisual static latent variable ($\mathbf{w}$), the audiovisual dynamical latent variable at time index t ($\mathbf{z}_t^{(av)}$), and the audio-only (resp. visual-only) dynamical latent variable at time index t ($\mathbf{z}_t^{(a)}$, resp. $\mathbf{z}_t^{(v)}$). In particular, we see that $\mathbf{w}$ is involved in generating the complete audiovisual speech sequence ($\mathbf{x}^{(a)},\mathbf{x}^{(v)}$). All latent variables are assumed independent, and the autoregressive structure of the priors for the dynamical variables in equations 4.4-4.6 is inspired by DSAE (Y. Li & Mandt, 2018), which is discussed in Subsection 2.4.6 in page 55. Following standard DVAEs (Girin et al., 2021b), each conditional distribution that appears in a product over the time indices in equations 4.2-4.6 is modeled as a Gaussian with a diagonal covariance, and its parameters are provided by deep neural networks (decoders) that take as input the variables after the conditioning bars. For the distributions in equations 4.2-4.3, the variance coefficients are fixed to one, while for the distributions in equations 4.4-4.6, the variance coefficients are learned. Standard feed-forward fully-connected neural networks can be used to parametrize conditional distributions over the observed audiovisual speech variables. The autoregressive structure of the priors over the latent dynamical variables requires the use of RNNs. Finally, the prior over the static latent variable $\mathbf{w}$ is a Gaussian with zero mean and identity covariance matrix. More details about the decoder network architectures can be found in B.1.

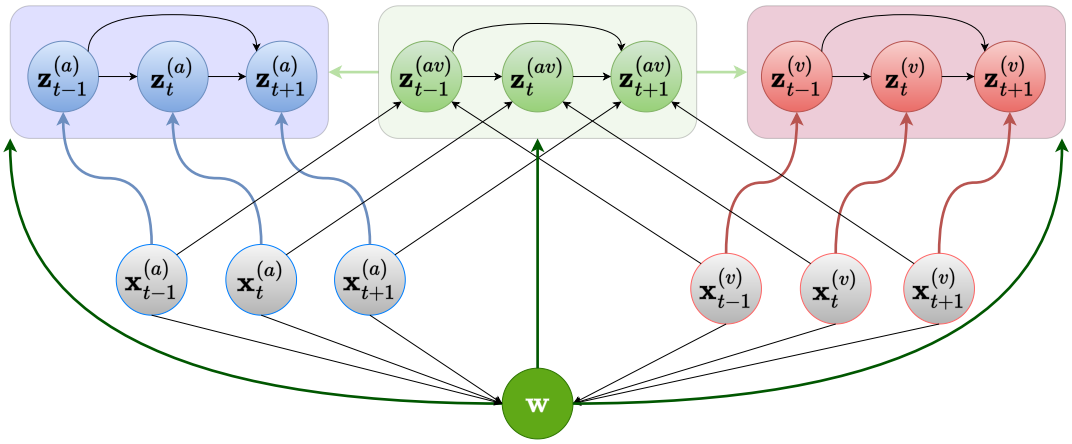


Figure 4.2 – MDVAE inference probabilistic graphical model.

4.2.3 Inference model

As in the standard VAE, the exact posterior distribution of the latent variables in the MDVAE model is intractable, we thus need to define an inference model $q_\phi(\mathbf{z}|\mathbf{x}) \approx p_\theta(\mathbf{z}|\mathbf{x})$. However, it is not because the exact posterior distribution is intractable that we cannot look at the structure of the exact posterior dependencies. Actually, using the Bayesian network of the model, the chain rule of probabilities, and D-separation (Bishop & Nasrabadi, 2006; Geiger et al., 1990), it is possible to analyze how the observed and latent variables depend on each other in the exact posterior, and define an inference model with the same dependencies. An extensive discussion of D-separation in the context of DVAEs can be found in Girin et al., 2021b. The Bayesian network corresponding to our MDVAE model is represented in Figure 4.1. For this model, it is relevant to factorize the inference model as follows:

$$\begin{aligned} q_\phi(\mathbf{z}|\mathbf{x}) = q_\phi\left(\mathbf{w}|\mathbf{x}^{(a)}, \mathbf{x}^{(v)}\right) q_\phi\left(\mathbf{z}^{(av)}|\mathbf{x}^{(a)}, \mathbf{x}^{(v)}, \mathbf{w}\right) \\ \times q_\phi\left(\mathbf{z}^{(a)}|\mathbf{x}^{(a)}, \mathbf{z}^{(av)}, \mathbf{w}\right) q_\phi\left(\mathbf{z}^{(v)}|\mathbf{x}^{(v)}, \mathbf{z}^{(av)}, \mathbf{w}\right), \end{aligned} \tag{4.7}$$

where

$$q_\phi\left(\mathbf{z}^{(av)}|\mathbf{x}^{(a)}, \mathbf{x}^{(v)}, \mathbf{w}\right) = \prod_{t=1}^{T} q_\phi\left(\mathbf{z}_t^{(av)} \mid \mathbf{z}_{1:t-1}^{(av)}, \mathbf{x}_{t:T}^{(a)}, \mathbf{x}_{t:T}^{(v)}, \mathbf{w}\right); \tag{4.8}$$

$$q_\phi\left(\mathbf{z}^{(a)}|\mathbf{x}^{(a)}, \mathbf{z}^{(av)}, \mathbf{w}\right) = \prod_{t=1}^{T} q_\phi\left(\mathbf{z}_t^{(a)} \mid \mathbf{z}_{1:t-1}^{(a)}, \mathbf{x}_{t:T}^{(a)}, \mathbf{z}_t^{(av)}, \mathbf{w}\right); \tag{4.9}$$

$$q_\phi\left(\mathbf{z}^{(v)}|\mathbf{x}^{(v)}, \mathbf{z}^{(av)}, \mathbf{w}\right) = \prod_{t=1}^{T} q_\phi\left(\mathbf{z}_t^{(v)} \mid \mathbf{z}_{1:t-1}^{(v)}, \mathbf{x}_{t:T}^{(v)}, \mathbf{z}_t^{(av)}, \mathbf{w}\right). \tag{4.10}$$

This factorization is consistent with the exact posterior dependencies between the latent and observed variables, i.e., no approximation was made as we followed the principle of D-separation. However, to lighten the inference model architecture, we choose to omit the non-causal dependencies on the observations in 4.8, 4.9 and 4.10. In these equations, we thus replace $\mathbf{x}_{t:T}^{(a)}$ with $\mathbf{x}_t^{(a)}$ and $\mathbf{x}_{t:T}^{(v)}$ with $\mathbf{x}_t^{(v)}$, and the equalities become approximations. In this inference model, $q_\phi\left(\mathbf{w}|\mathbf{x}^{(a)}, \mathbf{x}^{(v)}\right)$ and each conditional distribution that appears in a product over the time indices in equations 4.8-4.10 is modeled as a Gaussian with a diagonal covariance, and its parameters (mean vector and variance coefficients) are provided by deep neural networks (encoders) that take as input the variables after the conditioning bars. In practice, the MDVAE encoder can be decomposed into four sub-encoders, each dedicated to the inference of a specific latent variable. Distinct conditioning variables are concatenated at the input of these sub-encoders depending on the structure of the corresponding inference model. For instance, when inferring $\mathbf{w}$, we concatenate $\mathbf{x}^{(a)}$ and $\mathbf{x}^{(v)}$ along the feature dimension. More details about the encoder network architectures can be found in B.1.

The probabilistic graphical model of MDVAE during inference is represented in Figure 4.2, and it corresponds to the factorization in 4.7. It can be interpreted as follows: First, we infer the static audiovisual latent variable $\mathbf{w}$ from the observed audiovisual speech sequence, which corresponds to the computation of $q_\phi(\mathbf{w}|\mathbf{x}^{(a)}, \mathbf{x}^{(v)})$. Next, we infer the audiovisual dynamical latent variable $\mathbf{z}^{(av)}$ from the previously inferred variable $\mathbf{w}$ and the observed audiovisual speech, which corresponds to the computation of $q_\phi(\mathbf{z}^{(av)}|\mathbf{x}^{(a)}, \mathbf{x}^{(v)}, \mathbf{w})$. Indeed, we need the static audiovisual information to infer the dynamical audiovisual information from the audiovisual speech observations. Finally, we infer the audio-only (resp. visual-only) dynamical latent variables $\mathbf{z}^{(a)}$ (resp. $\mathbf{z}^{(v)}$) from the audio (resp. visual) speech observations $\mathbf{x}^{(a)}$ (resp. $\mathbf{x}^{(v)}$) and the previously-inferred audiovisual latent variables $\mathbf{w}$ and $\mathbf{z}^{(av)}$, which corresponds to the computation of $q_\phi(\mathbf{z}^{(a)}|\mathbf{x}^{(a)}, \mathbf{z}^{(av)}, \mathbf{w})$ (resp. $q_\phi(\mathbf{z}^{(v)}|\mathbf{x}^{(v)}, \mathbf{z}^{(av)}, \mathbf{w})$). This is logical, as to infer the latent information that is specific to one modality, we require the observations of that modality and also the latent information

that is shared with the other modality, which is captured by $\mathbf{w}$ and $\mathbf{z}^{(av)}$.

4.2.4 Training of MDVAE

The evidence lower bound of MDVAE

As in standard (D)VAEs (Girin et al., 2021a; Kingma & Welling, 2014; Rezende et al., 2014), learning the MDVAE generative and inference model parameters consists in maximizing the evidence lower-bound (ELBO):

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right] - D_{\mathrm{KL}} \left(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z})\right). \tag{4.11}$$

The first term in 4.11 is the reconstruction accuracy term, which aims to maximize the conditional log-likelihood over a training dataset. The input and output data can take any form, including raw images for the visual modality and speech power spectra for the audio modality, or can be replaced by *any representation from another pre-trained model*. The second term is the latent space regularization term, which encourages the latent variables to conform to the prior distribution.

We present the derivation to obtain the ELBO for the MDVAE:

$$\begin{aligned}
\log p_\theta(\mathbf{x}) &= \log \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] \\
&\geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] \\
\mathcal{L}(\theta, \phi) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \underbrace{\frac{p_\theta(\mathbf{x}^{(a)}, \mathbf{x}^{(v)}, \mathbf{z}^{(av)}, \mathbf{z}^{(a)}, \mathbf{z}^{(v)}, \mathbf{w})}{q_\phi(\mathbf{z}^{(av)}, \mathbf{z}^{(a)}, \mathbf{z}^{(v)}, \mathbf{w}|\mathbf{x}^{(a)}, \mathbf{x}^{(v)})}}_{\text{(F)}}\right]
\end{aligned} \tag{4.12}$$

The fraction above (F) can be simplified using the equations 4.1 and 4.7 as follows:

$$\begin{aligned}
(F) = p_\theta(\mathbf{x}^{(a)}, \mathbf{x}^{(v)}|\mathbf{w}, \mathbf{z}^{(av)}, \mathbf{z}^{(a)}, \mathbf{z}^{(v)}) \cdot \frac{p(\mathbf{w})}{q_\phi\left(\mathbf{w}|\mathbf{x}\right)} \cdot \frac{p_\theta(\mathbf{z}^{(av)})}{q_\phi\left(\mathbf{z}^{(av)}|\mathbf{x}, \mathbf{w}\right)} \\
\times \frac{p_\theta(\mathbf{z}^{(a)})}{q_\phi\left(\mathbf{z}^{(a)}|\mathbf{x}^{(a)}, \mathbf{z}^{(av)}, \mathbf{w}\right)} \cdot \frac{p_\theta(\mathbf{z}^{(v)})}{q_\phi\left(\mathbf{z}^{(v)}|\mathbf{x}^{(v)}, \mathbf{z}^{(av)}, \mathbf{w}\right)}
\end{aligned} \tag{4.13}$$

We incorporate equation 4.13 into equation 4.12. The writing of the ELBO can be

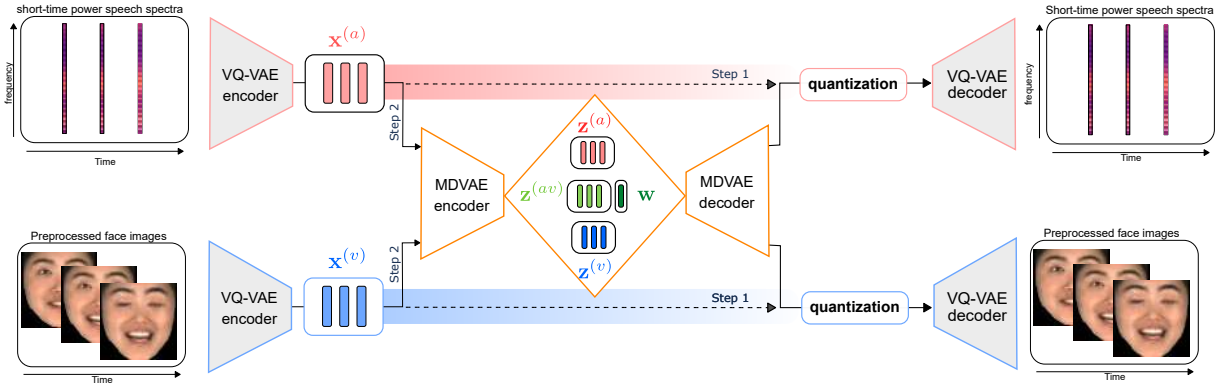


Figure 4.3 – The overall architecture of VQ-MDVAE. During the first step of the training process, we learn a VQ-VAE independently on each modality, without any temporal modeling. During the second step of the training process, we learn the MDVAE model on the latent representation provided by the frozen VQ-VAE encoders, before quantization.

simplified as below:

$$
\begin{aligned}
\mathcal{L}(\theta,\phi) = {} & \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta\left(\mathbf{x}^{(a)} \mid \mathbf{w}, \mathbf{z}^{(av)}, \mathbf{z}^{(a)}\right)\right] + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta\left(\mathbf{x}^{(v)} \mid \mathbf{w}, \mathbf{z}^{(av)}, \mathbf{z}^{(v)}\right)\right] \\
& - D_{KL}(q_\phi(\mathbf{w}|\mathbf{x}^{(a)}, \mathbf{x}^{(v)}) \parallel p(\mathbf{w})) \\
& - \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[D_{KL}\left(q_\phi\left(\mathbf{z}^{(av)}|\mathbf{x}^{(a)}, \mathbf{x}^{(v)}, \mathbf{w}\right) \parallel p_\theta\left(\mathbf{z}^{(av)}\right)\right)\right] \\
& - \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[D_{KL}\left(q_\phi\left(\mathbf{z}^{(a)}|\mathbf{x}^{(a)}, \mathbf{w}, \mathbf{z}^{(av)}\right) \parallel p_\theta\left(\mathbf{z}^{(a)}\right)\right)\right] \\
& - \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[D_{KL}\left(q_\phi\left(\mathbf{z}^{(v)}|\mathbf{x}^{(v)}, \mathbf{w}, \mathbf{z}^{(av)}\right) \parallel p_\theta\left(\mathbf{z}^{(v)}\right)\right)\right].
\end{aligned} \tag{4.14}
$$

This simplified ELBO 4.14 is excluded from any time dependency. In a direct way, we could expand it further by introducing the temporal dependencies exposed in the subsections 4.2.2 and 4.2.3.

Two stage training

Unlike GANs (Goodfellow et al., 2014), VAEs often produce poor reconstructions that lack realism, which also affects the generation of new data. Improving the quality of VAE reconstruction or generation is an active area of research. One issue with VAE is that using an information bottleneck in combination with a pixel-wise reconstruction error can result in blurry, unrealistic images. This problem also exists with the audio modality, where VAE-generated sound is often unnatural, mainly when using a time-frequency representation. To address this problem, several solutions have been proposed. One approach is to combine

VAEs and GANs, where the discriminator replaces the standard reconstruction error and provides improved realism (Larsen et al., 2016). Another solution is to build a hierarchical VAE, with a more complex structure for the latent space (Vahdat & Kautz, 2020). Other methods incorporate regularization techniques, such as using a perceptual loss for the image modality, to ensure that VAE outputs have similar deep features to their corresponding inputs (Hou et al., 2019; Pihlgren et al., 2020). In this work, we focus on using a VQ-VAE model (Van den Oord et al., 2017), which is a deterministic autoencoder with a discrete latent space. In the VQ-VAE, the continuous latent vector provided by the encoder is quantized using a discrete codebook before being fed to the decoder network. The codebook is jointly learned with the network architecture. The VQ-VAE model has been shown to produce higher-quality generations than VAEs or GANs (Razavi et al., 2019). For more information about VQ-VAE, please refer to Subsection 2.9 on Page 67. Therefore, as illustrated in Figure 4.3, we propose a two-stage training approach to the MDVAE model to improve its reconstruction and generation quality.

The first stage involves learning a VQ-VAE model independently on the visual and audio modalities and without temporal modeling. The training procedure of the VQ-VAEs, including the loss functions, is the same as originally proposed in (Van den Oord et al., 2017), using an exponential moving average for the codebook updates. The VQ-VAE loss function includes a reconstruction term, which corresponds to the pixel-wise mean squared error for the visual modality and to the Itakura-Saito divergence (Févotte et al., 2009; Girin et al., 2019a) for the audio modality. In the second stage, we learn the MDVAE model on the continuous representations obtained from the pre-trained VQ-VAE encoders before quantization, instead of working directly on the raw audiovisual speech data. The disentanglement between static versus dynamic and modality-specific versus audiovisual latent factors occurs during this second training stage. This is because the VQ-VAEs are learned independently on each modality and without temporal modeling. To reconstruct the data, the continuous representations from the MDVAE are quantized and decoded by the pre-trained VQ-VAE decoders. This approach will be referred to as VQ-MDVAE in the following.

The first stage of this two-stage approach can be seen as learning audiovisual speech features in an unsupervised manner using a VQ-VAE. This feature extraction procedure is pseudo-invertible, as we can go from the raw data to the features with the VQ-VAE encoder and from the features to the raw data with the VQ-VAE decoder. This two-stage learning approach not only improves the reconstruction/generation quality, but also speeds

up the training as the MDVAE model is learned from a compressed representation of the audiovisual speech data.

4.3 Experiments on audiovisual speech

This section presents three sets of experiments conducted with the VQ-MDVAE model for audiovisual speech processing. First, we analyze qualitatively and quantitatively the learned representations by manipulating audiovisual speech sequences in the MDVAE latent space. Second, we explore the use of the VQ-MDVAE model for audiovisual facial image denoising, showing that the model effectively exploits the audio modality to reconstruct facial images where the mouth region is corrupted. Finally, we show that using the static audiovisual latent representation learned by the VQ-MDVAE model leads to state-of-the-art results for audiovisual speech emotion recognition.

4.3.1 Expressive audiovisual speech dataset

The VQ-MDVAE model is trained on the multi-view emotional audiovisual dataset (MEAD) (K. Wang et al., 2020). It contains talking faces comprising 60 actors and actresses speaking with eight different emotions at three levels of intensity. We keep only the frontal view for the visual modality. 75%, 15%, and 10% of the dataset are used respectively for the training, validation, and test, with different speakers in each split. This corresponds to approximately 25h, 5h, and 3h of audiovisual speech, respectively. For the visual modality, face images in the MEAD dataset are cropped, resized to a 64x64 resolution, and aligned using Openface (Baltrušaitis et al., 2016). For the audio modality, power spectrograms are computed using the short-time Fourier transform (STFT). The STFT parameters are chosen such that the audio frame rate is equal to the visual frame rate (30 fps), which leads to an STFT analysis window length of 64 ms (1024 samples at 16 kHz) and a hop size of 52.08% of the window length.

4.3.2 Training VQ-MDVAE

The architecture of the MDVAE and VQ-VAE models are described in detail in B.1. This section only provides an overview of the training pipeline.

The pre-processed facial images and the power spectrograms are used to train the visual and audio VQ-VAEs, respectively. The two VQ-VAEs do not include any temporal model, i.e., the audio and visual frames of an audiovisual speech sequence are processed independently. The VQ-VAE for the visual modality takes as input and outputs an RGB image of dimension $64 \times 64 \times 3$. This image is mapped by the encoder to a latent representation corresponding to a 2D grid of 8×8 codebook vectors of dimension 32. The visual codebook contains a total number of 512 vectors. The VQ-VAE for the audio modality takes as input and outputs a speech power spectrum of dimension 513. This power spectrum is mapped by the encoder to a latent representation corresponding to a 1D grid of 64 codebook vectors of dimension 8. The audio codebook contains a total number of 128 vectors. The VQ-VAEs consist of convolutional layers for both the visual and audio modalities. Since the quantization operation is non-differentiable, the codebooks for each modality are learned using the stop gradient trick (Van den Oord et al., 2017).

The audio and visual observed data $\mathbf{x}^{(a)} \in \mathbb{R}^{d_a \times T}$ and $\mathbf{x}^{(v)} \in \mathbb{R}^{d_v \times T}$ that are used to train the MDVAE model are taken from the flattened output of the pre-trained and frozen VQ-VAE encoders before quantization, with $d_a = 512$ (64×8) and $d_v = 2048$ ($8 \times 8 \times 32$). The sequence length is fixed to $T = 30$ for training. The MDVAE model is composed of dense and recurrent layers. The dimensions of the latent variables in the VQ-MDVAE model are as follows: the static latent vector ($\mathbf{w} \in \mathbb{R}^w$) has a dimension of $w = 84$, the audiovisual dynamical latent vectors ($\mathbf{z}^{(av)} \in \mathbb{R}^{l_{av} \times T}$) have a dimension of $l_{av} = 16$, and both the audio and visual dynamical latent vectors ($\mathbf{z}^{(v)} \in \mathbb{R}^{l_v \times T}, \mathbf{z}^{(a)} \in \mathbb{R}^{l_a \times T}$) have a dimension of $l_v = l_a = 8$. The models are trained using the Adam optimizer (Kingma & Ba, 2015).

Remark. The MDVAE model was trained on other modalities to demonstrate its effective generalization capabilities (different views of the visual modality, visual modality and landmarks). Refer to the Appendix 3.4.2 for further details and insights.

4.3.3 Analysis-resynthesis

We first present the results of an *analysis-resynthesis* process on the audiovisual speech data. The *analysis* step involves performing inference on audiovisual speech sequences

Table 4.2 – Speech performance of the MDVAE model tested in the *analysis-resynthesis* experiment. The STOI, PESQ, and MOSnet scores are averaged over the test subset of the MEAD dataset.

Method	STOI ↑	PESQ ↑	MOSnet ↑	SI-SDR ↑
VQ-VAE-audio	0.91 ±0.02	3.49 ±0.25	3.60 ±0.15	6.67 ±1.18
DSAE-audio	0.79 ±0.05	2.10 ±0.31	1.88 ±0.30	-1.20 ±1.58
MDVAE	0.82 ±0.03	2.90 ±0.23	2.35 ±0.18	2.21 ±1.30
VQ-DSAE-audio	0.84 ±0.03	2.12 ±0.24	3.05 ±0.20	6.12 ±1.10
VQ-MDVAE	0.85 ±0.04	2.43 ±0.28	3.54 ±0.20	6.85 ±1.15

Table 4.3 – Visual performance of the MDVAE model tested in the *analysis-resynthesis* experiment. The MSE, PSNR, SCC and SSIM scores are averaged over the test subset of the MEAD dataset.

Method	MSE ↓	PSNR ↑	SCC ↑	SSIM ↑
VQ-VAE-visual	0.0016 ±0.0002	27.2 ±0.70	0.70 ±0.01	0.85 ±0.01
DSAE-visual	0.023 ±0.03	15.8 ±2.9	0.58 ±0.07	0.47 ±0.03
MDVAE	0.010 ±0.008	20.3 ±1.3	0.62 ±0.03	0.58 ±0.03
VQ-DSAE-visual	0.0018 ±0.0005	25.3 ±1.23	0.70 ±0.01	0.82 ±0.04
VQ-MDVAE	0.0017 ±0.0007	26.8 ±0.72	0.72 ±0.01	0.84 ±0.02

that were not seen during training to obtain the latent vectors, while the *resynthesis* step involves generating the sequence from the obtained latent vectors without any modification, with the goal of faithfully reconstructing the input sequence.

Methods For this experiment, we compare MDVAE and VQ-MDVAE to VQ-VAE (Van den Oord et al., 2017) and DSAE (Y. Li & Mandt, 2018), which are unimodal generative models. DSAE also includes a temporal model that separates sequential information from static information. The VQ-VAE does not include any temporal model. The VQ-VAE and DSAE are both trained separately on the audio and visual modalities. For a fair comparison, we consider the original DSAE and its improved version VQ-DSAE obtained by training the model in two stages like VQ-MDVAE (see Section 4.2.4). This experimental comparison therefore corresponds to an ablation study: If we take VQ-MDVAE and remove the multimodal modeling we obtain VQ-DSAE. If we further remove the temporal model we obtain VQ-VAE. It will also allow us to assess the impact of the proposed two-stage training process on both DSAE and MDVAE.

Evaluation metrics The average quality performance for the speech and visual modalities is evaluated using the MEAD test dataset. Four metrics are used to assess the quality of the resynthesized audio speech data:

— The Short-Time Objective Intelligibility (STOI) measure is an intrusive metric (i.e., it requires the original reference speech signal) that assesses how intelligible the resynthesized speech is (Taal et al., 2010);
— The Perceptual Evaluation of Speech Quality (PESQ) measure is an intrusive metric that evaluates the perceived quality of the resynthesized speech (Rix et al., 2001). It accounts for factors like distortion, noise, and other artifacts that can affect the overall perceived quality;
— The Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) is an intrusive metric defined as the power ratio between the original speech signal and the distortion caused by the resynthesis process (Le Roux et al., 2019); it is made invariant to signal amplitude rescaling.
— MOSnet is a learning-based non-intrusive metric that predicts human-rated quality scores for speech (Lo et al., 2019).

Four metrics are also used to assess the quality of the resynthesized visual data:

— The Mean Square Error (MSE) computes the average squared difference between the pixel values of the original and resynthesized visual data;
— The Peak Signal-to-Noise Ratio (PSNR) considers both the image fidelity and the level of noise or distortion introduced during resynthesis;
— The Spatial Correlation Coefficient (SCC) evaluates how well the structures and patterns in the images match using the correlation;
— The Structural Similarity Index Measure (SSIM) assesses the structural similarity between the original and resynthesized images. It takes into account luminance, contrast, and structure, providing a comprehensive measure of image quality (Z. Wang et al., 2004).

Discussions Tables 4.2 and 4.3 respectively show the reconstruction quality of the audio and visual modalities for this *analysis-resynthesis* experiment. The proposed VQ-MDVAE method outperforms MDVAE alone, as evidenced by the improvement of 0.03, 0.47, 1.19, and 4.64 for STOI, PESQ, MOSnet, and SI-SDR, respectively, for the audio modality. Similarly, for the visual modality, VQ-MDVAE yields a gain of 6.5, 0.1, and 0.26 for PSNR, SCC, and SSIM, respectively. These results validate the proposed two-step training approach, demonstrating a significant improvement in reconstruction quality. This is

confirmed when comparing the results of DSAE with those of VQ-DSAE. In addition, for both modalities, MDVAE and VQ-MDVAE outperform DSAE and VQ-DSAE, respectively. However, the proposed method (VQ-MDVAE) shows a decrease in reconstruction quality compared to using the VQ-VAE alone, especially for the PESQ metric. This can be attributed to the fact that the VQ-MDVAE, with its temporal dependencies, acts as a temporal filter. Despite this, we can leverage these temporal dependencies and the hierarchy provided by the MDVAE model for other applications, as discussed in the following sections.

4.3.4 Analysis-transformation-synthesis

This section aims to analyze the latent representations learned by the MDVAE model. We want to study what high-level characteristics of audiovisual speech are encoded in the different latent variables of the model. The experiments involve exchanging latent variables between a sequence named (A) and sequences named (B) through an *analysis-transformation-synthesis process*. The analysis step involves performing inference separately on two audiovisual speech sequences (A) and (B). Then, the values of certain latent variables from (A) are replaced with the values of the same latent variable from (B). Finally, the output sequence is reconstructed from the combined set of latent variables. The resulting sequence is expected to be a mixed sequence whose features correspond to sequence (A) for the unmodified latent variables and sequence (B) for the modified latent variables.

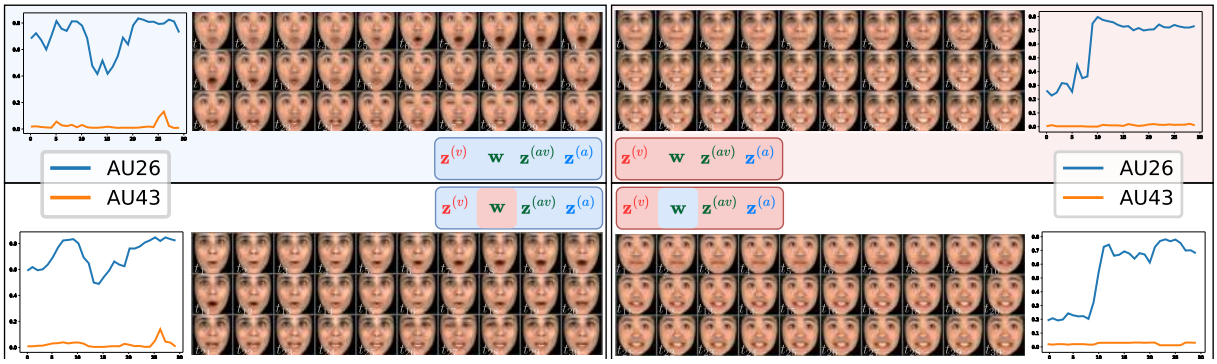


Figure 4.4 – Visual sequences generated using the *analysis-transformation-synthesis* experiment. The top two sequences depict original image sequences of two distinct individuals, while the bottom two sequences were generated by swapping the latent variable $\mathbf{w}$ between the two original sequences.

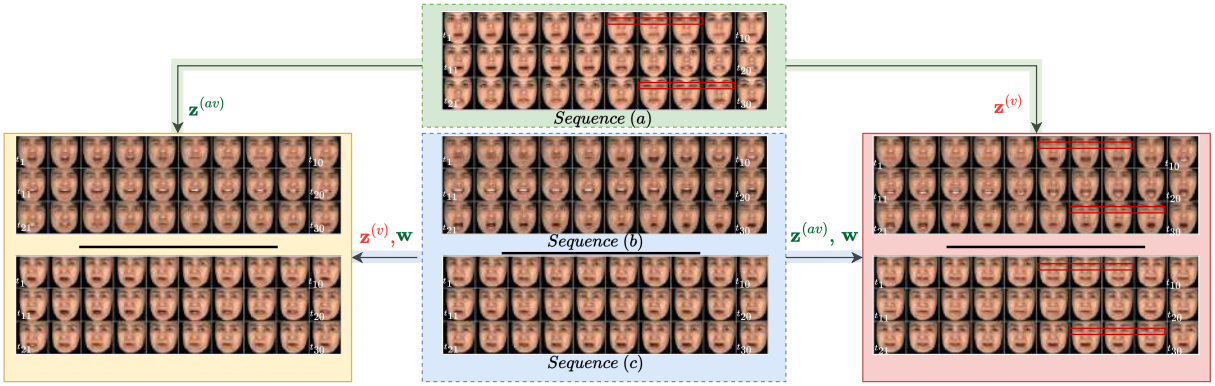


Figure 4.5 – This figure demonstrates the qualitative significance of each latent space for visual data using the *analysis-transformation-synthesis* experiment. The sequences in the yellow box (left) were generated using $\mathbf{z}^{(av)}$ from sequence (a) and $\mathbf{z}^{(v)}$, $\mathbf{w}$ from sequences (b) and (c). The sequences in the red box (right) were generated using $\mathbf{z}^{(v)}$ from the sequence (a), and $\mathbf{z}^{(av)}$, $\mathbf{w}$ from sequences (b) and (c).

Figure 4.6 – The first row represents a sequence of face images for an individual whose emotion is neutral. The rows below are generated with VQ-MDVAE, keeping all the dynamical latent variables of the first sequence and replacing the static latent variable with that of sequences from the same person but with different emotions (from top to bottom: fear, sad, surprised, angry, and happy).

Qualitative results

Visual modality Figure 4.4 illustrates visual sequences generated using the *analysis-transformation-synthesis* method, each accompanied by two curves representing the intensity of two facial action units (AUs), namely jaw drop (AU_{26}) and eyes closed (AU_{43}), plotted as a function of the frame index. AUs are the smallest components of facial

expression, involving coordinated contractions of facial muscles that produce recognizable and measurable changes in the face (Ekman & Friesen, 1978). These AUs were extracted from the visual sequences using Py-Feat (Muhammod et al., 2019). The top two sequences depict original visual sequences of different subjects exhibiting varying facial expressions. Conversely, the bottom sequences display the results when the variable $\mathbf{w}$ values are swapped between the two original sequences. We observe that the bottom-left sequence has the same facial movements as the top-left sequence, but the speaker identity is that of the top-right sequence. The curves of AU_{43} and AU_{26} for the bottom-left sequence are similar to those of the top-left sequence. A noticeable blink of the eyes occurs between frames 26 and 28, which is depicted by a peak in the AU_{43} curve. Similarly, the bottom-right sequence has the same facial movements as the top-right sequence, but the speaker identity is that of the top-left sequence. This disentanglement of dynamic facial movements from static speaker identity reveals that $\mathbf{w}$ encodes the visual identity of the speaker, among other information.

Figure 4.5 illustrates what other latent variables encode using the *analysis-transformation-synthesis* method. The figure shows three sequences of visual data, labeled as sequence *(a)* in the green box and sequences *(b)* and *(c)* in the blue box. First, two sequences on the left are reconstructed by combining $\mathbf{z}^{(av)}$ of sequence (a) with $\mathbf{w}$ and $\mathbf{z}^{(v)}$ of sequences (b) and (c). The speaker identity of sequences (b) and (c) is preserved in the output sequences, but the movement of the lips follows that of sequence (a). This shows that $\mathbf{z}^{(av)}$ encodes the lip movement. Second, two other sequences on the right are reconstructed by combining $\mathbf{z}^{(v)}$ of sequence (a) with $\mathbf{w}$ and $\mathbf{z}^{(av)}$ of sequences (b) and (c). The speaker identity and the movement of the lips of sequences (b) and (c) are preserved in the output sequences, but the movement of the eyes and eyelids (e.g., the blink of the eyes, as seen in the red rectangle) follows that of sequence (a). This indicates that $\mathbf{z}^{(v)}$ encodes eye and eyelid movements. It also appears that the head orientation in the bottom right output sequence is different from that of the original sequence (c), which was not the case for the bottom left output sequence. This indicates that $\mathbf{z}^{(v)}$ also encodes the head pose. From this example, we can also confirm that $\mathbf{w}$ encodes the speaker's identity.

Figure 4.6 shows that $\mathbf{w}$ also encodes the global emotional state. Each line in the figure is a reconstruction created by combining the dynamical latent variables of the sequence labeled as neutral in terms of emotion (first row) with $\mathbf{w}$ of other sequences of the same person labeled with different emotions (from top to bottom: fear, sad, surprised, angry, and happy). The emotion changes between the different rows, but the visual dynamics remain

the same as in the first row, indicating that the static audiovisual variable $\mathbf{w}$ encodes both the identity and the global emotion in the input sequence.

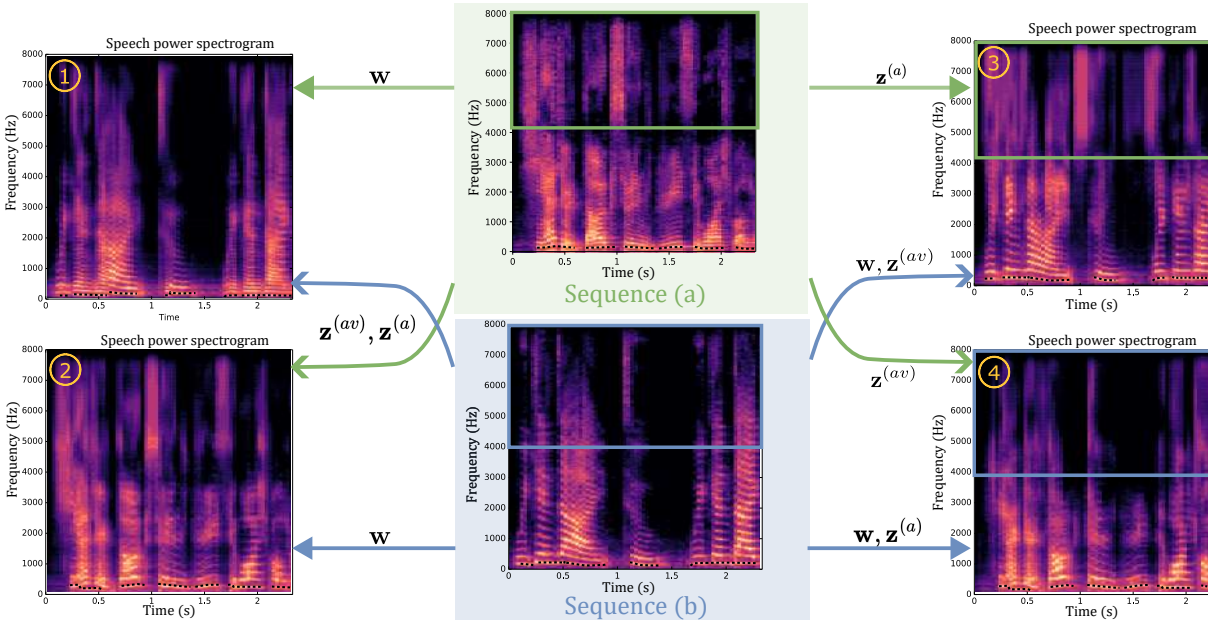


Figure 4.7 – Audio spectrograms generated from *analysis-transformation-synthesis* between sequence (a) in green and sequence (b) in blue. The spectrograms (1), (2), (3), and (4) are synthesized by swapping latent variables between sequence (a) and sequence (b). The black dotted line corresponds to the pitch contour.

Audio modality As for the visual modality, Figure 4.7 illustrates audio sequences (speech power spectrograms) generated using the *analysis-transformation-synthesis* method. In this figure, sequence (a) (green box) represents the power spectrogram and the pitch contour of a speech signal spoken by a male speaker, and sequence (b) (blue box) represents the power spectrogram and the pitch contour of a speech signal spoken by a female speaker. The pitch contour is extracted using CREPE (J. W. Kim et al., 2018). The generated spectrogram (1) (top left) is derived from $\mathbf{w}$ of sequence (a), and the dynamical latent variables $\mathbf{z}^{(av)}, \mathbf{z}^{(a)}$ of sequence (b). Comparing the resulting spectrogram with that of sequence (b), we can deduce that they have the same phonemic structure, but the pitch has been shifted downwards, as can be seen from the pitch contour and the spacing between the harmonics. Similarly, the reconstructed spectrogram (2) (bottom left) is derived from $\mathbf{w}$ of sequence (b), and the dynamical latent variables $\mathbf{z}^{(av)}, \mathbf{z}^{(a)}$ are from the sequence (a). Here, we notice that the pitch shifts upwards while preserving the phonemic structure of

sequence (a). Therefore, the static latent variable $\mathbf{w}$ encodes the average pitch value related to the speaker's identity. The generated spectrograms (3) (top right) and (4) (bottom right) reveal that the dynamical latent variables $\mathbf{z}^{(a)}$ and $\mathbf{z}^{(av)}$ have distinct roles in capturing the phonemic content. Specifically, $\mathbf{z}^{(a)}$ predominantly captures the high frequency, while $\mathbf{z}^{(av)}$ encodes the low frequency, which also corresponds to the lower formants. This finding is noteworthy as research has shown that the lower formants are highly correlated with the lip configuration (Arnela et al., 2016). Moreover, it is particularly interesting that the two correlated factors (lower formants and lip movements) are found in the same latent dynamical variable, $\mathbf{z}^{(av)}$, especially since the MDVAE was trained in an unsupervised manner.

Additional qualitative results Additional qualitative results, such as audiovisual animations, analysis-transformation-synthesis, interpolation on the static latent space, and audiovisual speech generation conditioned on specific latent variables, can be found online [1] or in Appendix B.3 and B.4. To demonstrate the generalizability, we have trained the proposed model on other modalities, as shown in Appendix B.5. A graphical user interface analysis-transformation-synthesis process has also been developed (see Appendix D.2).

Quantitative Results

The aim of this section is to complement the above qualitative analysis with quantitative metrics by measuring the ability of the VQ-MDVAE model to modify facial and vocal attributes through manipulations of the different latent variables.

Experimental setup and metrics The evaluation protocol for this experiment involves using a sequence (labeled as (A)) and 50 other sequences selected randomly from the test dataset (labeled as (B)). The protocol is based on the *analysis-transformation-synthesis* framework described in Section 4.3.4. It involves reconstructing sequences (B) using one of the latent variables (among $\{\mathbf{w}, \mathbf{z}^{(av)}, \mathbf{z}^{(a)}, \mathbf{z}^{(v)}\}$) taken from the sequence (A) and comparing audio and visual attributes extracted from the output sequences to the same attributes extracted from the original sequence (A). This comparison is done using the mean absolute error (MAE) and the Pearson correlation coefficient (PCC). If the MAE metric (resp. the PCC metric) is low (resp. high) for the swapping of a given latent variable, it indicates that the attribute was transferred from the sequence (A) to sequences (B); the swapped variable thus encodes the attribute. For the visual modality, the attributes being considered include the action units (ranging from 0 (not activated) to 1 (very

1. https://samsad35.github.io/site-mdvae/

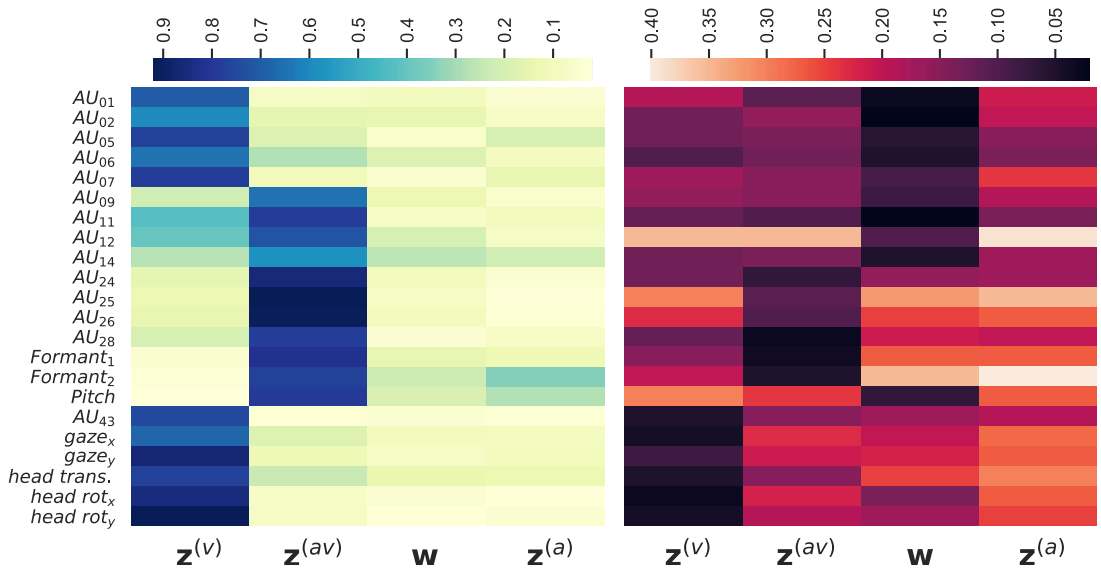


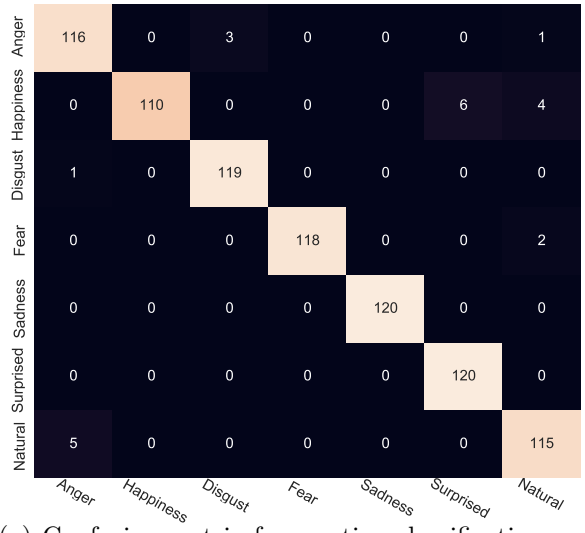
Figure 4.8 – Relationship between the audio/visual attributes and the latent variables of VQ-MDVAE. (left) Pearson correlation coefficient (PCC), (right) mean absolute error (MAE).

activated)), the angle of the gaze, and the head pose. These factors are estimated using Py-Feat (Muhammod et al., 2019) and Openface (Baltrušaitis et al., 2016). For the audio modality, we consider the first two formant frequencies (in Hz) and the pitch (in Hz), estimated using Praat (Boersma & Weenink, 2021b) and CREPE (J. W. Kim et al., 2018). Note that all these attributes are time-varying. The PCC is computed after centering the data (by subtracting the time average of the factor), which is not the case for the MAE. Therefore, contrary to the MAE, the PCC will not be affected by a time-invariant shift of the attribute.

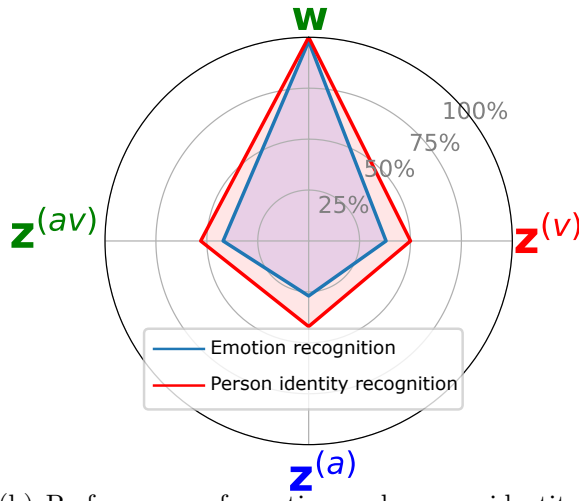
Discussion Figure 4.8 presents the average results obtained by repeating the protocol 50 times, i.e., using 50 different sequences (A). From this figure, we draw four main conclusions. First, the action units related to the lips and jaw (lip press AU_{24}, lip parts AU_{25}, jaw drop AU_{26}, and lip suction AU_{28}) and the first two formant frequencies all show high PCC values and low MAE values when performing transformations with the latent variable $\mathbf{z}^{(av)}$. It indicates that this audiovisual dynamical latent variable plays a significant role in globally controlling these factors. This is very interesting, considering that the lips and jaw are two important speech articulators whose movement induces variations of the shape of the vocal tract and thus also variations of the formant frequencies (the resonance frequencies of the vocal tract). The VQ-MDVAE model thus managed to encode highly-correlated visual and

audio factors in the same audiovisual dynamical latent variable. Secondly, the pitch factor shows a high PCC value when manipulating the dynamical audiovisual latent variable $\mathbf{z}^{(av)}$. However, it shows a low MAE value when manipulating the static audiovisual latent variable $\mathbf{w}$. This indicates that $\mathbf{w}$ encodes the average pitch value while $\mathbf{z}^{(av)}$ captures the temporal variation of the pitch around this center value (we remind that the PCC is computed from centered data but not the MAE). The fluctuations in pitch around the average value are encoded in the audiovisual latent variable $\mathbf{z}^{(av)}$ rather than the audio-specific one $\mathbf{z}^{(a)}$. This finding is supported by a recent study (Berry et al., 2022) that demonstrates a significant correlation between pitch and the lowering of the jaw. Then, the action unit associated with the closing of the eyes (AU_{43}), the angle of the gaze as well as the pose of the head show a high PCC and low MAE when manipulating the visual dynamical latent variable $\mathbf{z}^{(v)}$. This suggests that $\mathbf{z}^{(v)}$ plays a significant role in globally controlling the movement of the eyelids, the gaze, and the head movements. These factors are indeed much less correlated with the audio than the lip and jaw movements, which explains why they are encoded in the visual dynamical latent variable $\mathbf{z}^{(v)}$ and not in the audiovisual dynamical latent variable $\mathbf{z}^{(av)}$. Finally, action units such as the inner brow raiser (AU_{01}), outer brow raiser (AU_{02}), upper lid raiser (AU_{05}), cheek raiser (AU_{06}), and lid tightener (AU_{07}) on one side, and nose wrinkler (AU_{09}), nasolabial deepener (AU_{11}), lip corner puller (AU_{12}), and dimpler (AU_{14}) on the other side, show high PCC values with respect to $\mathbf{z}^{(v)}$ and $\mathbf{z}^{(av)}$, respectively, but low MAE values with respect to $\mathbf{w}$. We argue that this result is related to the encoding of the speaker's emotional state in the latent space of the VQ-MDVAE model. Indeed, we have shown qualitatively that the static audiovisual latent variable $\mathbf{w}$ encodes the global emotional state of a speaker, which explains why it also encodes the average activation level (as indicated by the low MAE values) of the above-mentioned action units that are important for emotions. In contrast, the dynamical latent variables $\mathbf{z}^{(av)}$ and $\mathbf{z}^{(v)}$ capture the temporal variations around this average value (as indicated by the high PCC values). As an illustration, we can think of an audiovisual speech utterance spoken by a happy speaker. The global emotional state (happy) would be encoded in $\mathbf{w}$, leading to high constant average values of the cheek raiser (AU_{06}) and lip corner puller (AU_{12}) action units, and these values would be modulated temporally by the movement of the speech articulators, as encoded in $\mathbf{z}^{(av)}$.

In Section 4.3.4, we showed qualitatively that the static audiovisual latent variable $\mathbf{w}$ encodes the speaker's identity and global emotion. This paragraph aims to quantify this with two complementary approaches. The first approach operates in the reconstructed



(a) Confusion matrix for emotion classification on the VQ-MDVAE output images after perturbation of the dynamical latent variables.



(b) Performance of emotion and person identity recognition for each latent variable of the VQ-MDVAE model.

Figure 4.9 – Analysis of the latent variables of the VQ-MDVAE model in terms of emotion and person identity.

image space at the output of the VQ-MDVAE model, while the second approach operates in the latent space of the model. To investigate the emotions in the VQ-MDVAE output images, we randomly select an audiovisual sequence (A) from the test data that is labeled with a specific emotion. We then perturb the dynamical latent variables of (A) by replacing them with those of sequences (B) whose emotions are different from that of sequence (A), while keeping the static audiovisual latent variable $\mathbf{w}$ of sequence (A) unchanged. We evaluate the performance of an emotion classification model (ResMaskNet (Pham et al., 2021)) on the VQ-MDVAE output images produced by this experiment and repeat the process 120 times for each emotion. The results are summarized in a confusion matrix shown in Figure 4.9(a). This matrix is mainly diagonal, indicating that as long as the static audiovisual latent variable $\mathbf{w}$ is not changed, the overall emotion is not changed. This is consistent with the discussion in the previous paragraph, where $\mathbf{w}$ was shown to control the average value of certain action units. In the second approach, we use the latent variable of the VQ-MDVAE model to recognize emotions and identities using a Support Vector Machine (SVM) classifier. The training and test datasets for the SVM comprised 70% and 30% of the combined test and validation data from the MEAD dataset, respectively. The dataset consisted of 11 speakers and included eight emotions. The performance accuracy for both classification tasks is shown in Figure 4.9(b), for different latent variables used as input

to the classifier. The results show that emotional and identity information are encoded in the static audiovisual latent variable $\mathbf{w}$, with 98% and 100% correct classification, respectively. B.2 contains visualizations of the static latent space, while the aforementioned companion website provides qualitative results of interpolations on $\mathbf{w}$, demonstrating how we can modify the emotion within an audiovisual speech sequence without altering the identity, and vice versa.

4.3.5 Audiovisual facial image denoising

Experimental set-up This section focuses on denoising audiovisual facial videos. The denoising approach consists of encoding and decoding corrupted visual speech sequences with autoencoder-based models (see next paragraph) pre-trained on the clean MEAD dataset. We intentionally introduced two types of perturbations, strategically located around the eyes and mouth. Specifically, we chose to perturb the sequences using centered isotropic Gaussian noise, and we studied the impact of different levels of noise variance. Our analysis was performed on sequences consisting of ten images, where only the six central images were corrupted.

Methods In this experiment, we compare the performance of VQ-MDVAE, which uses both audio and visual modalities and includes a hierarchical temporal model, with three other models: VQ-VAE (Van den Oord et al., 2017), a unimodal model only trained on the visual modality and without temporal modeling; DSAE (Y. Li & Mandt, 2018), a unimodal model only trained on the visual modality and with the same temporal hierarchical model as the proposed VQ-MDVAE; and JointMVAE (Suzuki et al., 2016), a multimodal model without temporal modeling. To ensure a fair comparison, we trained the DSAE and JointMVAE models in two stages, similar to the VQ-MDVAE model. It is important to mention that the VQ-VAE used in this experiment is identical to the one used in VQ-MDVAE, VQ-DSAE, and VQ-JointMVAE.

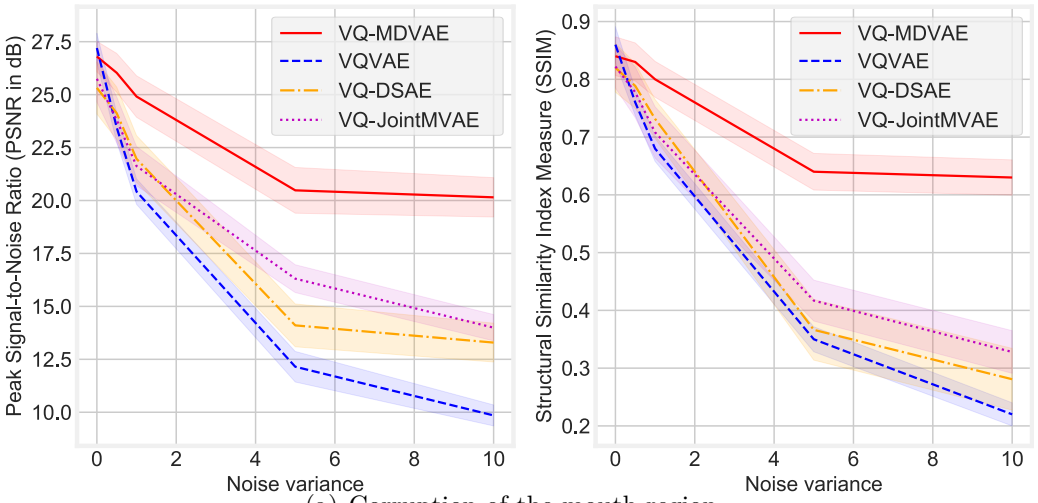
Metrics To evaluate the denoising performance, we consider again the PSNR and SSIM metrics. These are calculated on the corrupted region of the image, and provide a quantitative measure of the quality and similarity of the denoised image compared to the original. The higher the PSNR and SSIM values, the better the denoising performance.

Discussion Figures 4.10 and 4.11 present the qualitative and quantitative results for

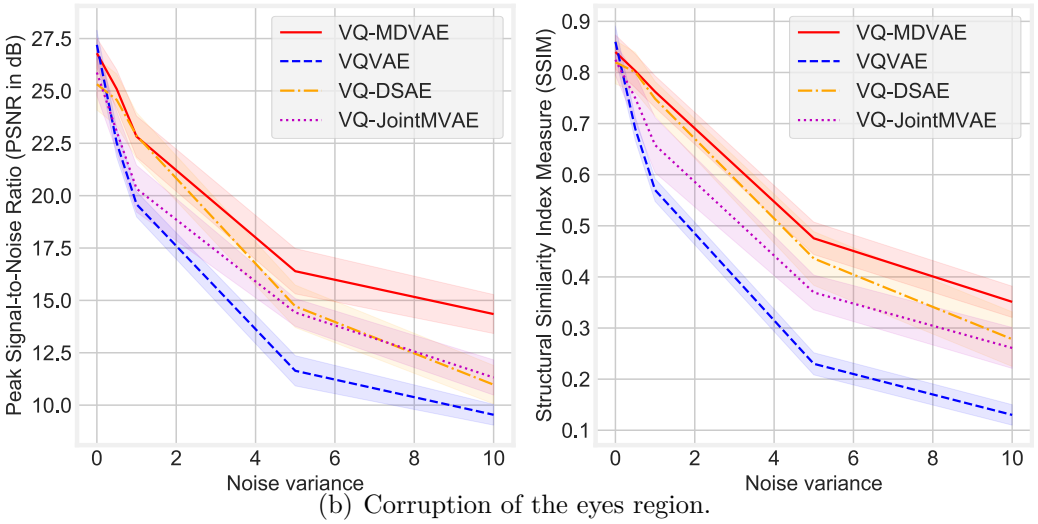
(a) Corruption of the eyes region.

(b) Corruption of the mouth region.

Figure 4.10 – Qualitative comparison of the denoising results. From top to bottom: perturbed sequences; sequences reconstructed with VQ-VAE; sequences reconstructed with DSAE; and sequences reconstructed with VQ-MDVAE.



(a) Corruption of the mouth region.



(b) Corruption of the eyes region.

Figure 4.11 – (For better visibility, please zoom in.) Quantitative results of audiovisual facial image denoising. (a) PSNR (left) and SSIM (right) are plotted as a function of the noise variance when the noise is applied to the mouth region. (b) PSNR (left) and SSIM (right) are plotted as a function of the noise variance when the noise is applied to the eyes region.

the denoising experiment, respectively. The mean and standard deviation of the metrics computed over 200 test sequences for the mouth and eyes corruptions are shown in Figures 4.11(a) and 4.11(b), respectively. Overall, the VQ-MDVAE, VQ-JointMVAE, and VQ-DSAE models outperform the VQ-VAE for both types of perturbations. In the case of mouth corruption, the VQ-MDVAE and VQ-JointMVAE models perform better than the unimodal VQ-DSAE model, demonstrating the benefit of multimodal modeling. These

models use the audio modality to denoise the mouth, resulting in a notable 7 dB increase in PSNR at a variance of 10 for VQ-MDVAE compared to VQ-DSAE. As expected, the audio modality is less useful for denoising the eyes, resulting in a smaller advantage for multimodal models in this case. In fact, the PSNR improvement with VQ-MDVAE for the corruption of the eyes is only 3 dB compared to the unimodal VQ-DSAE model. It can also be seen that VQ-MDVAE consistently outperforms VQ-JointMVAE, which shows the benefit of temporal modeling in multimodal models.

4.3.6 Audiovisual speech emotion recognition

This section presents emotion recognition experiments based on the static audiovisual representation $\mathbf{w}$ learned by VQ-MDVAE in an unsupervised manner. We consider two problems: estimating the emotion category and the emotional intensity level.

Experimental set-up We assess the effectiveness of the proposed model on two different datasets: MEAD (K. Wang et al., 2020) and RAVDESS (Livingstone & Russo, 2018). The MEAD dataset was presented in Section 4.3.1. The RAVDESS dataset contains 1440 audio files recorded by 24 professional actors, each labeled with one of eight different emotions: neutral, calm, happy, sad, angry, fearful, disgusted, or surprised. We conduct two types of evaluations to measure performance. The first evaluation involves recognizing emotions and their intensity levels where individuals can be seen during the training phase (*person-dependent evaluation*). For this evaluation, we randomly divide the dataset into 70% training data and 30% testing data. The second evaluation involves recognizing emotions and their intensity levels when individuals are not seen during the training phase (*person-independent evaluation*). To perform this evaluation, we use a 5-fold cross-validation approach to separate the speakers' identities between the training and evaluation sets. Through these evaluations, we are able to assess the ability of the models to detect emotions and their intensity levels in both person-dependent and person-independent scenarios using two different datasets.

Methods We compare the performance of VQ-MDVAE with three methods from the literature. First, the VQ-DSAE-audio and VQ-DSAE-visual models, which correspond to the VQ-DSAE model already discussed in the previous experiments, here trained either on the audio modality or on the visual modality. We remind that VQ-DSAE is an improved version of DSAE (Y. Li & Mandt, 2018) that uses the 2-stage training process proposed

in the present study. VQ-MDVAE can be seen as a multimodal extension of VQ-DSAE because both methods share the same hierarchical temporal model, including a static and a dynamical latent variable. Comparing VQ-MDVAE with the two VQ-DSAE models will thus allow us to fairly assess the benefit of a multimodal approach to emotion recognition. Second, the wav2vec model (Schneider et al., 2019), which is a self-supervised unimodal representation learning approach. Wav2vec is trained on the audio speech signals of the Librispeech dataset (Panayotov et al., 2015), which includes 960 hours of unlabeled speech data, with 2338 different speakers. Finally, we also include in this experiment two state-of-the-art supervised multimodal approaches (Chumachenko et al., 2022; Tsai et al., 2019), which are based on an audiovisual transformer architecture. The method of (Chumachenko et al., 2022) will be referred to as "AV transformer". It also relies on transfer learning using EfficientFace (Z. Zhao et al., 2021), a model pre-trained on AffectNet (Mollahosseini et al., 2017), the largest dataset of in-the-wild facial images labeled in emotions. The method of (Tsai et al., 2019) will be referred to as "MULT" for multimodal transformer.

AV transformer and MULT are fully supervised, trained, and evaluated on RAVDESS. This contrasts with wav2vec, VQ-DSAE-audio, VQ-DSAE-visual, and VQ-MDVAE, which are pre-trained in a self-supervised or unsupervised manner and then used as frozen feature extractors to train a small classification model on top of the extracted representation of (audiovisual) speech. For VQ-DSAE-audio, VQ-DSAE-visual, and VQ-MDVAE, only the global latent variable ($\mathbf{w}$) is fed to the classifier. For wav2vec, a temporal mean-pooling layer is added before the classifier as in (Pepino et al., 2021). Depending on the feature extraction method and evaluation configuration (person independent or dependent), we consider different classification models: a simple multinomial logistic regression (MLR) implemented with a single linear layer followed by a softmax activation function, or a multilayer perceptron (referred to as MLP) with two hidden layers followed by a linear layer and a softmax activation function. In the person-dependent setting, we explore a third approach (referred to as DA + MLR) that involves transforming the test data using an unsupervised domain adaptation method (DA) before classification with the MLR model. Unsupervised domain adaptation is here used to compensate for the domain shift due to the fact that speakers are different in the training and testing sets. This is further discussed below.

Discussion We start by comparing VQ-MDVAE with its two unimodal counterparts,

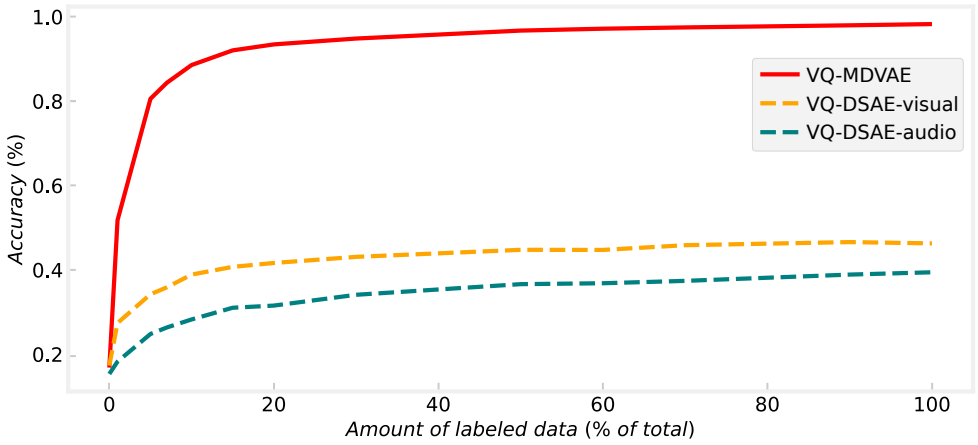


Figure 4.12 – Accuracy for emotion category classification as a function of the amount of labeled data used to train the MLR classification model on the MEAD dataset in the person-dependent evaluation setting.

VQ-DSAE-audio and VQ-DSAE-visual, for the emotion category classification task on the MEAD dataset. In Figure 4.12, we show the classification accuracy as a function of the amount of labeled training data used to train the MLR classification model. VQ-MDVAE and the VQ-DSAE models are all pre-trained in an unsupervised manner on the MEAD dataset. Using the exact same experimental protocol, we observe that when using 100% of the labeled data the VQ-MDVAE model outperforms its two unimodal counterparts by about 50% of accuracy, which clearly demonstrates the interest of a multimodal approach to emotion recognition from latent representations learned with dynamical VAEs. Another interesting observation is that we need less than 10% of the labeled data to reach 90% of the maximal performance of the VQ-MDVAE model.

Table 5.1 compares the emotion category and intensity level classification performance of the proposed VQ-MDVAE method and the previously mentioned methods from the literature. We report the accuracy (in %), defined as the ratio of correctly predicted instances to the total number of instances, and the F1-score (in %), defined as the harmonic mean of the precision and recall. For the person-dependent evaluation ("PD" section of the table), VQ-MDVAE demonstrates superior performance in recognizing emotion categories (resp. emotion levels) on the MEAD dataset, outperforming VQ-DSAE-audio by 57.8% (resp. 35.2%), VQ-DSAE-visual by 47.6% (resp. 40.6%), and wav2vec by 22% (resp. 28.5%) of accuracy. On the RAVDESS dataset, it can be observed that VQ-MDVAE pre-trained on MEAD and finetuned on RAVDESS (in an unsupervised manner) outperforms the fully-supervised state-of-the-art method (Chumachenko et al., 2022) (AV transformer) by

Table 4.4 – Accuracy (%) and F1-score (%) results of emotion category and intensity level recognition in the person-dependent (PD) and person-independent (PI) evaluation settings for the MEAD and RAVDESS datasets. The best scores are in bold and second best scores are underlined. For the VQ-MDVAE model evaluated on RAVDESS, two scores are reported. The first one corresponds to VQ-MDVAE trained on MEAD only, and the second one to the same model fine-tuned (in an unsupervised manner) on RAVDESS.

	Model		Emotion category				Emotion intensity level			
			MEAD		RAVDESS		MEAD		RAVDESS	
	classification	representation	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
PD	MLR	VQ-DSAE-audio (Y. Li & Mandt, 2018)	40.4	39.3	-	-	48.7	45.7	-	-
		VQ-DSAE-visual (Y. Li & Mandt, 2018)	50.6	51.1	-	-	43.3	44.2	-	-
		wav2vec (Schneider et al., 2019)	76.2	75.0	74.3	75.5	55.0	54.6	76.5	76.3
		VQ-MDVAE (our)	**98.2**	**98.3**	81.9 / **89.4**	82.9 / **89.6**	**83.9**	**83.1**	78.0 / **80.1**	77.2 / **79.8**
	AV transformer (Chumachenko et al., 2022)		-	-	89.2	88.6	-	-	-	-
PI	MLR	wav2vec (Schneider et al., 2019)	68.4	64.5	69.5	68.6	51.8	50.3	76.6	75.6
		VQ-MDVAE (our)	73.2	72.5	68.8 / 71.4	68.5 / 70.5	63.8	61.7	73.8 / 77.2	75.7 / 77.6
	MLP	wav2vec (Schneider et al., 2019)	70.9	70.8	70.2	70.6	53.7	53.9	76.6	76.3
		VQ-MDVAE (our)	80.0	80.5	77.5 / 78.7	78.0 / 78.1	71.5	72.2	77.4 / 77.4	77.6 / 77.7
	DA + MLR	wav2vec (Schneider et al., 2019)	71.0	69.9	71.6	71.2	53.5	52.9	76.8	76.5
		VQ-MDVAE (our)	**83.1**	**82.2**	78.1 / **79.3**	78.0 / **80.7**	**77.5**	**78.0**	78.1 / **79.0**	78.5 / **79.1**
	MULT (Tsai et al., 2019)		-	-	76.6	77.3	-	-	-	-
	AV transformer (Chumachenko et al., 2022)		-	-	79.2	78.2	-	-	-	-

0.2% of accuracy and 1.0% of F1-score. Note that the AV transformer cannot be trained simultaneously on MEAD and RAVDESS because the emotion labels in these two datasets are different. On the contrary, the proposed VQ-MDVAE model can be pre-trained on any emotional audiovisual speech dataset, precisely because it is unsupervised. The learned representation can then be used to train a supervised classification model. This evaluation confirms that the static audiovisual latent variable $\mathbf{w}$ learned by the proposed VQ-MDVAE is an effective representation for audiovisual speech emotion recognition. Indeed, as shown in Figure B.2 of B.2, emotion categories and levels form distinct clusters in the static audiovisual latent space of the VQ-MDVAE model.

For the person-independent evaluation, we only compare VQ-MDVAE, wav2vec, MULT and AV transformer, as the person-dependent evaluation showed that VQ-MDVAE outperforms its two unimodal counterparts based on VQ-DSAE. Compared with the person-dependent setting, we observe in the "PI" section of Table 5.1 a decrease in performance for all methods using an MLR classification model. For VQ-MDVAE, this decline can be analyzed through visual representations of the static audiovisual latent space, as shown in Figure B.2 of B.2. This figure highlights the hierarchical structure of the static latent audiovisual space in terms of identity, emotion, and intensity level. In this structure, identities are represented by clusters, each of which is made up of several emotion clusters.

These clusters represent eight distinct emotions distributed in a range of intensity levels from weak to strong. As a result, each identity is associated with its own representation of emotions, which means that the emotion clusters differ from one identity to another. By incorporating the identity information as in the previous evaluation approach, we can more accurately classify the emotion categories. Consequently, a simple linear model (MLR) is sufficient for classifying both the emotions and their levels. To improve generalization to test data where speakers were not seen during training, we propose two solutions. First, we improve the classification model by replacing the linear MLR classifier by a non-linear MLP classifier, which results in a substantial increase in accuracy for the VQ-MDVAE model: +6.8% and +7.3% for emotion category classification on the MEAD and RAVDESS datasets, respectively (using the finetuned model for RAVDESS). We observe a similar trend with the wav2vec + MLP model, which leads to an improvement in performance compared to using the MLR classifier. Second, we keep the MLR classification model but apply unsupervised domain adaptation to the test data using an optimal transport approach (Courty et al., 2017). Domain adaptation has been shown to be effective when dealing with domain shifts caused by unknown transformations, such as changes in identity, gender, age, ethnicity, or other factors (D. Kim & Song, 2022; Wei et al., 2018). To adapt our model to a new domain, we use optimal transport to map the probability distribution of the source domain ($\mathbf{w}$ of seen identities) to that of the target domain ($\mathbf{w}$ of unseen identities). This is accomplished by minimizing the earth mover's distance between the two distributions (Courty et al., 2017). By finding an optimal transport plan, we can transfer knowledge from the source domain to the target domain in an unsupervised manner (i.e., emotion labels are not used), resulting in a large improvement in accuracy for both the wav2vec and VQ-MDVAE models compared to when no domain adaptation is performed: +9.9% and +7.9% for emotion category classification on the MEAD and RAVDESS datasets with the VQ-MDVAE model (using the finetuned model for RAVDESS), and +2.6% and +2.1% with the wav2vec model. It can also be seen that the MLR linear classification model with domain adaptation is more effective than the MLP non-linear classification approach. Finally, for emotion category classification on RAVDESS, we see that the proposed VQ-MDVAE (finetuned) with domain adaptation and MLR outperforms the state-of-the-art fully-supervised AV transformer and MULT methods by 0.1% and 2.7% of accuracy. This is particularly interesting considering that most of the proposed model parameters have been learned in an unsupervised manner. Indeed, only the MLR classification model, which includes 680 $(84 \times 8 + 8)$ trainable parameters, is learned using

labeled emotional audiovisual speech data.

4.4 Conclusion of the chapter

Deep generative modeling is a powerful unsupervised learning paradigm that can be applied to many different data types. In this chapter, we proposed the VQ-MDVAE model to learn structured and interpretable representations of multimodal sequential data. A key to learn a meaningful representation in the proposed approach is to structure the latent space into different latent variables that disentangle static, dynamical, modality-specific and modality-common information. By defining appropriate probabilistic dependencies between the observed data and the latent variables, we could learn structured and interpretable representations in an unsupervised manner. Trained on an expressive audiovisual speech dataset, the same VQ-MDVAE model was used to address several tasks in audiovisual speech processing. This versatility contrasts with task-specific supervised models. The experiments have shown that the VQ-MDVAE model effectively combines the audio and visual information in static ($\mathbf{w}$) and dynamical ($\mathbf{z}^{(av)}$) audiovisual latent variables, while characteristics specific to each individual modality are encoded in dynamical modality-specific latent variables ($\mathbf{z}^{(a)}$ and $\mathbf{z}^{(v)}$). Indeed, we have shown that lip and jaw movements can be synthesized by transferring $\mathbf{z}^{(av)}$ from one sequence to another, while preserving the speaker's identity, emotional state and visual-only facial movements. For denoising, we have shown that the audio modality provides robustness with respect to the corruption of the visual modality on the mouth region. Finally, we proposed to use the static audiovisual latent variable $\mathbf{w}$ for emotion recognition. This approach was effective with only a few labeled data and obtained much better accuracy than unimodal baselines. Experimental results have also shown that the proposed unsupervised representation learning approach outperforms state-of-the-art fully-supervised emotion recognition methods based on an audiovisual transformer.

Unfortunately, the two modalities are not always available in audiovisual speech processing. For instance, the audio modality might be missing due to highly intrusive noise, and the visual modality might be missing due to low-lighting conditions. A robust multimodal information retrieval system should be able to handle such a situation where some modalities are temporarily missing. In the current configuration of the

MDVAE model, the proposed approach relies on both modalities for inference of the latent variables. Nevertheless, MDVAE could be extended to accommodate single-modality inference using the "sub-sampled training" approach proposed in M. Wu and Goodman, 2018, or maybe using the multimodal masking strategies proposed in Bachmann et al., 2022. Moreover, inferring all latent variables from one single modality would allow the model to be used for cross-modality generation, i.e., generating one modality given another.

Chapter 5

A VECTOR QUANTIZED MASKED AUTOENCODER FOR AUDIOVISUAL SPEECH EMOTION RECOGNITION

Contents

Summary

The previous chapter presented a multimodal dynamical VAE applied to unsupervised audiovisual speech representation learning. This chapter shifts our focus towards another learning paradigm centered around *self-supervision*. This chapter can be summarized in three points: Firstly, we introduce the VQ-MAE-AV model, which stands for a vector quantized masked autoencoder designed specifically for learning audiovisual speech representations. Secondly, we present two distinct data fusion techniques based on attention mechanisms, implemented in both the encoder and the decoder model. Lastly, we discuss the application of our method in the domain of emotion recognition tasks, where the proposed method outperforms state-of-the-art audiovisual speech emotion recognition methods.

5.1 Introduction

In recent years, advances in artificial intelligence technology and hardware acceleration have shifted towards multimodal processing in emotion recognition (Ramachandram & Taylor, 2017; Schoneveld et al., 2021; Tsai et al., 2019). Supervised learning using large annotated datasets can result in valuable representations. However, collecting a substantial amount of emotionally labeled data can be expensive, time-consuming, and sometimes impractical. Most datasets in emotion recognition tasks rely on actors to display various emotions at specific intensities, which requires considerable acquisition time and resources as discussed in Section 1.4, page 18. Unsupervised and self-supervised learning approaches are natural ways to address these issues and have been explored in recent studies (Bengio et al., 2006; L.-W. Chen & Rudnicky, 2021; Dib et al., 2023; Y. Wang et al., 2021). SSL offers the advantage of high scalability, as the SSL task can be carried out on large amounts of unlabeled speech data, reducing the need for labeled data (S. Liu et al., 2022; C. Zhang et al., 2022). These methods involve pre-training models with pretext tasks and then fine-tuning them on a smaller set of labeled data for the emotion recognition task (Gong, Lai, et al., 2022; Jegorova et al., 2023; Pepino et al., 2021).

SSL models can be broadly categorized into discriminative and generative approaches (C. Zhang et al., 2022). Discriminative SSL focuses on creating pairs or groups of data

samples and formulating loss functions that enable the model to distinguish or group these samples, which can later benefit downstream tasks (T. Chen et al., 2020; X. Chen et al., 2020). Pretext tasks can consist of solving jigsaw puzzles (M. Noroozi & Favaro, 2016) or predicting image rotations (Gidaris et al., 2018) for example. Contrastive learning has emerged as the predominant paradigm in discriminative SSL (Akbari et al., 2021; Alayrac et al., 2020). In contrast, generative SSL involves generating or reconstructing segments of unlabeled and potentially corrupted data using an autoencoder model (Bao et al., 2021; He et al., 2022; Z. Xie et al., 2022). The latent representation produced by the encoder can subsequently be used for downstream tasks.

The present study focuses on the MAE, an SSL generative model that uses an asymmetric encoder-decoder architecture with input masking (He et al., 2022). The MAE approach is inspired by masked language modeling (Devlin et al., 2019) and has been successfully applied to image modeling thanks to the development of Vision Transformers (Dosovitskiy et al., 2020). A description of the MAE paradigm is introduced in Section 2.5 on page 68. The MAE has recently been adapted for audio using a 2D time-frequency representation (Baade et al., 2022; Gong, Lai, et al., 2022; Xu et al., 2022). It was recently shown that combining the task of reconstructing masked tokens with contrastive learning can improve the representation learned by an MAE (Gong, Lai, et al., 2022; Z. Huang et al., 2022).

A recent extension of the MAE was presented in (Feichtenhofer et al., 2022; Tong et al., 2022) for modeling image sequences, called Video-MAE. It uses the same architecture as the vanilla MAE (He et al., 2022) but incorporates a masking process from video ViT (ViViT) (Arnab et al., 2021). Since videos often contain redundant information, particularly in scenes with no motion, the authors propose a cubic masking approach along the temporal axis, combined with a high masking ratio of 90%. Other works have extended the MAE to handle multimodal data (Bachmann et al., 2022; Geng et al., 2022; Gong, Rouditchenko, et al., 2022). MultiMAE (Bachmann et al., 2022) encodes a small random subset of visible tokens from multiple modalities (RGB, depth, and semantic images) and is trained to reconstruct the missing ones. M3AE (Geng et al., 2022) is a unified MAE architecture for two input modalities (image and text). The main difference in architecture between M3AE and MultiMAE lies in the architecture of the decoder. The MulitMAE approach introduces individual decoders for each modality, enhancing their fusion by incorporating a cross-attention layer at the outset of each decoder. On the other hand, the M3AE approach uses a single decoder that takes the concatenated tokens from all modalities.

In the literature, MAEs are typically trained using the *L1* or *L2* losses, which can

negatively affect the reconstruction quality of the masked tokens, resulting, for instance, in blurred or noisy images or sounds. It has been shown that improving the quality of MAE reconstructions can be beneficial in terms of downstream task performance (He et al., 2022). Several approaches have been proposed in that sense, which include adding a perceptual loss (Dong et al., 2021) or using discrete representations obtained from vector quantized generative adversarial networks (VQ-GANs) (Esser et al., 2021) or variational autoencoders (VQ-VAEs) (Van den Oord et al., 2017) to train the MAE (T. Li et al., 2022; Sadok, Leglaive, & Séguier, 2023). These works only considered an unimodal setting, while the present study proposes a multimodal MAE for audiovisual speech representation learning.

We introduce the VQ-MAE-AV model, a vector quantized MAE for audiovisual speech representation learning applied to audiovisual SER. The overall architecture of the model is illustrated in Figure 5.1 and 5.2. To the best of our knowledge, this is the first multimodal SSL approach based on MAEs that is proposed for audiovisual SER. Moreover, unlike existing multimodal MAEs proposed in other application contexts, the VQ-MAE-AV model operates on the discrete latent representation of the modalities and introduces global tokens that are learned using contrastive learning to capture the overall sequence-wise information from each modality. The VQ-MAE-AV model is pre-trained on the VoxCeleb2 dataset (Chung et al., 2018) and fine-tuned on three standard emotional audiovisual speech datasets. The experimental results show that the proposed VQ-MAE-AV model consistently outperforms the state-of-the-art audiovisual SER methods across the three test datasets. Extensive ablation experiments are also presented to study the impact of different model designs. The code and qualitative results are available online [1].

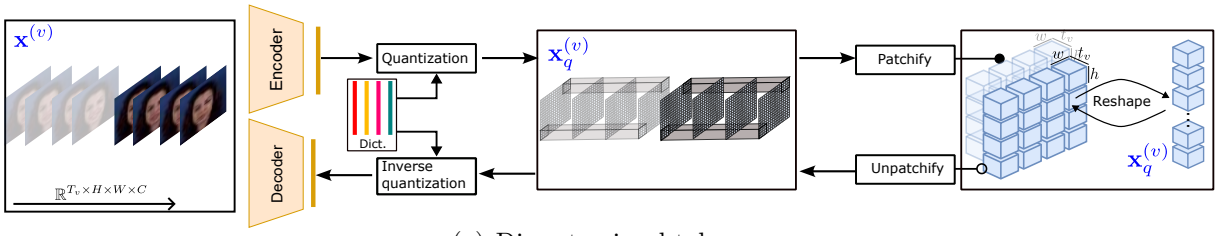
5.2 The VQ-MAE-AV model

This section presents the VQ-MAE-AV model. The overall approach is illustrated in Figure 5.1 and 5.2 and it can be summarized as follows:

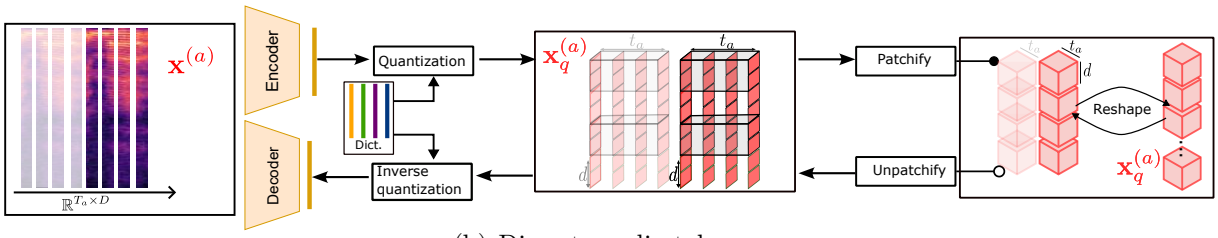
— Fully-convolutional VQ-VAEs are trained independently on the audio and visual modalities (see Section 5.2.1);

— Discrete audio and visual tokens are built from the quantized representations provided by the frozen VQ-VAE encoders (see Section 5.2.2);

1. https://anonymous-35.github.io/VQ-MAE-AV



(a) Discrete visual tokens.



(b) Discrete audio tokens.

Figure 5.1 – Discrete audio and visual tokens creation: (i) fully-convolutional VQ-VAEs are trained independently on the audio and visual modalities (see Section 5.2.1); (ii) discrete audio and visual tokens are built from the quantized representations provided by the frozen VQ-VAE encoders (see Section 5.2.2).

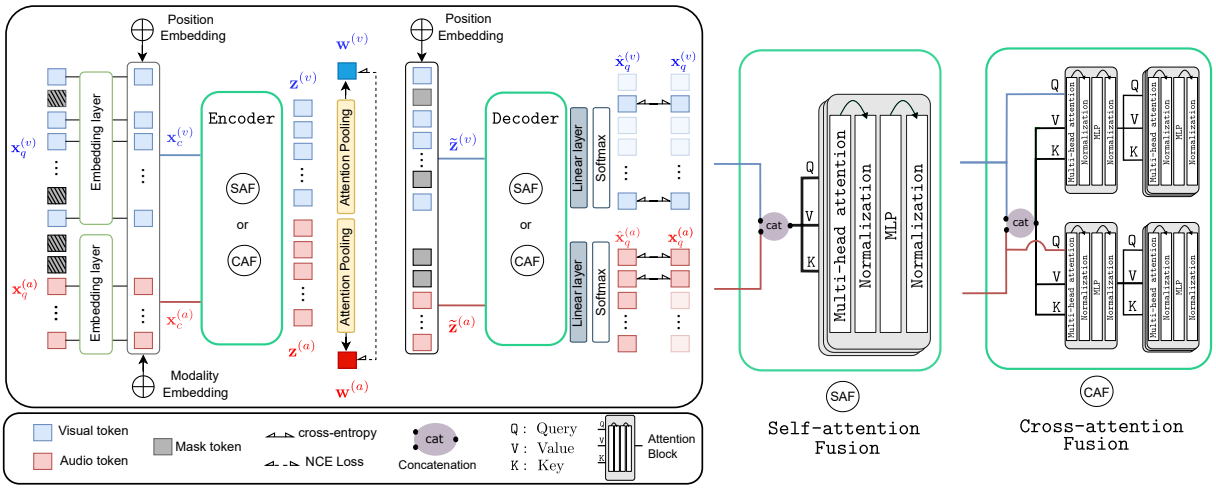


Figure 5.2 – VQ-MAE-AV model structure. See the first paragraph of Section 5.2 for a complete description of the pipeline.

— A proportion of the discrete audio and visual tokens is masked out, using a coupled masking strategy between the two modalities (see Section 5.2.3);

— The visible audio and visual tokens are replaced with trainable continuous embedding vectors (see Section 5.2.4), which are fed to the VQ-MAE-AV encoder (see Sections 5.2.5 and 5.2.5, where we present two strategies based on attention

mechanisms to fuse the modalities);
- Attention pooling is used to compute global sequence-wise tokens that are specific to each modality (see Section 5.2.5);
- The token-wise representation obtained from the encoder is combined with mask tokens and fed to the VQ-MAE-AV decoder, which tries to reconstruct the original non-masked discrete audio and visual tokens (see Section 5.2.5);
- The VQ-MAE-AV model is trained in a self-supervised manner to minimize (i) the cross-entropy loss between the reconstructed and original tokens and (ii) a contrastive loss between the audio and visual global tokens (see Section 5.2.6);
- After self-supervised learning, the VQ-MAE-AV encoder and attention pooling layers are fine-tuned for supervised audiovisual SER (see Section 5.2.7).

This section will present each above-listed aspect of the model in more detail.

5.2.1 Vector quantized variational autoencoder

The proposed multimodal self-supervised approach uses the discrete latent representation of two pre-trained and frozen VQ-VAEs (Van den Oord et al., 2017). For more information about VQ-VAE, please consult Subsection 2.9 on Page 67. Specifically, as illustrated in Figure 5.1, we use the VQ-VAE-audio and VQ-VAE-visual encoders to obtain compressed and quantized representations of the input speech power spectrogram $\mathbf{x}^{(a)} \in \mathbb{R}^{T_a \times D}$ where T_a and D correspond to the time and frequency dimensions, and of the input image sequence $\mathbf{x}^{(v)} \in \mathbb{R}^{T_v \times H \times W \times C}$ where T_v, H, W and C correspond to the time, height, width, and channel dimensions. The audio and visual quantized representations are denoted by $\mathbf{x}_q^{(a)} \in \mathbb{Z}^{T_a \times D'}$ and $\mathbf{x}_q^{(v)} \in \mathbb{Z}^{T_v \times H' \times W'}$, respectively. Each entry of $\mathbf{x}_q^{(a)}$ and $\mathbf{x}_q^{(v)}$ corresponds to the index of a vector in the VQ-VAE codebooks. Notably, $\mathbf{x}_q^{(a)}$ retains the time-frequency structure of the original spectrogram, while $\mathbf{x}_q^{(v)}$ retains the spatio-temporal structure of the original sequence images. This is because the VQ-VAE-audio and VQ-VAE-visual models are designed to be fully convolutional on the frequency and spatial axes, respectively, and they process the frames within a sequence independently. Therefore, compression occurs along the frequency axis ($D' \ll D$) for $\mathbf{x}_q^{(a)}$ and along the x and y-axes of the image ($H' \ll H$, $W' \ll W$) for $\mathbf{x}_q^{(v)}$. As shown in Figure 5.2 and discussed in the following subsections, the proposed MAE-based self-supervised learning approach operates on these discrete and compressed representations before audiovisual speech reconstruction using the VQ-VAE decoders.

The training procedure of the VQ-VAEs follows the original approach presented in (Van

den Oord et al., 2017). In particular, the VQ-VAE loss functions involve a reconstruction term between the original and reconstructed data, which corresponds to the mean squared error for the visual modality and to the Itakura-Saito divergence for the audio modality (Févotte et al., 2009). More details are provided in Section 5.3.1.

5.2.2 Discrete audio and visual tokens

As shown in Figure 5.1, the audio and visual quantized representations $\mathbf{x}_q^{(a)} \in \mathbb{Z}^{T_a \times D'}$ and $\mathbf{x}_q^{(v)} \in \mathbb{Z}^{T_v \times H' \times W'}$ from the output of the VQ-VAE encoders are divided into non-overlapping patches to build discrete tokens $\mathbf{x}_q^{(a)} \in \mathbb{Z}^{(n_{t_a} \cdot t_a) \times (n_d \cdot d)}$ and $\mathbf{x}_q^{(v)} \in \mathbb{Z}^{(n_{t_v} \cdot t_v) \times (n_h \cdot h) \times (n_w \cdot w)}$, where $T_a = n_{t_a} \cdot t_a$, $D' = n_d \cdot d$, $T_v = n_{t_v} \cdot t_v$, $H' = n_h \cdot h$, and $W' = n_w \cdot w$. These representations are reshaped to $\mathbf{x}_q^{(a)} \in \mathbb{Z}^{(n_{t_a} \cdot n_d) \times (t_a \cdot d)}$ and $\mathbf{x}_q^{(v)} \in \mathbb{Z}^{(n_{t_v} \cdot n_h \cdot n_w) \times (t_v \cdot h \cdot w)}$, which are seen as sequences of $n_{t_a} \cdot n_d$ and $n_{t_v} \cdot n_h \cdot n_w$ tokens of dimension $t_a \cdot d$ and $t_v \cdot h \cdot w$, respectively.

The use of discrete audio and visual tokens in VQ-MAE-AV has several motivations. Firstly, by dividing the audiovisual data into spatio-spectro-temporal patches, the method could learn to relate audio tokens to visual tokens. For example, the audio tokens are expected to correlate strongly with the visual tokens corresponding to the mouth area (Arnela et al., 2016; Sadok, Leglaive, Girin, et al., 2023b). The proposed method can potentially learn a representation that captures shared and distinctive information between the two modalities by exploiting their complementarity. Additionally, the use of discrete tokens can reduce the computational cost of the method as it involves working with a reduced representation of the data, which allows us to increase the number of tokens (i.e., manipulate longer audiovisual speech sequences) without exploding in the number of trainable parameters compared to the multimodal MAE in the literature (Bachmann et al., 2022).

5.2.3 Masking

We apply masking to the audio and visual sequences of tokens. The MAE aims to reconstruct the masked tokens from the visible tokens to learn a semantic representation of the data. The masking strategy impacts the performance of downstream tasks. In the original MAE, the masked tokens are randomly drawn with a target ratio of the entire set of tokens, typically 75%. In multimodal scenarios, the basic strategy would be to do the same for each modality, but it has been shown that it is more interesting to implement a

coupled masking strategy between the modalities (Bachmann et al., 2022). This involves randomly drawing a masking proportion $p_a \in [0, 1]$ for the audio modality and $p_v \in [0, 1]$ for the visual modality such that $p_a + p_v = 1$. To do this, a Dirichlet distribution is used: $(p_a, p_v) \sim Dir(\alpha_a, \alpha_v)$ where $\alpha_a, \alpha_v > 0$ are the concentration parameters of the distribution. When $\alpha_a = \alpha_v = 1$, the distribution is uniform overall points in its support (the 1-simplex), and $\alpha_a \gg \alpha_b$ results in a sampling behavior where most of the tokens are taken from the audio modality, and vice versa. These parameters can be used to reduce the dominance of one modality over the other. This strategy allows the reconstruction of missing information from one modality by relying on another. Therefore, the model will leverage the less masked modality for the reconstruction of the more masked one. For this study, we set $\alpha_a = \alpha_v = 1$.

5.2.4 Continuous embedding vectors

The discrete tokens correspond to the indices obtained through the quantization step of the pretrained VQ-VAE encoder. Before being input to the VQ-MAE-AV encoder, these discrete tokens are replaced with trainable continuous embedding vectors taken from an audio codebook in $\mathbb{R}^{k_a \times e_a}$ and from a visual codebook in $\mathbb{R}^{k_v \times e_v}$, where $k_{a/v}$ is the number of codes in the codebook and $e_{a/v}$ is the dimension of each code. This is simply achieved by replacing the indices of a discrete token with the corresponding vectors of dimension $e_{a/v}$ in the codebook. After this embedding process, the sequences of discrete tokens $\mathbf{x}_q^{(a)} \in \mathbb{Z}^{(n_{t_a} \cdot n_d) \times (t_a \cdot d)}$ and $\mathbf{x}_q^{(v)}$ in $\mathbb{Z}^{(n_{t_v} \cdot n_h \cdot n_w) \times (t_v \cdot h \cdot w)}$ are transformed into sequences of continuous tokens $\mathbf{x}_c^{(a)} \in \mathbb{R}^{(n_{t_a} \cdot n_d) \times (t_a \cdot d \cdot e_a)}$ and $\mathbf{x}_c^{(v)} \in \mathbb{Z}^{(n_{t_v} \cdot n_h \cdot n_w) \times (t_v \cdot h \cdot w \cdot e_v)}$.

5.2.5 VQ-MAE-AV encoder and decoder

Attention Block

The VQ-MAE-AV encoder and decoder are built with multi-head Attention blocks similar to those used in the Vision Transformer (ViT) (Dosovitskiy et al., 2020). Each block comprises a multi-head attention layer, normalization layers, and a Multi-layer Perceptron (MLP). These layers are interconnected with residual connections, as depicted in Figure 5.2, and they will be used to capture the inter and intra-relationships between audio and visual tokens. This attention block is inspired by the attention layer in the original transformer (Vaswani et al., 2017). To simplify the reading afterward, we denote the attention block by $\texttt{Attention}(Q, V, K)$, where Q, V, K are the query, value, and key,

respectively. The self-attention mechanism uses the same input vector for the query, key, and value vectors. In the case of cross-attention, the query and key are different to enable attention across multiple modalities *or* inputs.

Encoders

We propose two fusion strategies of the audio and visual speech data, resulting in two architectures for the VQ-MAE-AV encoder. The first fusion strategy is called *self-attention fusion*(SAF). As represented in Figure 5.2, this fusion consists of a concatenation of token sequences for the two modalities, followed by L self-attention blocks. This concatenation operates on the first dimension of the variables, i.e., we obtain a sequence of $(n_{t_a} \cdot n_d) + (n_{t_v} \cdot n_h \cdot n_w)$ tokens after concatenation.

The second fusion strategy is called *cross-attention fusion* (CAF). As shown in Figure 5.2, the sequences of audio and visual tokens are used separately as the queries of two separate attention blocks, which share the same keys and values corresponding to the concatenation of the modalities. These two cross-attention blocks are then followed by a stack of L self-attention blocks.

For both fusion strategies, the encoder outputs one sequence of tokens for each modality, denoted by $\mathbf{z}^{(a)}$ and $\mathbf{z}^{(v)}$.

Global tokens

In addition to the token-wise representations $\mathbf{z}^{(a)}$ and $\mathbf{z}^{(v)}$, we learn two sequence-wise global tokens, denoted by $\mathbf{w}^{(a)} \in \mathbb{R}^{1\times(t_a \cdot d \cdot e_a)}$ and $\mathbf{w}^{(v)} \in \mathbb{R}^{1\times(t_v \cdot h \cdot w \cdot e_v)}$, which can be thought of as similar to [CLS] tokens (He et al., 2022). These global modality-specific tokens are introduced to aggregate the spectro-temporal and spatio-temporal information in the two modalities, which can be useful for downstream tasks involving predictions at the sequence level, such as audiovisual SER. The global tokens are computed as follows:

$$\mathbf{w}^{(a)} = \texttt{Attention}\left(Q_{(a)}, V_{(a)}, K_{(a)}\right); \tag{5.1}$$

$$\mathbf{w}^{(v)} = \texttt{Attention}\left(Q_{(v)}, V_{(v)}, K_{(v)}\right), \tag{5.2}$$

where $Q_{(a)} = \mathcal{Z}^{(a)} \in \mathbb{R}^{1\times(t_a \cdot d \cdot e_a)}$ and $Q_{(v)} = \mathcal{Z}^{(v)} \in \mathbb{R}^{1\times(t_v \cdot h \cdot w \cdot e_v)}$ represent respectively trainable audio and visual tokens, as proposed in Touvron et al., 2021, $V_{(a)} = K_{(a)} = \mathbf{z}^{(a)}$, and $V_{(v)} = K_{(v)} = \mathbf{z}^{(v)}$.

Decoders

The token-wise representation obtained from the encoder is combined with mask tokens and fed to the VQ-MAE-AV decoder along with additional position embeddings, as denoted by $\tilde{\mathbf{z}}^{(a)}$ and $\tilde{\mathbf{z}}^{(v)}$ and illustrated in Figure 5.2. The mask tokens actually correspond to one single trainable vector as proposed in the original MAE (He et al., 2022). Similarly as for the encoder, the audio and visual inputs of the VQ-MAE-AV decoder can be fused using either self-attention fusion or cross-attention fusion. The number of attention blocks L' in the decoder is chosen to be lower compared to that of the encoder ($L' < L$).

A linear layer is added at the end of the decoder, which maps to the size of the VQ-VAE codebooks. The output of this linear layer corresponds to the logits of the discrete tokens. After applying an *softmax* operation, we obtain reconstructions $\hat{\mathbf{x}}_q^{(a)}$ and $\hat{\mathbf{x}}_q^{(v)}$ of the indices $\mathbf{x}_q^{(a)}$ and $\mathbf{x}_q^{(v)}$ that were provided by the VQ-VAE-audio and VQ-VAE-visual encoders, respectively.

5.2.6 VQ-MAE-AV loss functions

Generative loss function

To train the VQ-MAE-AV model, we minimize the cross-entropy loss applied only to the masked discrete tokens:

$$\mathcal{L}_{rec} = \texttt{cross-entropy}\left(\mathbf{x}_q^{(a)}\left(\Omega_{\mathcal{M}}^{(a)}\right),\ \hat{\mathbf{x}}_q^{(a)}\left(\Omega_{\mathcal{M}}^{(a)}\right)\right) + \texttt{cross-entropy}\left(\mathbf{x}_q^{(v)}\left(\Omega_{\mathcal{M}}^{(v)}\right),\ \hat{\mathbf{x}}_q^{(v)}\left(\Omega_{\mathcal{M}}^{(v)}\right)\right) \tag{5.3}$$

where $\mathbf{x}\left(\Omega_{\mathcal{M}}^{(\cdot)}\right)$ denotes the set of masked tokens in $\mathbf{x}$. Another benefit of manipulating discrete representations for multimodal inputs is the homogeneity of the losses, which does not require balancing the losses between the two modalities.

Contrastive loss function

Building upon the approaches presented in (Akbari et al., 2021; Alayrac et al., 2020), the global tokens are learned using noise contrastive estimation. This approach enhances the alignment of audiovisual speech pairs by grouping together embeddings that belong to the same time sequence and separating them from those that do not correspond to the same sequence. This approach involves minimizing the loss function $\mathcal{L}_{NCE}(\mathbf{w}^{(a)}, \mathbf{w}^{(v)})$,

which is defined by:

$$\mathcal{L}_{NCE}(\mathbf{u}, \mathbf{v}) = -\log\left(\frac{\exp\left(\mathbf{u}^\top \mathbf{v}/\tau\right)}{\exp\left(\mathbf{u}^\top \mathbf{v}/\tau\right) + \sum\limits_{(\mathbf{u}',\ \mathbf{v}')\in\mathcal{N}} \exp\left(\mathbf{u}'^\top \mathbf{v}'/\tau\right)}\right) \tag{5.4}$$

To form positive pairs ($\mathbf{w}^{(a)}$, $\mathbf{w}^{(v)}$) for both audio and visual modalities, we select corresponding streams from the same temporal location in the video. Conversely, negative pairs ($\mathbf{w}'^{(a)}$, $\mathbf{w}'^{(v)}$) are formed by selecting *non*-corresponding streams drawn from a set $\mathcal{N}$ of different temporal locations for each batch. The sensitivity of the NCE loss in distinguishing between positive and negative pairs is regulated by a temperature parameter τ.

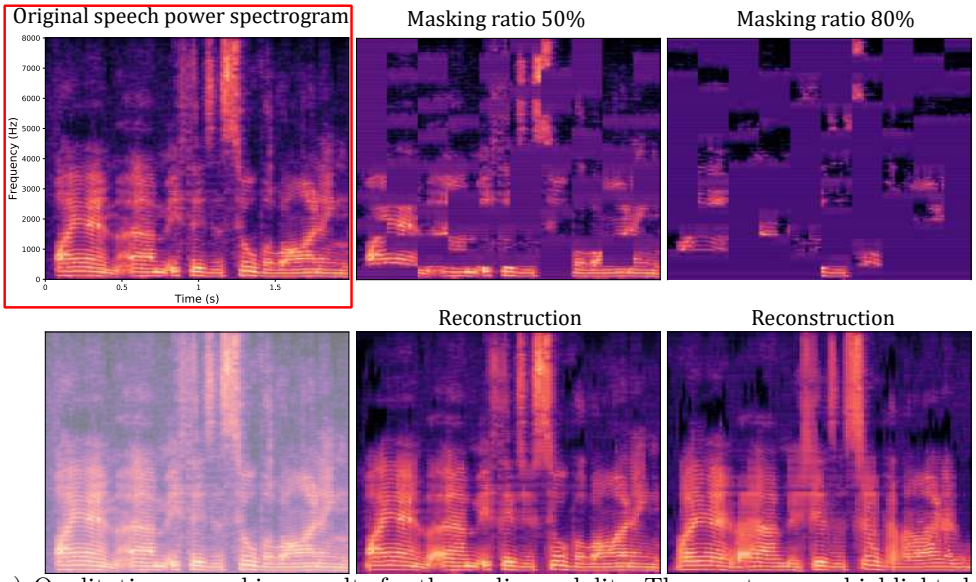
Final loss function

The final loss function corresponds to the sum of the generative loss function in (5.3) and of the contrastive loss function in (5.4):

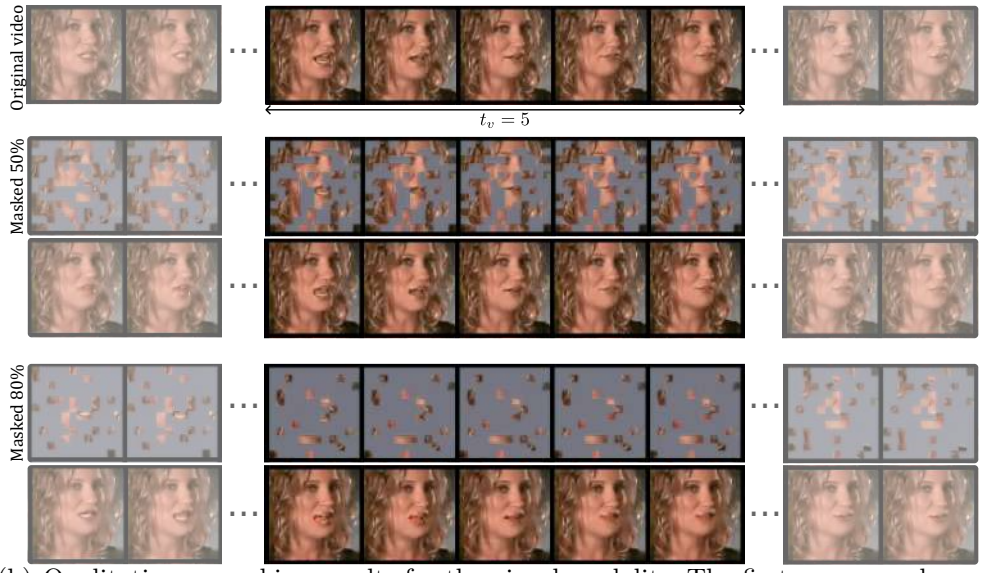
$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{NCE} \tag{5.5}$$

5.2.7 Fine-tuning for audiovisual SER

After pre-training VQ-MAE-AV on token unmasking, the model is fine-tuned for audiovisual SER. We propose three different approaches for fine-tuning the model. The first two approaches rely on a pooling operation (of the encoder outputs $\mathbf{z}^{(a)}$ and $\mathbf{z}^{(v)}$) to extract the global audio and visual tokens, which is either a simple *mean pooling* or the *attention pooling* discussed in Section 5.2.5. The extracted tokens are then concatenated and passed through a single linear layer, followed by an *argmax* operation to classify the emotion. The third proposed approach, referred to as *Query2Emo* and inspired by (S. Liu et al., 2021), involves cross-attention between all audio and visual tokens (concatenation of $\mathbf{z}^{(a)}$, $\mathbf{z}^{(v)}$) as key and value, and the emotion classes represented by trainable embeddings as the query. Query2Emo has a single attention block for both the encoder and decoder. The extracted embeddings are then concatenated and passed through a single linear layer. We adopt the asymmetric loss (Ridnik et al., 2021) for all these approaches between the predicted emotion and the ground-truth emotion.



(a) Qualitative unmasking results for the audio modality. The spectrogram highlighted in the red box represents the original spectrogram. The two spectrograms on the top right represent the spectrograms masked at 50% and 80%, respectively. The reconstructions using VQ-MAE-AV can be seen directly below these masked spectrograms.



(b) Qualitative unmasking results for the visual modality. The first sequence shows the original video, followed by the next two sequences representing the masked video with a ratio of 50% and its reconstruction using VQ-MAE-AV. The last two sequences represent the masked video with a ratio of 80% and its reconstruction using VQ-MAE-AV.

Figure 5.3 – Quantitative results of the audio reconstruction (a) and visual reconstruction (b) using VQ-MAE-*12*.

5.3 Experiments

This section presents our study's experimental setup and results, as well as a discussion of their implications. We evaluate the effectiveness of the proposed approach for emotion recognition on three standard audiovisual speech databases and compare it with state-of-the-art methods. Additionally, we perform an ablation study to analyze the impact of various hyperparameters and architectures on the performance of our method.

We present qualitative audio and visual reconstruction results in Figure 5.3(a) and Figure 5.3(b), respectively. Furthermore, we conducted a study to measure the reconstruction quality at different masking ratios of the audiovisual speech data. This analysis is detailed in Appendix C.1 and demonstrates the effectiveness of leveraging multimodality to enhance reconstruction quality. Additional qualitative results can also be found on the website [2].

Remark. A preliminary study focusing solely on audio speech modality has been carried out, and the results were published in a workshop (Sadok, Leglaive, & Séguier, 2023). A subsequent study is under development to *generalize and improve* the approach introduced in this chapter on other types of input data, including visual modality and action units. Detailed information on these two studies is available in appendix C.3 and C.4, respectively.

5.3.1 Experimental setup

Datasets and preprocessing

Dataset for self-supervised training To pre-train VQ-MAE-AV, we use the VoxCeleb2 dataset (Chung et al., 2018), which offers a broad range of audiovisual speech data from open-source media, with each video featuring a single speaker. We restricted our dataset use to a subset of around 1000 hours of audiovisual speech, encompassing 2170 different speakers. The test set includes about 100 hours of audiovisual speech data, with 117 different speakers.

Data pre-processing The VQ-VAE-audio and VQ-VAE-visual models are trained on the VoxCeleb2 dataset. The former is trained on STFT power spectrograms ($\mathbf{x}^{(a)}$), while the latter uses RGB sequence images ($\mathbf{x}^{(v)}$) that are cropped on the face to

2. https://samsad35.github.io/VQ-MAE-AudioVisual

reach a resolution of 96×96. To compute the STFT, a Hann window of 64 ms (1024 samples at 16 kHz) and a 68% overlap are used, resulting in a sample rate of 50 Hz, which is twice the sample rate of the visual modality. This leads to sequences of $D = 513$ Fourier coefficients.

Emotional audiovisual speech databases We fine-tune and evaluate the proposed approaches on three emotional audiovisual speech databases.

— **RAVDESS** (Livingstone & Russo, 2018): This English database consists of 1440 videos recorded by 24 professional actors and labeled with eight different emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprised).
— **CREMA-D** (H. Cao et al., 2014): It is an English dataset of 7442 videos recorded by 91 actors. Actors spoke from a selection of 12 sentences. The sentences used one of six emotions (anger, disgust, fear, happy, neutral, and sad) with four different intensities (low, medium, high, and unspecified).
— **eNTERFACE05** (Livingstone & Russo, 2018): The audiovisual acted dataset comprises recordings of six distinct emotions: anger, disgust, fear, joy, sadness, and surprise. These emotional expressions were eliparencited from a diverse group of 43 individuals representing 14 nationalities. The dataset encompasses a total of 1290 video samples. Each participant was instructed to listen to six short stories that eliparencited a specific emotion.

We selected these databases as they are commonly used for the emotion recognition task, and their raw data is accessible. To ensure a fair comparison with previous works, we performed *6*-fold, *10*-fold and *5*-fold cross-validation for the RAVDESS, CREAM-D and eNTERFACE05 datasets, respectively. We carefully partitioned the datasets to ensure speaker identity separation between fine-tuning and evaluation phases.

Model architecture

VQ-VAE architectures The VQ-VAE-audio (respectively VQ-VAE-visual) architecture is symmetrical concerning the encoder and the decoder, with three 1D (respectively 2D) convolution for the encoder or transposed convolution for the decoder layers on the frequency axis and a residual convolution layer. The VQ-VAE models

process each frame independently with no time dependency. For each speech power spectrogram frame of size $D = 513$, the VQ-VAE-audio encoder compresses it into a discrete latent vector (a column of $\mathbf{x}_q^{(a)}$) of size $D' = 64$. For each image frame of size $(H = 96, W = 96, C = 3)$, the VQ-VAE-visual encoder compresses it into a discrete latent representation of size $(H' = 24, W' = 24)$. The VQ-VAE-audio and the VQ-VAE-visual codebooks contain, respectively, $k_a = 256$ and $k_v = 512$ codes of dimension $e_a = 8$ and $e_v = 4$. Such a low dimension is chosen to increase the use of the different codes in the codebook (Yu et al., n.d.). Please refer to the supplementary materials for more information about the VQ-VAE architectures.

VQ-MAE-AV architectures The VQ-MAE-AV model uses $L = 12$ attention blocks in the encoder and $L' = 4$ in the decoder. Each self-attention layer of a block is divided into *four* heads. By default, the parameters of the discrete audio and visual tokens (d, h, and w) are set to 4. We set $t_a = 10$ and $t_v = 5$ because the sampling rate of $\mathbf{x}^{(a)}$ is twice the sampling rate of $\mathbf{x}^{(v)}$. In the ablation study (Section 5.3.3), we will explore all possible combinations of the VQ-MAE-AV encoder and decoder, according to the two fusion strategies (*self-attention fusion* and *cross-attention fusion*). We will also evaluate the impact of the pooling strategy and contrastive learning. Additional ablation studies are presented in the supplementary material.

Training and fine-tuning details

Self-supervised training details The VQ-MAE-AV is trained using the AdamW optimizer (Loshchilov & Hutter, 2017) with a cosine scheduler to adjust the learning rate, with a 100-epoch warm-up period. The parameters of the optimizer, similar to (He et al., 2022), are $\beta_2 = 0.9$, $\beta_2 = 0.95$, and `weight_decay`$= 0.05$. The base learning rate follows the linear scaling rule (Goyal et al., 2017) $lr = (\texttt{base\_lr} = 1e-3) \times (\texttt{batchsize} = 128)/256$. We distributed the pre-training of VQ-MAE-AV on 4 NVIDIA HGX A100. Over the course of 140 epochs, with each epoch having a duration of approximately 15 minutes.

Fine-tuning details For the fine-tuning process, we also use the AdamW optimizer (Loshchilov & Hutter, 2017) with a cosine scheduler to adjust the learning rate and

with a 40-epoch warm-up period. The parameters of the optimizer are the same as those used for the pre-training. The base learning rate is `1e-4`.

Table 5.2 – Performance of VQ-MAE-AV using the *self-attention fusion* strategy for both encoder and decoder, and it is fine-tuned using the attention pooling strategy. 'Pre-train' refers to the training of the VQ-MAE-AV for the unmasking task on the VoxCeleb2 database. 'Freeze' refers to the freezing of the VQ-MAE-AV encoder.

Method	Pre-train	Freeze	Accuracy (%)
VQ-MAE-AV	✗	✗	29.6
VQ-MAE-AV	✓	✓	70.5
VQ-MAE-AV	✓	✗	81.5

Table 5.3 – Performance of VQ-MAE-AV using the *self-attention fusion* strategy for both encoder and decoder, and it is fine-tuned using the Query2Emo pooling strategy. 'Generative' corresponds to the loss function in Eq. 5.3 and 'Contrastive' corresponds to the loss function in Eq. 5.4.

Method	Contrastive	Generative	Accuracy (%)
VQ-MAE-AV	✓	✗	75.2
VQ-MAE-AV	✗	✓	84.3
VQ-MAE-AV	✓	✓	84.8

Table 5.4 – Performance of VQ-MAE-AV without contrastive learning for different encoder and decoder architectures, and it is fine-tuned using the attention pooling strategy. *SAF* stands for *self-attention fusion* and *CAF* stands for *cross-attention fusion*.

Method	Param. (M)	Encoder	Decoder	Acc. (%)
VQ-MAE-AV	13.5	*SAF*	*SAF*	81.5
VQ-MAE-AV	25.0	*CAF*	*SAF*	82.8
VQ-MAE-AV	18.4	*SAF*	*CAF*	81.8
VQ-MAE-AV	30.0	*CAF*	*CAF*	83.0

Table 5.5 – Performance of VQ-MAE-AV using the *self-attention fusion* strategy for both encoder and decoder and without contrastive learning for different pooling strategies on emotion recognition.

Pooling strategy	Accuracy (%)	F1-score (%)
Mean Pooling	78.1	78.4
Attention Pooling	81.5	80.1
Query2Emo	84.3	84.8

Table 5.1 – Accuracy (%) and F1-score (%) results of audiovisual SER. VQ-MAE-AV is pre-trained with both the generative (5.3) and contrastive (5.4) loss functions using the *cross-attention fusion* strategy for both the encoder and decoder and it is fine-tuned using the Query2Emo pooling strategy. In the modality column, A and V stand for audio and visual, respectively.

RAVDESS			
Method	Modality	Acc	F1
VO-LSTM (Ghaleb et al., 2019)	V	60.50	-
MAE-DFER (Sun et al., 2023)	V	75.56	-
AV-LSTM (Ghaleb et al., 2019)	A+V	65.80	-
MuLT (Tsai et al., 2019)	A+V	76.60	77.30
MDVAE (Sadok, Leglaive, Girin, et al., 2023b)	A+V	79.30	80.70
AVT (Chumachenko et al., 2022)	A+V	79.20	78.20
Ours	A	73.20	72.80
Ours	V	74.10	73.90
Ours	A+V	**84.80**	**84.50**
CREMA-D			
Method	Modality	Acc	F1
VO-LSTM (Ghaleb et al., 2019)	V	66.80	-
MAE-DFER (Sun et al., 2023)	V	77.38	-
AV-LSTM (Ghaleb et al., 2019)	A+V	72.90	-
MATER (Ghaleb et al., 2020)	A+V	67.20	-
AV-Gating (Ghaleb et al., 2019)	A+V	74.00	-
RAVER (Goncalves & Busso, 2022)	A+V	77.30	-
Ours	A	72.10	71.60
Ours	V	76.50	76.70
Ours	A+V	**80.40**	**80.00**
eNTERFACE05			
Method	Modality	Acc	F1
VO-LSTM (Ghaleb et al., 2019)	V	66.80	-
MAE-DFER (Sun et al., 2023)	V	77.38	-
AV-LSTM (Ghaleb et al., 2019)	A+V	72.90	-
MATER (Ghaleb et al., 2020)	A+V	67.20	-
AV-Gating (Ghaleb et al., 2019)	A+V	74.00	-
RAVER (Goncalves & Busso, 2022)	A+V	77.30	-
Ours	A	72.10	71.60
Ours	V	76.50	76.70
Ours	A+V	**80.40**	**80.00**

5.3.2 Audiovisual speech emotion recognition

Table 5.1 compares the emotion recognition performance (accuracy and F1-score metrics) of the proposed VQ-MAE-AV model (using the *cross-attention fusion* strategy for both the encoder and decoder) with the performance of several state-of-the-art methods. For instance, the audiovisual transformer (AVT) (Chumachenko et al., 2022) is a supervised method that employs self-attention fusion and modality dropout for audiovisual SER; Robust Audiovisual Emotion Recognition (RAVER) (Goncalves & Busso, 2022) is a supervised approach designed to address challenges in modality alignment, temporal information capture, and missing features handling; MAE dynamic facial expression recognition (MAE-DFER) (Sun et al., 2023) is a self-supervised technique that employs a local-global interaction Transformer as the encoder pretrained on VoxCeleb2; multimodal dynamical VAE (MDVAE) (Sadok, Leglaive, Girin, et al., 2023b) is an unsupervised audiovisual speech representation learning technique, using a hierarchical latent space that separates static from dynamical information and modality-common from modality-specific information.

The VQ-MAE-AV model with Query2Emo outperforms the state-of-the-art methods. VQ-MAE-AV achieves 8.2%, 5.6%, and 5.5% better accuracy than the MULT, AVT, and MDVAE methods for the RAVDESS dataset, respectively. Regarding the CREMA-D dataset, VQ-MAE-AV achieves 13.2%, 7.5%, and 3.1% better accuracy than the MATER, AV-Gating, and RAVER method. On the eNTERFACE05 dataset, VQ-MAE-AV achieves 8.2% and 5.5% better accuracy than the FAN and Graph-Tran methods. Overall, the VQ-MAE-AV model consistently outperforms state-of-the-art methods across all datasets, which demonstrates the effectiveness of the proposed audiovisual speech self-supervised representation learning technique for SER.

Table 5.1 also compares VQ-MAE-AV with its unimodal versions VQ-MAE-V and VQ-MAE-A. The VQ-MAE-AV model exhibits improved performance compared to VQ-MAE-V and VQ-MAE-A, with gains of 3.9% and 8.3% in accuracy on the CREMA-D dataset, respectively. This shows the effectiveness of incorporating audiovisual information in self-supervised learning for improved SER performance.

5.3.3 Ablation study and model properties

We conduct a series of experiments to evaluate the impact of various hyperparameters and model designs on the emotion recognition performance of the proposed VQ-MAE-AV

model. In the following paragraphs, we will present and discuss the findings of these experiments. All the ablation study is done on RAVDESS.

Impact of pre-training and fine-tuning

Table 5.2 shows the significance of pre-training and fine-tuning the VQ-MAE-AV model for audiovisual SER. Pre-training the model for the unmasking task on the VoxCeleb2 database substantially improves the emotion recognition performance, with the accuracy rising from 29.6% to 81.5%. Fine-tuning the encoder is also essential, as keeping it frozen leads to a 11% drop in accuracy.

Impact of the contrastive learning

The influence of contrastive learning on emotion recognition is illustrated in Table 5.3. When training VQ-MAE-AV exclusively with the contrastive loss, the accuracy achieved is 75.2%, which is 9.1% lower than when using only the generative loss. Notably, the fusion of both losses, contrastive and generative, leads to improved emotion recognition, resulting in a 0.5% accuracy boost compared to using the generative loss alone. This underscores the importance of integrating both contrastive and generative learning strategies.

Impact of the encoder/decoder architecture

Table 5.4 shows the performance of the four different configurations of the VQ-MAE-AV model described in Section 5.2.5, including all possible combinations of the fusion strategies for the VQ-MAE-AV encoder and decoder:

— *Self-attention fusion* for both the encoder and decoder (SAF-SAF);
— *Self-attention fusion* for the encoder and *cross-attention fusion* for the decoder (SAF-CAF);
— *Cross-attention fusion* for the encoder and *Self-attention fusion* for the decoder (CAF-SAF);
— and *Cross-attention fusion* for both the encoder and decoder (CAF-CAF).

Among these configurations, CAF-CAF achieves the highest accuracy with a 1.5% improvement over SAF-SAF, followed by CAF-SAF with a 1.3% improvement over SAF-SAF, and then SAF-CAF with only a 0.3% improvement. The *cross-attention fusion* encoder architecture achieves the best performance in emotion recognition, as shown by the CAF-CAF and CAF-SAF configurations. However, there exists a trade-off between performance

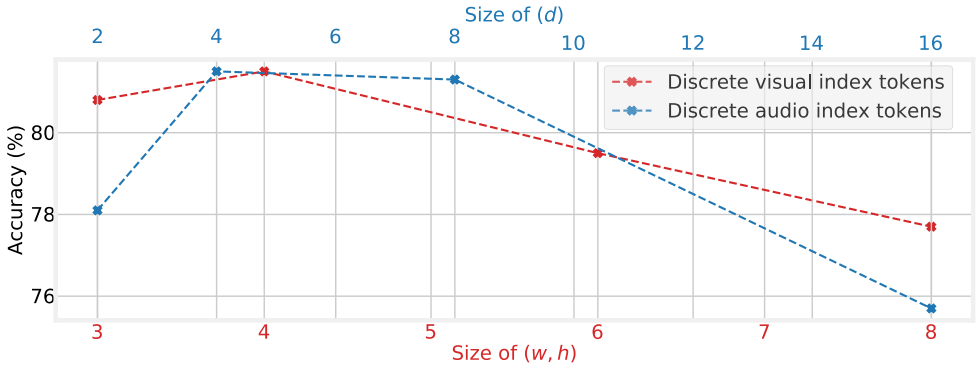


Figure 5.4 – Impact of the discrete audio and visual token size on emotion recognition.

and the number of model parameters. Notably, CAF-CAF involves slightly more than twice as many parameters as SAF-SAF.

Impact of the pooling strategy

Table 5.5 presents the impact of various pooling strategies when fine-tuning the model for SER. The results reveal that attention-based pooling techniques, such as attention pooling and Query2Emo, outperform mean pooling. Among the two attention-based methods, Query2Emo outperforms attention pooling, with an accuracy gain of 2.8%.

Impact of the audio and visual discrete token size

Figure 5.4 shows the impact of the dimensions of the discrete tokens for the visual (h and w) and audio (d) modalities on the SER performance. The values of h and w represent the visual token size on the horizontal and vertical axes, respectively, while d represents the token size on the frequency axis of the audio modality. Our study reveals that the performance of emotion recognition is impacted by both (h, w) and d values. Therefore, it is important to select these parameters carefully. Based on our experiments, we recommend fixing them to $(h = w = 4,\ d = 4)$.

Other ablation studies

We conducted additional ablation studies to investigate the impacts of the encoder depth and of the masking strategy. Detailed results and insights from these studies can be found in Appendix C.2.

5.4 Conclusion of the chapter

Masked autoencoder modeling is a versatile self-supervised learning approach that can be adapted to various types of data. This chapter introduced the VQ-MAE-AV model for learning representations of audiovisual speech data, which could be extended to other multimodal sequential data. VQ-MAE-AV took as input a discrete audio representation and a discrete visual representation obtained via two separate VQ-VAEs. These reduced representations were then divided into multiple discrete tokens, with spatio-temporal tokens for the visual modality and spectro-temporal tokens for the audio modality. Pre-trained on the VoxCeleb2 dataset and fine-tuned on the standard expressive audiovisual speech dataset, the experiments showed that the VQ-MAE-AV model effectively combined the audio and visual modalities for SER, outperforming several state-of-the-art methods across multiple datasets. For future work, we plan to investigate other masking strategies and increase the resolution of the images to further improve emotion recognition performance.

Chapter 6

Conclusion

In Chapter 1, we highlighted the challenges encountered by supervised models in the domain of ER and affective computing in general. Specifically, we pointed out the detrimental effects of biases in training data and supervised methods, ultimately hampering robustness and the ability to generalize effectively.

Given the exponential growth of non-annotated audiovisual data, there arises a need to explore alternative training paradigms that rely minimally or not on labeled data. This becomes particularly pertinent due to emotionally annotated data's inherent ambiguities and limitations (Section 1.3). These limitations encompass data volume constraints, recording hours and participant numbers, and the challenge of capturing intricate emotional nuances.

In Chapter 1, specifically in Section 1.1.2, we highlighted the essential criteria for an effective ER system, which encompass multimodality, robustness and accuracy, generality, sensitivity to dynamics, and contextual awareness. The following section will analyze the thesis' main points to determine how much our contributions align (or almost) with these criteria.

6.1 Main thesis points: exploration and analysis

Unraveling disentangled representations Disentangled representation is a focal point in my research trajectory. Chapter 3 introduces a straightforward approach to acquiring and manipulating speech variation factors (such as pitch and formant) within a VAE method. Moving forward to Chapter 4, we present another approach also based on VAE, where disentanglement is achieved across modalities (specific or shared) as well as temporally (static or dynamic).

Throughout these studies, we have demonstrated the significant advantages of disentangled representations, especially within downstream tasks such as ER. These advantages span several key areas, including enhanced sample efficiency, improved robustness, and

generalization capabilities.

Shifting from supervised to unsupervised and self-supervised approaches This transition aligns with a broader shift in artificial intelligence, where the pursuit of increased adaptability and versatility has conducted researchers toward alternative paradigms, diverging from the confines of supervised learning. These approaches can potentially harness the immense volume of unlabeled data available on the internet, a resource-rich in information but often needing more precise annotations. This shift is driven by the fact that traditional supervised models may falter when encountering new emotions, leaving them vulnerable to overfitting. The exploration of weakly supervised and unsupervised techniques represents a strategic response to this challenge, intending to fortify the resilience and scalability of ER models.

Through our contributions, we want to show that this transition to less supervised approaches can benefit ER task, making ER systems more robust and generable. While the shift from supervised to unsupervised methods may not yet be a groundbreaking revolution in ER, it represents a strategic change with significant potential. These non-supervised methods aim to establish a strong foundation by learning intricate data representations initially. These acquired representations become valuable for subsequent tasks, moving from solving isolated issues to fostering a more universally applicable framework.

Dynamical and multimodal input synergy This thesis aims to comprehensively understand human emotions by developing both temporal and multimodal (audiovisual) representations. Emotions are intricate, dynamic phenomena that evolve over time. Moreover, emotions are frequently conveyed through various sensory inputs. Visual data, for instance, can capture essential aspects of emotional expression, such as facial expressions, body language, and gestures. On the other hand, audio data can offer valuable insights into elements like tone of voice and speech patterns. Chapters 4 and 5 show that combining these modalities enriches the available information, leading to a more comprehensive understanding of emotions than the unimodal/static method.

6.2 Ethical considerations and perspectives

A typical doctoral thesis spans an average duration of three years, encompassing the processes of conceptualization, implementation, analysis, and interpretation. However,

this rigorous journey is also a period during which we occasionally step back from our research. Particularly in the context of this thesis, I acknowledge that the subject matter can evoke apprehension among those outside the field. As profoundly personal experiences, emotions raise valid concerns regarding privacy violations, emotional manipulation, and biases among users.

6.2.1 Ethical concerns with emotion recognition

As ER technology advances, it raises ethical concerns (Denning & Denning, 2020; Latif et al., 2022). Its rapid proliferation has sparked discussions about potential risks and challenges across diverse domains. These include the fear of exploitative manipulation, where access to intimate emotions could be misused to influence beliefs and behaviors, particularly in areas like politics, markets, and social interactions. Furthermore, the technology's often-invasive nature, with little regard for consent and privacy, raises significant concerns about personal autonomy and privacy rights. Additionally, the vulnerability of ER systems to adversarial attacks introduces risks, particularly in critical applications like mental health diagnosis. Issues of bias and fairness are also at the forefront, given the potential for biased training data and inaccurate emotion labeling to perpetuate unfair outcomes. Lastly, the reliance on reductionist emotional models can oversimplify the complexity of human emotions, potentially leading to misinterpretations and inadequate support in various contexts.

Using ER for emotional surveillance raises important ethical questions. To ensure the ethical treatment of individuals and prevent exploitation or manipulation, we can draw inspiration from well-established ethical principles, such as those outlined in the Belmont Report 1980 (on Aging et al., 1980). The Belmont Report identifies three key principles: *Respect for Persons, Beneficence, and Justice*. These principles serve as a foundation for guiding ethical conduct in research involving human subjects (Sims, 2010).

In the context of AI and ER, there has been a growing focus on ethics, resulting in numerous AI ethics principles and codes. Notably, Jobin et al., 2019 identified 84 such codes related to AI in 2019. Four high-level ethical principles commonly found in these declarations are beneficence, non-maleficence, autonomy, and justice. Additionally, Floridi and Cowls, 2022; Floridi et al., 2021 emphasized the principle of explicability, advocating that AI models should not operate as inscrutable black boxes.

6.2.2 Future directions

In terms of methodology and paradigm

Explainability and interpretability The vulnerability of ER systems largely stems from the inherent black-box nature of the machine learning models. These models often lack transparency and the ability to provide interpretable insights into their outcomes. It is vital to render them explainable in a manner comprehensible to humans. In recent years, extensive research efforts have been dedicated to developing methods for interpreting the workings of ML models, including methods explanations that delve into model internals to reveal the most effective features contributing to a specific prediction (Abnar & Zuidema, 2020; Chefer et al., 2021; Zeiler & Fergus, 2014).

Our thesis emphasized the acquisition of disentangled representations aiming for human-comprehensible models. Specifically in ER, this approach unveils the factors influencing emotion prediction transparently and interpretably. By doing so, it fosters trust and user acceptance. Given the numerous advantages of disentangled representations, the keyword "disentangled representation" will receive primary attention in future research efforts.

Integrating multidisciplinary insights ER systems can significantly benefit from an interdisciplinary approach incorporating psychology, neuroscience, and computer science expertise. This inclusive perspective provides a holistic understanding of the complex phenomenon of human emotions. Psychologists and neuroscientists get valuable insights into the underlying mechanisms governing emotional experiences, expressions, and regulation in individuals.

Sen YAN's doctoral thesis, "Personalizing facial expressions by exploring emotional mental prototypes", showcases the power of interdisciplinary collaboration in advancing ER. The thesis focuses on the idea of a unique mental prototype in each person. Sen YAN combines psychology and artificial intelligence in a novel way. The key innovation is blending generative adversarial networks with the psychophysical reverse correlation process. This method involves adjusting facial expressions by controlling specific action units in a GAN. Through iterative refinement guided by human feedback, the system identifies the most important action units linked to a particular emotion for an individual. This process customizes emotional prototypes beyond generic models to offer personalized insights into emotional expressions (Yan, 2023).

This pipeline can also be used for audio speech by incorporating the model from

Chapter 3. This adapted model allows us to control pitch and formant in the pipeline, giving us a more precise way to adjust audio speech attributes. By manipulating these factors and others, we can create personalized interpretations of speech-related emotions that match an individual's unique expression. Additionally, this pipeline has potential in multimodal ER. Combining insights from visual and audio modalities allows us to develop emotional prototypes tailored to each person. This approach can help us better understand how individuals express and perceive emotions across different modalities.

Context matters Context provides invaluable information that facilitates the interpretation of emotions. For example, when someone smiles while receiving a gift, the context of the situation (i.e., a gift-giving occasion) helps us deduce that the expressed emotion is likely happiness or gratitude. Contextual clues serve as a guide for ascribing meaning to emotional expressions. Including additional context through other modalities, such as text (scene or environmental descriptions), can enhance ER.

Recently, a growing interest has been in enhancing social reasoning abilities (J. Li et al., 2022a; B. Xie & Park, 2023). B. Xie and Park, 2023 explore the integration of multimodal inputs and domain expertise to advance the comprehension of common knowledge. They introduce a question-answering model conditioned on contextual embeddings from multimodal inputs, showing promising results in social intelligence learning.

Improving and evolving our ongoing research Technical limitations are inherent in our research, often serving as catalysts for improvement. However, one persisting limitation throughout this thesis has been the challenge of synthesizing high-quality audio waveforms. This thesis primarily represents audio speech using power spectrograms, omitting phase information, which is crucial for waveform reconstruction. Techniques like Griffin-Lim algorithm (Griffin & Lim, 1984), commonly used for phase reconstruction, have limitations and may introduce artifacts. However, high-quality speech synthesis is rapidly evolving, offering promising avenues for ongoing research. We are contemplating replacing the source-filter VAE, VQ-MDVAE, or VQ-MAE-AV decoder with a higher-quality decoder to enhance waveform speech synthesis (using neural vocoders like Hifi-GAN (J. Kong et al., 2020)). This upgrade enables tasks like speaker conversion or speech emotion conversion.

An essential aspect to explore is the adaptability of our methods across diverse tasks. While our unsupervised and self-supervised models were primarily developed for effective representation learning in emotion recognition, their application extends to various other

domains. One such domain is speech and speaker recognition, where MDVAE, for instance, could be used to enrich speaker data. By altering factors like head position, eye movements, and emotions while preserving mouth movements (content), these data augmentations could substantially enhance the performance of speech and speaker recognition tasks.

In terms of validation

Benchmarking and Comparison In validating our models, it is essential to establish a universally accepted protocol and a set of validation metrics to ensure fair benchmarking. Researchers and developers can use standardized validation frameworks to assess their ER systems against standard criteria. This practice not only enables equitable comparisons between different systems but also fosters healthy competition, ultimately stimulating innovation in the field. A lack of a universally accepted benchmarking protocol hinders the fair evaluation of ER systems. Establishing such standards is an ongoing challenge for the field.

Cross-corpus validation Different ER datasets often vary significantly regarding data collection conditions, participants' demographics, cultural backgrounds, and emotional expressions. Models need to generalize across this variability. While significant strides have been made in cross-corpus validation, the field still faces challenges in developing methods that can robustly handle diverse datasets (Milner et al., 2019; B. Schuller et al., 2010). Additionally, ongoing efforts are required to ensure that models are designed with cross-corpus generalization in mind.

In terms of emotion/attitude representation

In Section 1.3, we explored the two most popular ways of representing emotions: the categorical and the continuous modelization based on arousal and valence. It is important to mention that there are alternative ways of representing emotions, such as *Plutchik's wheel* of Emotions. Much like colors, Plutchik's model represents emotions as a wheel, with eight primary emotions arranged in pairs of opposites. It also considers the intensity and combinations of these primary emotions. Moine and Obin, 2020 proposed an alternative representation focusing on social attitudes, categorized into four attributes: friendly, seductive, dominant, or distant. These social attitudes are distinct from emotions, which encapsulate the internal state of a speaker, and propositional attitudes, which reflect a speaker's disposition toward an utterance.

In our research, we focused on the recognition of categorical emotions (Chapters 4 and 5) or their discrete intensity (Chapter 4). We did not explore continuous emotion representation. However, since our goal is to acquire audiovisual representations, we could readily swap out the classifier for a regression model in the output of our unsupervised/self-supervised models (VQ-MAE-AV, VQ-MDVAE) to forecast either the arousal or valence value.

In terms of environmental issues

There is no doubt that the use of large-scale models and large data sets can dramatically improve a model's ability to learn high-quality representations. Currently, there is a race to see who can create the largest model (for example, large language models) with the largest dataset. However, it is imperative to consider the carbon footprint associated with these massive learning processes and their environmental consequences.

In 2019, researchers at the University of Massachusetts Amherst made a significant discovery when they trained several large language models (Strubell et al., 2019). Their investigation revealed that training a single large AI model could produce a staggering ≈ 300 tons of carbon emissions. To put this into perspective, these emissions are equivalent to what five cars would emit over their lifetimes. A more recent study focused on training GPT-3, a language model with a massive 175 billion parameters. The findings showed that this training process consumed 1287 MWh of electricity. Furthermore, it led to the release of 502 metric tons of carbon emissions, a figure that can be likened to the carbon footprint generated by 112 gasoline-powered cars operating for a year (Tamburrini, 2022).

Addressing environmental issues related to AI methods requires a multi-faceted approach. In addition to dedicated hardware resources for AI architecture and efficient model optimization techniques, several other strategies can contribute to environmental sustainability. These include employing quantization and model pruning to reduce computational requirements (Liang et al., 2021), implementing federated learning to reduce data transfer energy consumption (Z. Yang et al., 2020), applying knowledge distillation to transfer insights from large models to smaller and more energy-efficient ones (Gou et al., 2021), and investing in green data centers powered by renewable energy sources (Bird et al., 2014). Additionally, exploring sparse models that train only essential neural network connections and developing dynamic resource allocation systems to prevent energy waste are essential. Research into energy-efficient hardware architectures designed specifically for AI tasks and incentives for prioritizing eco-friendly AI technologies are also essential components

(Tamburrini, 2022). Furthermore, considering carbon offset programs and advocating for policies and regulations that promote environmentally conscious AI development practices can collectively contribute to reducing the carbon footprint of AI methods (Wara & Victor, 2008). Collaboration across hardware, software, and policy domains is key to addressing these environmental challenges effectively.

Résumé en Français

Introduction à la thèse

Imaginez-vous lors d'un entretien d'embauche ou d'une session de recrutement vidéo. Vous êtes confronté à une question difficile qui provoque en vous des réactions physiologiques et émotionnelles telles que la transpiration, l'accélération du rythme cardiaque, une tension artérielle élevée, des bouffées de chaleur, ainsi qu'une gamme d'émotions allant du stress à l'anxiété, voire à la colère. Ces réactions, bien qu'elles soient normales, peuvent entraver votre capacité à communiquer efficacement et à démontrer vos compétences.

En de telles situations stressantes, imaginez avoir accès à un outil en cours de développement en collaboration avec Randstad et CentraleSupélec, conçu pour atténuer ces réponses émotionnelles et physiologiques. Cet outil agirait comme un coach numérique, proposant des interactions adaptées à votre état émotionnel, au contexte et à l'environnement, vous permettant de regagner confiance en vous et d'améliorer vos performances lors de l'entretien.

Contexte et problématique Pour assurer l'efficacité de cet outil, une *reconnaissance précise des émotions* est essentielle. C'est précisément l'application de ma thèse, qui s'inscrit dans le domaine de l'informatique affective. Cette discipline englobe le développement de technologies capables de reconnaître, interpréter et réagir aux émotions humaines (R. Picard, 1997). L'informatique affective trouve des applications variées, de la santé (Mano et al., 2016) à l'éducation (C.-H. Wu et al., 2016) en passant par le divertissement et le marketing. La reconnaissance précise des émotions revêt une importance capitale pour que les machines interagissent harmonieusement avec nos émotions. Pour être efficace, un système de reconnaissance des émotions doit être *multimodal, robuste, général, sensible aux dynamiques et conscient du contexte* (Pantic et al., 2005).

La pipeline d'un système de reconnaissance des émotions en utilisant une approche

supervisée est généralement structurée de la manière suivante : Tout d'abord, il faut constituer une base de données d'apprentissage comportant des enregistrements, tels que des données audio et/ou visuelles, étiquetés avec les émotions correspondantes. Ensuite, ces données sont prétraitées pour extraire des caractéristiques pertinentes, ce qui implique souvent une analyse spectrale pour les signaux audio et une analyse des images pour les signaux visuels. Les caractéristiques extraites sont ensuite utilisées pour entraîner un modèle d'apprentissage automatique, tel qu'un réseau de neurones, avec les étiquettes d'émotion comme cibles. Ce modèle est entraîné à reconnaître les schémas de caractéristiques associés à chaque émotion. Une fois le modèle entraîné, il peut être évalué sur des données de test pour mesurer ses performances en terme de précision de la reconnaissance émotionnelle.

Dans les systèmes de reconnaissance des émotions, les émotions sont souvent représentées par des étiquettes catégorielles ou des échelles numériques, mais ces représentations ne parviennent pas à saisir pleinement l'ambiguïté émotionnelle (Tran et al., 2022). Cette ambiguïté peut introduire des biais dans les données et par conséquent sur le modèle qui est entrainé de manière supervisé.

Existe-t-il d'autres paradigmes d'apprentissage permettant d'acquérir des représentations pertinentes de manière non supervisée, en vue de leur application ultérieure dans des tâches auxiliaires telles que la reconnaissance des émotions ?

Solutions et défis envisagés

Exploitation des données non étiquetées pour la reconnaissance des émotions L'utilisation de données non étiquetées pour la reconnaissance des émotions présente des avantages importants. Les méthodes d'apprentissage non supervisé et auto-supervisé permettent de réduire la dépendance aux données étiquetées, ce qui est crucial étant donné les limites et les biais potentiels des ensembles de données étiquetées. Ces méthodes impliquent généralement un processus d'apprentissage en deux étapes, où le modèle apprend d'abord à représenter les données de manière non supervisée ou auto-supervée sur une grande base de donnée non annotée, puis transfère ces connaissances à des tâches auxiliaires (e.g., reconnaissance des émotions).

Exploration de l'apprentissage multimodal pour la reconnaissance des émotions De plus, l'adoption de l'apprentissage multimodal pour l'affective computing est essentielle. Les émotions ne sont pas uniquement transmises par les mots, mais aussi par le ton de la voix, les expressions faciales et le langage corporel. En intégrant ces différentes modalités dans les systèmes de reconnaissance des émotions, on peut réduire l'incertitude et améliorer la précision (Abdullah et al., 2021; Sebe et al., 2005). Cela permet de capturer les émotions complexes qui se produisent dans des contextes réels, où les signaux émotionnels peuvent varier considérablement d'une personne à l'autre.

L'apprentissage non supervisé ou auto-supervisé des représentations multimodales et séquentielles présente-t-il des avantages pour la reconnaissance des émotions ?

Un apperçu rapide de l'état-de-l'art

Les modèles génératifs profonds, tels que l'autoencodeur variationnel (VAE), sont devenus récemment très efficaces pour l'apprentissage non supervisé de représentations latentes à partir de données complexes comme les images, l'audio et le texte (Goodfellow et al., 2014; Kingma & Welling, 2014; Rezende et al., 2014). Apprendre ces représentations est important, non seulement pour la synthèse des données, mais aussi pour leur analyse et leur transformation. Une représentation efficace doit saisir les caractéristiques clés des données tout en restant invariable face aux petites variations locales des données d'entrée, et elle doit être aussi démêlée que possible pour assurer l'explicabilité (Bengio, Courville, & Vincent, 2013; Van Steenkiste et al., 2019). Ces modèles génératifs ont grandement amélioré notre capacité à créer des représentations structurées et interprétables des données.

Le VAE permet l'apprentissage profond non supervisé dans un cadre bayésien. En général, une distribution gaussienne standard est choisie pour la distribution a priori sur la variable latente, favorisant l'indépendance (démêlement) entre les différentes dimensions de la représentation apprise. Cependant, les VAEs classiques présentent des limites en termes de démêlement, surtout avec des ensembles de données complexes. Pour améliorer cela, différentes approches ont été développées pour introduire des biais dans le modèle et/ou l'algorithme d'apprentissage, permettant de renforcer le démêlement. Ces méthodes incluent la modification de la borne infèrieur de la

vraissemblence (R. T. Chen et al., 2018; Higgins et al., 2017a; H. Kim & Mnih, 2018). De nouvelles approches se concentrent sur l'apprentissage faiblement supervisé (Locatello, Poole, et al., 2020; Sadok, Leglaive, Girin, et al., 2023a) ou semi-supervisé (Klys et al., 2018). Les VAE sont flexibles et ont été étendus à différentes formes de données, y compris le multimodal et le séquentiel.

Les VAEs ont suscité un intérêt significatif pour la modélisation de données multimodales. Ces VAEs, avec leurs encodeurs et décodeurs, sont stables à l'apprentissage contrairement aux réseaux génératifs antagonistes (GANs) (Goodfellow et al., 2014). Ils sont adaptés à la modélisation générative multimodale (Suzuki & Matsuo, 2022). Plusieurs approches ont été développées pour apprendre un espace latent commun pour plusieurs données d'entrée hétérogènes. Par exemple, PoE-VAE (M. Wu & Goodman, 2018) adopte le produit d'experts (PoEs) (Hinton, 2002) pour modéliser la distribution postérieure de données multimodales, tandis MoE-VAE (Shi et al., 2019) utilise un mélange d'experts (MoEs). Une autre approche (Sutter et al., 2021) combine ces deux méthodes pour améliorer la reconstruction des données. Néanmoins, des limites ont été démontrées et formalisées pour ces méthodes. Par exemple, les modèles VAE multimodaux produisent souvent des reconstructions de moindre qualité par rapport aux modèles VAEs unimodaux, en particulier pour les ensembles de données complexes.

Un autre domaine où les modèles VAEs ont connu des progrès significatifs est la modélisation de données séquentielles, où les variables latentes et/ou observées évoluent dans le temps. Les VAE dynamiques (DVAE) (Girin et al., 2021b) visent à traiter des données complexes de grande dimension présentant des corrélations temporelles ou spatiales à l'aide de réseaux bayésiens dynamiques profonds. Les réseaux neuronaux récurrents sont souvent utilisés à cette fin, et une large gamme de méthodes ont été développées, différant dans leur structure de modèle d'inférence et de modèle génératif. Ces modèles DVAE ont deux points en commun lors de la modélisation de données séquentielles : L'entrainement non supervisée est préservée, et la structure du VAE est maintenue ; cela signifie que les modèles d'inférence et génératifs sont appris conjointement en maximisant une borne inférieure du log de la probabilité marginale.

Le chapitre 2 (page 27) offre un aperçu approfondi de l'état de l'art concernant les modèles génératifs entraînés de manière non supervisée ou auto-supervisée. Il aborde l'apprentissage de représentations disentangled, multimodales, séquentielles, ainsi que d'autres aspects pertinents.

Contributions

Apprentissage et contrôle de la représentation avec un autoencodeur variationnel (Chapitre 3, page 75)

Comprendre et contrôler les représentations latentes dans les modèles génératifs profonds est un problème difficile mais important pour l'analyse, la transformation et la génération de divers types de données. Dans le traitement de la parole, s'inspirant des mécanismes anatomiques de la phonation, le modèle source-filtre (Fant, 1970) considère que les signaux de parole sont produits à partir de quelques facteurs latents continus indépendants et physiquement interprétables, parmi lesquels la fréquence fondamentale et les formants sont de première importance. Dans ce travail, nous montrons que le modèle source-filtre de la production de la parole apparaît naturellement dans l'espace latent d'un VAE (Kingma & Welling, 2014; Rezende et al., 2014) entraîné de manière non supervisée sur un ensemble de données de signaux de parole naturelle. En utilisant seulement quelques secondes de signaux étiquetés générés par un synthétiseur vocal artificiel, nous montrons expérimentalement que la fréquence fondamentale et les fréquences des formants sont encodées dans des sous-espaces orthogonaux de l'espace latent du VAE et nous développons une méthode faiblement supervisée pour contrôler de manière précise et indépendante ces facteurs de variation de la parole dans les sous- espaces latents appris. Sans nécessiter d'informations supplémentaires telles que du texte ou des données étiquetées manuellement, nous proposons un modèle génératif profond de spectrogrammes de parole qui est conditionné par la fréquence fondamentale et les fréquences des formants, et qui est appliqué à la transformation des signaux de parole.

Autoencodeur variationnel dynamique multimodal pour l'apprentissage de la représentation audiovisuelle de la parole (Chapitre 4, page 103)

Nous présentons un autoencoder variationelle multimodal et dynamique (MDVAE) appliqué à l'apprentissage non supervisé de la représentations de la parole audiovisuelle. L'espace latent est structuré pour dissocier les facteurs dynamiques latents partagés entre les modalités de ceux qui sont spécifiques à chaque modalité. Une variable latente

statique est également introduite pour encoder l'information qui reste constante au fil du temps au sein d'une séquence de parole audiovisuelle. Le modèle est entraîné de manière non supervisée sur un ensemble de données de parole émotionnelle audiovisuelle, en deux étapes: Dans la première étape, un autoencoder variationelle quantifié (VQ-VAE) (Van den Oord et al., 2017) est appris indépendamment pour chaque modalité, sans modèle temporel. La deuxième étape consiste à apprendre le modèle MDVAE sur la représentation intermédiaire des VQ-VAE avant la quantification. La dissociation entre l'information statique et dynamique, ainsi que entre l'information spécifique à chaque modalité et communes aux différentes modalités, se produit au cours de cette deuxième étape d'entraînement. Des expériences approfondies sont menées pour étudier comment les facteurs latents de la parole audiovisuelle sont encodés dans l'espace latent du MDVAE. Ces expériences comprennent la manipulation de la parole audiovisuelle, le débruitage audiovisuel d'images de visages et la reconnaissance des émotions dans la parole audiovisuelle. Les résultats montrent que le MDVAE combine efficacement l'information audio et visuelle dans son espace latent. Ils montrent également que la représentation statique apprise de la parole audiovisuelle peut être utilisée pour la reconnaissance des émotions avec peu de données annotées, et avec une meilleure précision par rapport aux modèles unimodaux de référence et à un modèle supervisé état de l'art.

Un autoencodeur masqué pour l'apprentissage de la représentation audiovisuelle de la parole (Chapitre 5, page 137)

Nous nous appuyons sur l'intérêt croissant pour les méthodes d'apprentissage auto-supervisé, qui offrent des solutions prometteuses aux limitations de l'apprentissage supervisé. Ces approches permettent d'apprendre à partir de vastes quantités de données non annotées, souvent disponibles dans divers domaines. Dans ce contexte, nous proposons le modèle VQ-MAE-AV, conçu spécifiquement pour l'apprentissage auto-supervisé de la représentation de la parole audiovisuelle. Contrairement aux MAE multimodaux existants qui traitent les données brutes de la parole audiovisuelle, le modèle VQ-MAE-AV adopte un paradigme d'auto-apprentissage basé sur des représentations discrètes de la parole audio et visuelle apprises par deux VQ-VAE pré-entraînés (Van den Oord et al., 2017). Pour évaluer l'efficacité de l'approche proposée, le modèle VQ-MAE-AV est pré-entraîné sur la base de données VoxCeleb2 (Chung et al., 2018) et

affiné sur des ensembles de données standard de parole audiovisuelle émotionnelle. Les résultats expérimentaux démontrent que la méthode proposée surpasse les méthodes actuelles de reconnaissance des émotions dans la parole audiovisuelle. Ces résultats soulignent le potentiel des approches d'apprentissage auto-supervisé et mettent en avant l'efficacité du modèle VQ-MAE-AV dans l'apprentissage de représentations robustes et efficaces de la parole audiovisuelle pour la reconnaissance des émotions.

Conclusion

Les principales conclusions de cette thèse peuvent être synthétisées en trois points :

L'apprentissage des représentations démêlées La représentation démêlée est un élément central de cette recherche. Dans le Chapitre 3, une méthode VAE est présentée pour apprendre et contrôler des facteurs de variation de la parole (comme le pitch et les formants). Le Chapitre 4 explore une autre approche basée VAE aussi, permettant le désentrelacement entre les modalités (spécifiques ou partagées) ainsi que temporellement (statique ou dynamique). Ces études montrent que les représentations démêlées appliquées à la reconnaissance des émotions (ER) offrent une meilleure efficacité, performance et généralisation.

Transition des approches supervisées vers les approches non supervisées et auto-supervisées Le passage à des approches non supervisées peut renforcer la reconnaissance des émotions, améliorant la robustesse et l'adaptabilité des systèmes ER. Bien que ce changement de méthodologie ne représente pas une révolution majeure dans le domaine, il constitue un changement stratégique avec un potentiel considérable. Ces méthodes non supervisées visent à établir des bases solides en apprenant des représentations complexes, passant de la résolution de problèmes spécifiques à la création d'un cadre plus universellement applicable.

Synergie dynamique et multimodale pour la reconnaissance des émotions Cette thèse explore les émotions humaines via des représentations temporelles et multimodales. Les émotions, phénomènes dynamiques, sont transmises par divers stimuli sensoriels (audio et visuelle). Les chapitres 4 et 5 démontrent que cette combinaison de modalités enrichit la compréhension émotionnelle.

Acronyms

AI Artificial Intelligence. 29

CNNs Convolutional Neural Networks. 16, 55

DNN Deep Neural Network. 28, 47, 48, 56, 60, 65

DNNs Deep Neural Networks. 33, 61, 205

DSAE Disentangled Sequential Autoencoders. 58, 206

DVAE Dynamical Variational Autoencoder. 56, 57, 59, 109

ELBO Evidence lower bound. 48–50, 53, 55, 57, 58, 61, 63, 64, 112

ER Emotion recognition. 28, 38, 159–164, 173

FA Factor Analysis. 46

GAN Generative Adversarial Network. 40, 42, 60, 113

i.i.d. independent and identically distributed. 42, 49

IB Information Bottleneck. 31, 32, 34, 36

KL Kullback-Leibler divergence. 42

MAE Masked Autoencoder. 70–74, 139, 140, 143, 154

MDVAE Multimodal Dynamical Variational Autoencoder. 104, 106–108, 113, 114, 116, 117, 123, 124, 131–133, 135, 208, 214

MEAD Multi-view emotional audiovisual dataset. 115, 117, 126, 127, 130, 132–134, 199, 209, 213, 214

MI Mutual information. 30, 32, 34, 53–55

MLP Multi-Layer Perceptron. 69, 131, 133, 134

MoE Mixture of Experts. 64

MSE Mean Squared Error. 73

MVAE Multimodal Variational Autoencoder. 60–62, 64, 65

PCA Principal Component Analysis. 41, 44–46

PoE Product of Experts. 63, 64

PPCA Probabilistic Variational Autoencoder. 45–47

ReLU Rectified Linear Unit. 69

RL Representation Learning. 28, 31, 51, 206

RNNs Recurrent Neural Networks. 16, 55–57, 59

SGD Stochastic gradient descent. 50, 51

SSL Self-supervised Learning. 32, 35, 36, 68–70, 138, 140, 205

STFT Short-Time Fourier Transform. 80, 83, 115, 149, 150, 177, 179

TC Total Correlation. 54, 55

VAE Variational Autoencoder. 40, 41, 47, 52, 57, 60, 61, 64, 66, 68, 76–84, 86, 90, 95, 98, 100, 101, 105, 106, 108, 113, 114, 159, 173

VQ-MAE-AV Vector quantized masked autoencoder for audiovisual representation. 7, 137, 138, 140–149, 151–155, 157, 209, 214

VQ-VAE Vector quantized VAE. 67, 113, 114, 116, 117, 127–129, 140–144, 146, 149–151, 157, 206, 208, 209

Appendix A

Appendix: Source-filter VAE

A.1 Experimental setup details

VAE training To train the IS-VAE model (Bando et al., 2018; Girin et al., 2019b; Leglaive et al., 2018), we use the Wall Street Journal (WSJ0) dataset (Garofalo et al., 1993), which contains 25 hours of speech signals sampled at 16 kHz, including 52 female and 49 male speakers. The time-domain speech signals are converted to power spectrograms using the STFT with a Hann analysis window of length 64 ms (1,024 samples) and an overlap of 75%. The VAE input/output dimension is $D = 513$ (we only keep the non-redundant part of the power spectrogram corresponding to positive frequencies) and the latent vector dimension is set to $L = 16$. The VAE encoder and decoder networks each have three dense layers. Their dimensions (input dimension, output dimension) are (513, 256), (256, 64) and (64, 2×16) for the encoder, and (16, 64), (64, 256) and (256, 513) for the decoder [1]. A hyperbolic tangent (tanh) activation function is used at each layer, except for the output layers of the encoder and decoder where we use the identity function. We train the model using the Adam optimizer (Kingma & Ba, 2015) with a learning rate equal to 0.001.

Artificially generated speech data For a given factor of variation, the corresponding latent subspace is learned using trajectories of speech power spectra generated with Soundgen (Anikin, 2019), all other factors being arbitrarily fixed (see Section 3.3.2). For f_0, the trajectory contains 226 points (which corresponds to 3.6 seconds of speech) evenly spaced in the range $[85, 310]$ Hz, f_1, f_2 and f_3 being fixed to 600 Hz, 2000 Hz, and 3000 Hz, respectively. For f_1, the trajectory contains 401 points (which corresponds to 6.4 seconds of speech) evenly spaced in the range $[200, 1000]$ Hz, f_0, f_2 and f_3 being fixed to 140 Hz, 1600 Hz, and 3200 Hz, respectively. For f_2, the trajectory contains 401 points evenly spaced in the range $[800, 2800]$ Hz, f_0, f_1 and f_3 being fixed to 140 Hz, 500 Hz,

1. We have two parameter vectors (mean and variance) for the distribution of $\mathbf{z}$ at the encoder, whereas we have only one single parameter vector (scale) for the distribution of $\mathbf{x}$ at the decoder, see 3.2 and 3.3.

and 3200 Hz, respectively. For f_3, the trajectory contains 241 points (which corresponds to 3.9 seconds of speech) evenly spaced in the range $[2000, 3200]$ Hz, f_0, f_1 and f_2 are fixed to 140 Hz, 500 Hz, and 1200 Hz, respectively. These four trajectories are illustrated in Figure A.1. The amplitude of the formants is fixed at 30dB, and their bandwidth is automatically calculated from the formant frequencies using a formula derived from phonetics studies. Quoting the documentation of Soundgen (Anikin, 2019), "above 500 Hz [the bandwidth] follows the original formula known as "TMF-1963" (Tappert et al., 1963), and below 500 Hz it applies a correction to allow for energy losses at low frequencies (Khodai-Joopari & Clermont, 2002). Below 250 Hz the bandwidth starts to decrease again, in a purely empirical attempt to achieve reasonable values even for formant frequencies below the ordinary human range. See the internal function soundgen:::getBandwidth()." The regression models used to control the speech factors of variation in the latent space (see Section 3.3.4) are learned on the same trajectories, but using the values of the Soundgen input parameters.

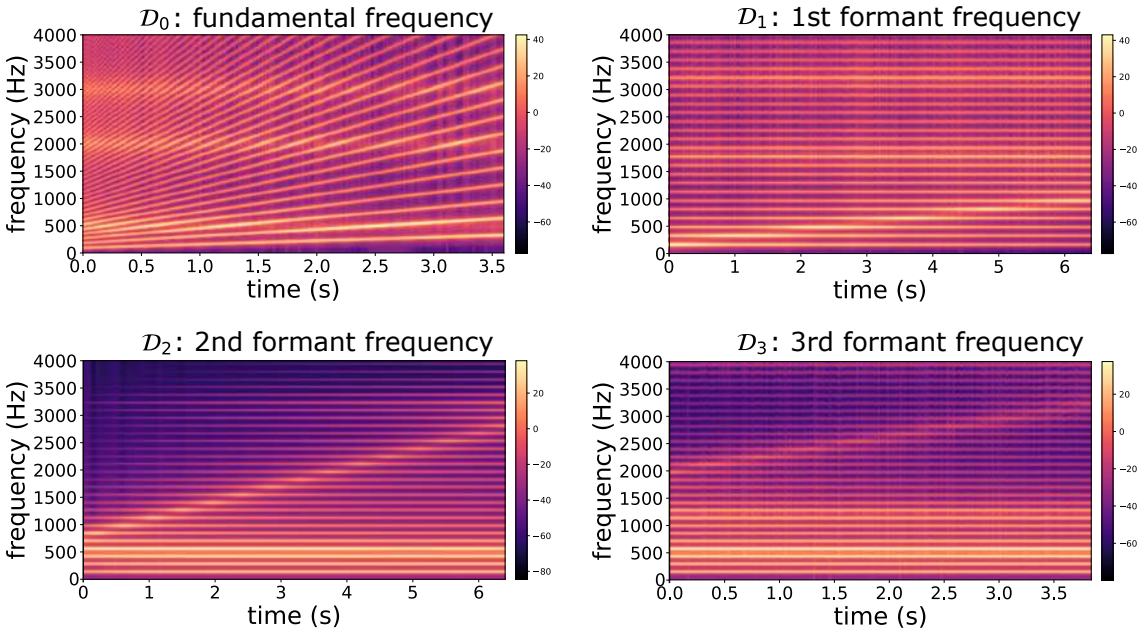
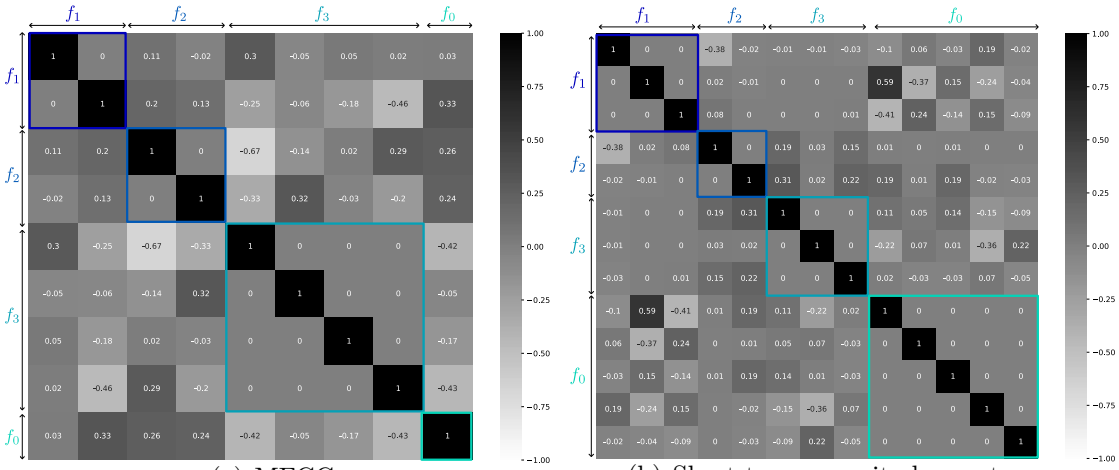


Figure A.1 – Trajectories of speech power spectra generated with Soundgen (Anikin, 2019), where only one factor of variation globally varies in each trajectory. From top to bottom and from left to right: the trajectory of the fundamental frequency f_0, the trajectory of the formant f_1, the trajectory of the formant f_2 and the trajectory of the formant f_3.

A.2 Correlation matrices obtained from MFCCs and short-term magnitude spectra



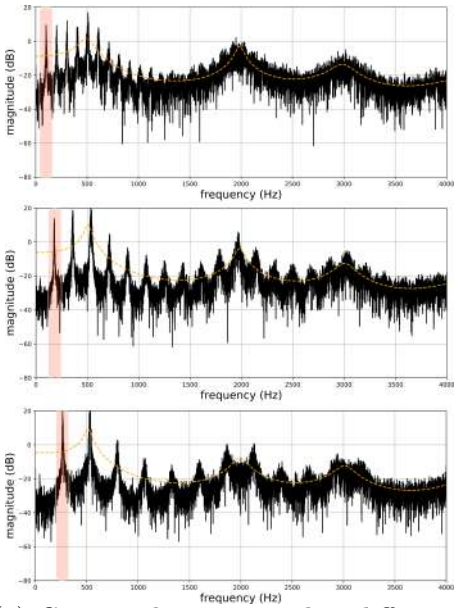
(a) MFCC (b) Short-term magnitude spectrum

Figure A.2 – Correlation matrix of the latent subspace basis vectors learned for MFCC (top) and short-term magnitude spectrum (bottom).

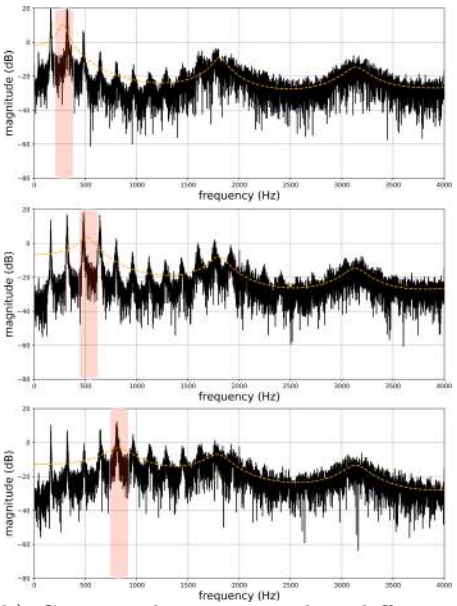
We conducted experiments with the proposed method (i.e., learning the subspace for each factor of variation and then learning a regression model to move in the learned subspace) on the following representations: MFCC and short-term magnitude spectrum (i.e., columns of the STFT magnitude spectrogram). We used the same artificial dataset as for the VAE latent space representation. Figures A.2(a) and A.2(b) below show the correlation matrix of the latent subspace basis vectors learned for the MFCC and short-term magnitude spectrum, respectively. Similarly to what we did with the VAE, the dimension of the subspaces is determined by applying a threshold on the data variance ($\geq$80%). For the dimension of the f_0 subspace with the MFCCs, we actually need 15 components to keep 80% of the data variance, so here we take only the principal one to facilitate the reading.

A.3 Additional qualitative results

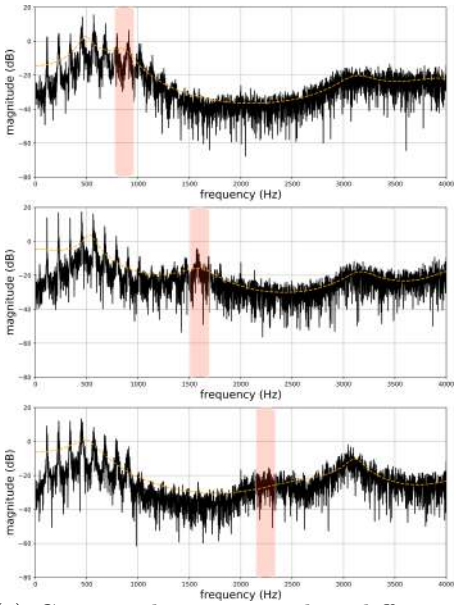
Examples of generated speech spectra



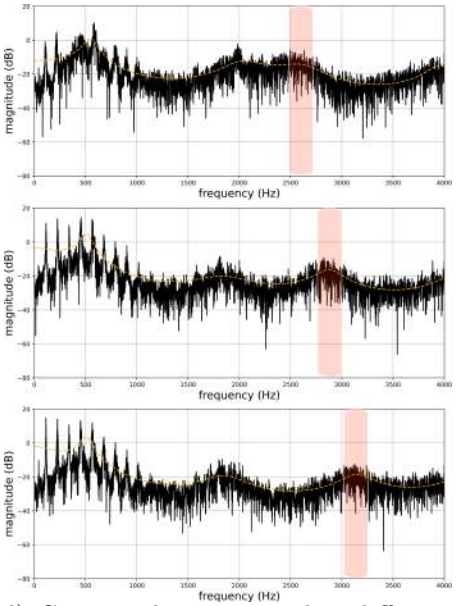
(a) Generated spectra with 3 different values of f_0.



(b) Generated spectra with 3 different values of f_1.



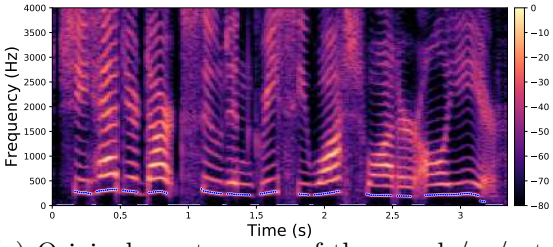
(c) Generated spectra with 3 different values of f_2.



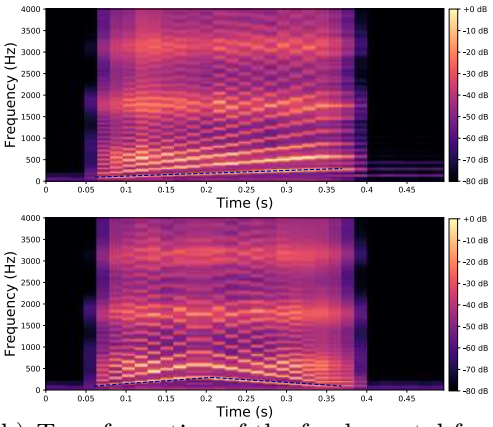
(d) Generated spectra with 3 different values of f_3.

Figure A.3 – Power spectra (solid black line) and spectral envelopes (dashed orange line) obtained using the conditional prior in 3.13 (generalized to conditioning on multiple factors). Each subfigure contains three plots where we vary the value of one single factor at a time: f_0 in (a), f_1 in (b), f_2 in (c), and f_3 in (d).

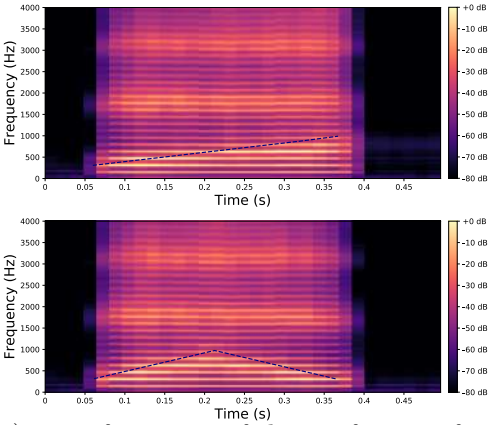
Examples of transformed speech spectrograms



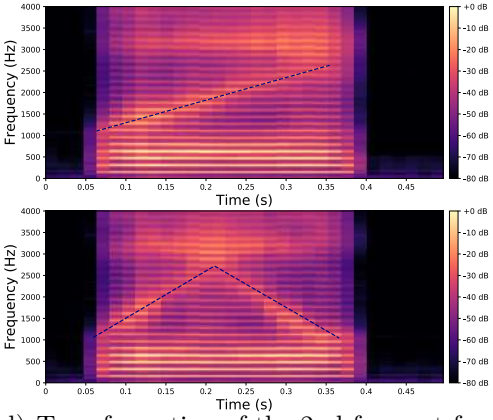
(a) Original spectrogram of the vowel /ae/ uttered by a male speaker.



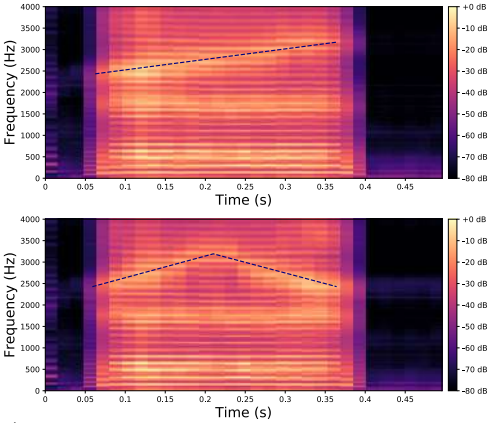
(b) Transformation of the fundamental frequency f_0.



(c) Transformation of the 1st formant frequency f_1.



(d) Transformation of the 2nd formant frequency f_2.



(e) Transformation of the 3rd formant frequency f_3.

Figure A.4 – Figure (a) shows the spectrogram of a vowel uttered by a male speaker. Figures (b), (c), (d) and (e) show transformations of this spectrogram with the proposed method, where we vary f_0, f_1, f_2, and f_3, respectively. The target value for these factors is indicated by the dashed blue line.

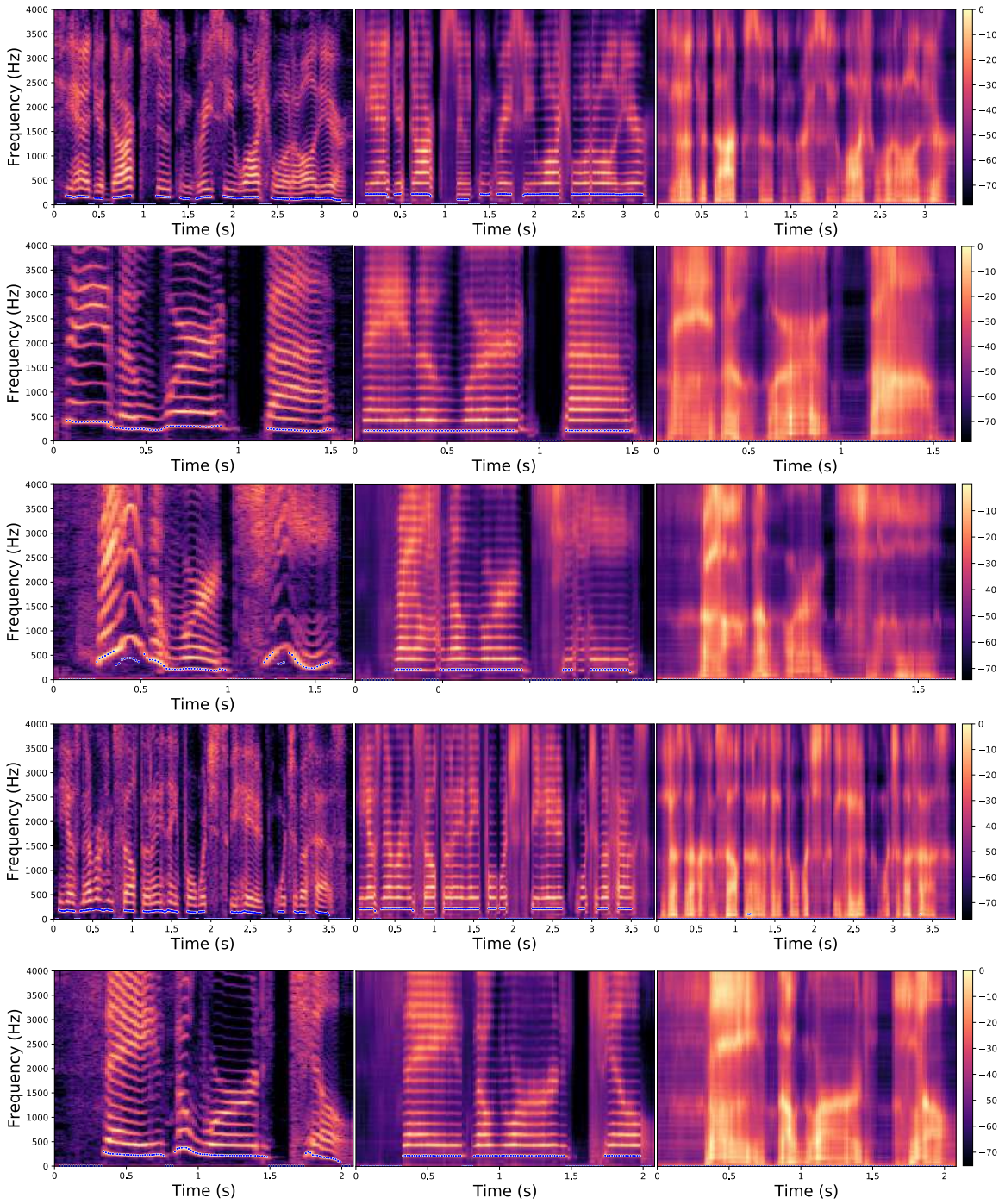


Figure A.5 – Each line in this figure corresponds to a speech signal uttered by a different speaker. Left: spectrogram of the original speech signal; Middle: transformed spectrogram where the fundamental frequency is set constant over time; Right: transformed spectrogram where the original voiced speech signal (left) is converted into a whispered speech signal (i.e., the fundamental frequency is removed).

Appendix B

Appendix: MDVAE

B.1 The detailed architecture of the vector quantized MDVAE

This section details the architecture of the VQ-MDVAE model, starting with the VQ-VAE and then the MDVAE.

Table B.1 – The architecture of the VQ-VAE-visual.

	Layer	Activation	Output dim
Input	-	-	3 × 64 × 64
Encoder	Conv2D(3, 64, 4, 2, 1)	ReLu	64 × 32 × 32
	Conv2D(64, 128, 4, 2, 1)	ReLu	128 × 16 × 16
	Conv2D(128, 128, 4, 2, 1)	ReLu	128 × 8 × 8
	2 × Residual Stack	ReLu	128 × 8 × 8
	Conv2D(128, 32, 1, 1)	-	32 × 8 × 8
Decoder	ConvT2D(32, 128, 1, 1)	-	128 × 8 × 8
	2 × Residual Stack (T)	ReLu	128 × 8 × 8
	ConvT2D(128, 64, 4, 2, 1)	ReLu	128 × 16 × 16
	ConvT2D(64, 64, 4, 2, 1)	ReLu	64 × 32 × 32
	ConvT2D(64, 3, 4, 2, 1)	-	3 × 64 × 64

Conv2D(in_channel, out_channel, kernel_size, stride, padding)
Residual Stack (T) = { 2 × Conv(T)2D(128, 128, 3, 1, 1)}

Table B.2 – The architecture of the VQ-VAE-audio.

	Layer	Activation	Output dim
Input	-	-	1 × 513
Encoder	Conv1D(1, 16, 4, 2, 1)	Tanh	16 × 256
	Conv1D(16, 32, 4, 2, 1)	Tanh	32 × 128
	Conv1D(32, 32, 3, 2, 1)	Tanh	32 × 64
	1 × Residual Stack	Tanh	32 × 64
	Conv1D(32, 8, 1, 1)	-	8 × 64
Decoder	ConvT1D(8, 32, 1, 1)	-	32 × 64
	1 × Residual Stack (T)	Tanh	32 × 64
	ConvT1D(32, 32, 3, 2, 1)	Tanh	32 × 128
	ConvT1D(32, 16, 4, 2, 1)	Tanh	16 × 256
	ConvT1D(16, 1, 4, 2, 0)	-	1 × 513

Conv1D(in_channel, out_channel, kernel_size, stride, padding)

Residual Stack (T) = { 2 × Conv(T)1D(32, 32, 3, 1, 1)}

VQ-VAE

The VQ-VAE developed for audio or images consists of three parts: (i) an encoder that maps an image to a sequence of continuous latent variables, referred to as the intermediate representation in the paper; (ii) a shared codebook that is used to quantize these continuous latent vectors to a set of discrete latent variables (each vector is replaced with the nearest vector from the codebook); and (iii) a decoder that maps the indices of the vectors from the codebook back to an image. The architectures of the visual and audio VQ-VAEs are described in tables B.1 and B.2, respectively.

MDVAE

MDVAE is decomposed into two models: (i) the inference model (encoder), further decomposed into four inferences for each latent variable, represented by Gaussian distributions whose parameters are determined via a neural network. The prior distributions for the dynamic latent variables are also trained, except for the static latent space, where the prior is assumed to be a standard normal distribution. (ii) The second part comprises

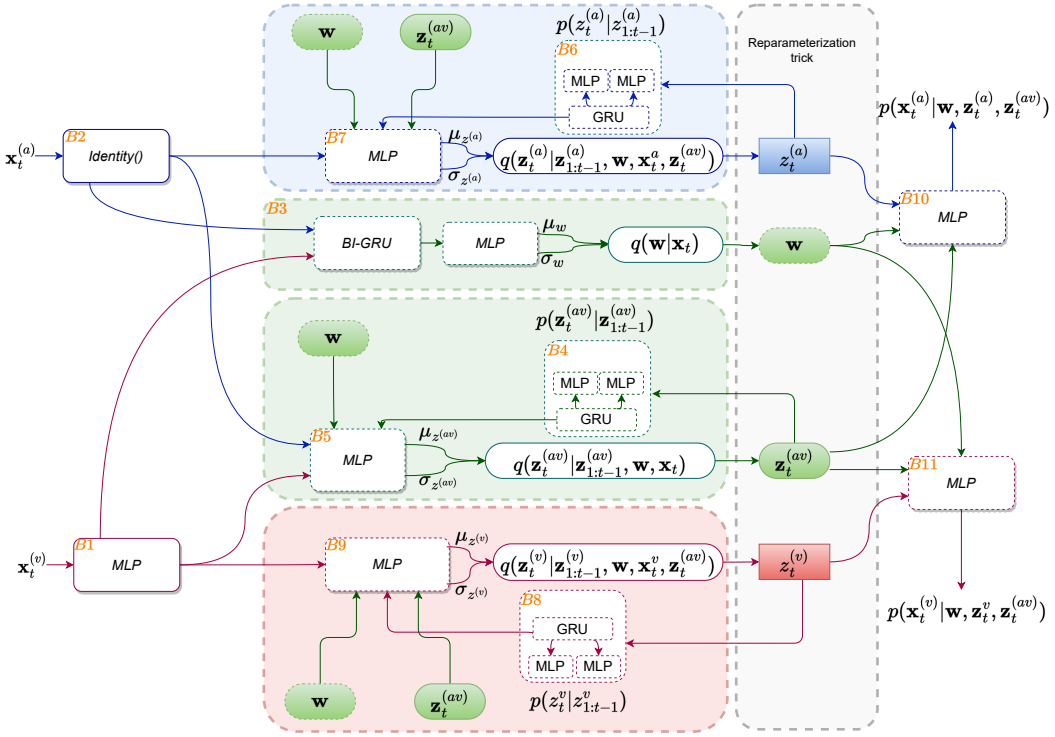


Figure B.1 – (Better zoom in) The overall architecture of the MDVAE.

two decoders, one for the visual and the other for the audio modality. Structured only with linear layers and non-linear activation functions, the input of these two decoders are the concatenation of $\mathbf{w}, \mathbf{z}_t^{(av)}, \mathbf{z}_t^{(v)}$ and $\mathbf{w}, \mathbf{z}_t^{(av)}, \mathbf{z}_t^{(a)}$ for the visual and audio modalities, respectively. Table B.3 and Figure B.1 present the details of the MDVAE architecture. The figure overviews the MDVAE architecture, including the connections between the blocks and the variables. The table complements the figure by detailing each block individually, including its dimensions, activation functions, and other relevant information. Together, the table and figure comprehensively describe the MDVAE architecture.

B.2 Visualization of the MDVAE static latent space

2D visualizations of the static latent space of the MDVAE are obtained using dimension reduction methods. Figure B.2(a) shows visualizations obtained with PCA and ISOMAP for one single speaker in the MEAD dataset, and the colors indicate the emotion labels. It can be seen that different emotions form different clusters, and the neutral emotion is approximately in the middle. Figure B.2(b) corresponds to the exact visualization, but

Table B.3 – The architecture details of the MDVAE. The blocks from B1 to B11 are illustrated in Figure B.1 to better understand their interactions.

Block	Layer	Activation	Output dim.
$B1$	`Linear(`$32 \cdot 8 \cdot 8$`, 1024)`	ReLu	1024
	`Linear(1024, 512)`	ReLu	$r_v = 512$
$B2$	`Identity`	-	$r_a = 512$
$B3$	`GRU(`$r_v + r_a$`, 256, 1, True)`	-	$2 \cdot 256$
	`Linear(2 · 256, 256)`	Tanh	256
	$\sigma_{\mathbf{w}}$`: Linear(256, `l_w`)`	-	l_w
	$\mu_{\mathbf{w}}$`: Linear(256, `l_w`)`	-	l_w
$B4$	`GRU(`l_{av}`, 128, 1, False)`	-	h_{av}
	`Linear(128, 64)`	ReLu	64
	`Linear(64, `l_{av}`)`	-	l_{av}
	`Linear(64, `l_{av}`)`	-	l_{av}
$B5$	`Linear(`$r_v + r_a + h_{av} + l_w$`, 256)`	ReLu	256
	`Linear(256, 128)`	ReLu	128
	$\sigma_{\mathbf{z}^{(av)}}$`: Linear(128, `$l_{av}$`)`	-	l_{av}
	$\mu_{\mathbf{z}^{(av)}}$`: Linear(128, `l_{av}`)`	-	l_{av}
$B6$	`GRU(`l_a`, 128, 1, False)`	-	h_a
	`Linear(128, 32)`	ReLu	32
	`Linear(32, `l_a`)`	-	l_a
	`Linear(32, `l_a`)`	-	l_a
$B7$	`Linear(`$r_a + h_a + l_w$`, 128)`	Tanh	128
	`Linear(128, 32)`	Tanh	32
	$\sigma_{\mathbf{z}^{(a)}}$`: Linear(32, `$l_a$`)`	-	l_a
	$\mu_{\mathbf{z}^{(a)}}$`: Linear(32, `l_a`)`	-	l_a
$B8$	`GRU(`l_v`, 128, 1, False)`	-	h_v
	`Linear(128, 64)`	ReLu	64
	`Linear(64, `l_v`)`	-	l_v
	`Linear(64, `l_v`)`	-	l_v
$B9$	`Linear(`$r_v + h_v + l_w$`, 256)`	ReLu	256
	`Linear(256, 128)`	ReLu	128
	$\sigma_{\mathbf{z}^{(v)}}$`: Linear(128, `$l_v$`)`	-	l_v
	$\mu_{\mathbf{z}^{(v)}}$`: Linear(128, `l_v`)`	-	l_v
$B10$	`Linear(`$l_v \cdot l_{av} \cdot l_w$`, 512)`	ReLu	512
	`Linear(512, 1024)`	ReLu	1024
	`Linear(1024, 2048)`	ReLu	2048
$B11$	`Linear(`$l_a \cdot l_{av} \cdot l_w$`, 128)`	Tanh	128
	`Linear(128, 256)`	Tanh	256
	`Linear(256, 512)`	Tanh	512

`GRU(input_size, hidden_size, num_layers, bidirectional)`
`Linear(input_size, output_size)`

the colors now indicate the emotion intensity levels. For each emotion, the intensity level increases continuously from the middle to the outside of the emotion cluster. Finally, Figure B.2(c) shows the identity clusters for six different speakers (left figure) and the emotion clusters for two speakers (figure on the right), both obtained using PCA. 3D visualizations and other dimension reduction methods are available on the companion website.

B.3 Interpolation in static latent space

In this section, we exploit the static latent space of MDVAE to perform temporal interpolation between two sequences. The interpolation is performed according to the following equation:

$$\tilde{\mathbf{w}}_t = \alpha_t \mathbf{w} + (1 - \alpha_t)\mathbf{w}'; \quad \alpha_t = (T - t)/(T - 1) \tag{B.1}$$

The visual sequence is generated as a following:

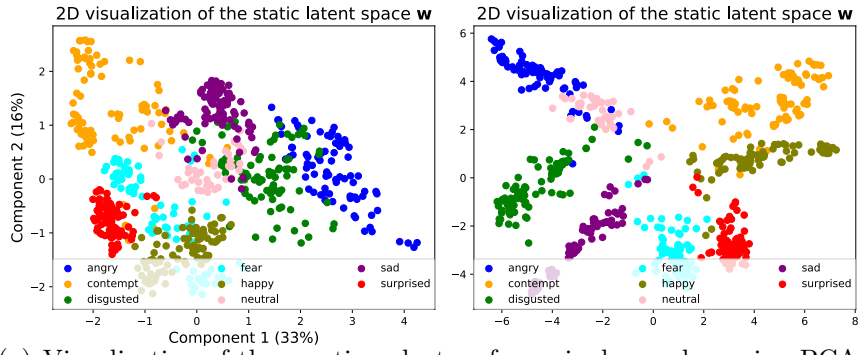
$$p_\theta(\mathbf{x}^{(v)}|\mathbf{z}^{(av)}, \mathbf{z}^{(v)}, \tilde{\mathbf{w}}) = \prod_t p_\theta(\mathbf{x}_t^{(v)}|\mathbf{z}_t^{(av)}, \mathbf{z}_t^{(v)}, \tilde{\mathbf{w}}_t) \tag{B.2}$$

We illustrate this interpolation with qualitative examples. Figure B.3(a) interpolates the emotions while keeping the same identity. Figure B.3(b) interpolates the identity while keeping the same emotion.

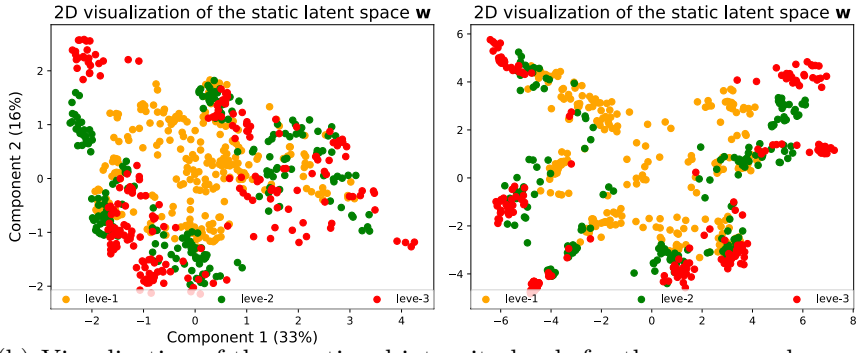
B.4 Conditional generation experiment

MDVAE is a generative model, able to generate new data by sampling following the distribution of the prior. Being hierarchical, we are also able to generate and edit some variation factors while keeping some others.

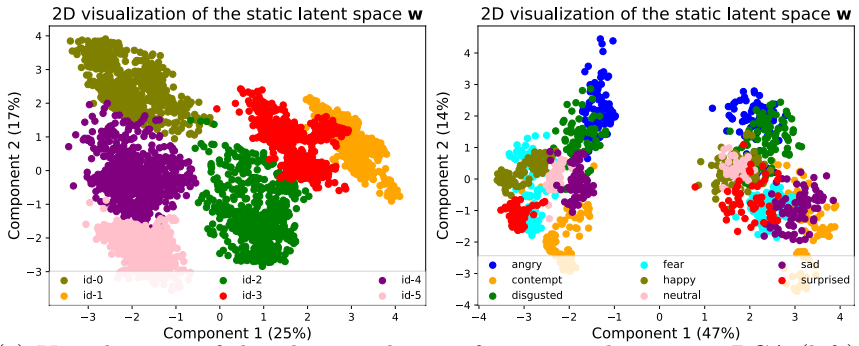
Generation of audio sequences We illustrate the ability of the MDVAE model to generate speech spectrograms with a qualitative example. Figures B.4(c) and B.4(b) show spectrograms generated by sampling according to the prior distribution of $\mathbf{z}^{(av)}$ and $\mathbf{z}^{(a)}$, respectively, conditioned on the first thirty frames (the lighted block in the figures). These first thirty frames (with a duration of 1s) are obtained through analysis-resynthesis. After



(a) Visualization of the emotion clusters for a single speaker using PCA (left) and ISOMAP (right).



(b) Visualization of the emotional intensity levels for the same speaker as in Figure B.2(a) using PCA (left) and ISOMAP (right).



(c) Visualization of the identity clusters for six speakers using PCA (left) and visualization of the emotion clusters for two speakers using PCA (right).

Figure B.2 – 2D visualizations of the static latent space.

1s, we switch the MDVAE to pure generation mode. We can see that the spectrograms generated with the prior distribution of $\mathbf{z}^{(av)}$ exhibit a harmonic structure.

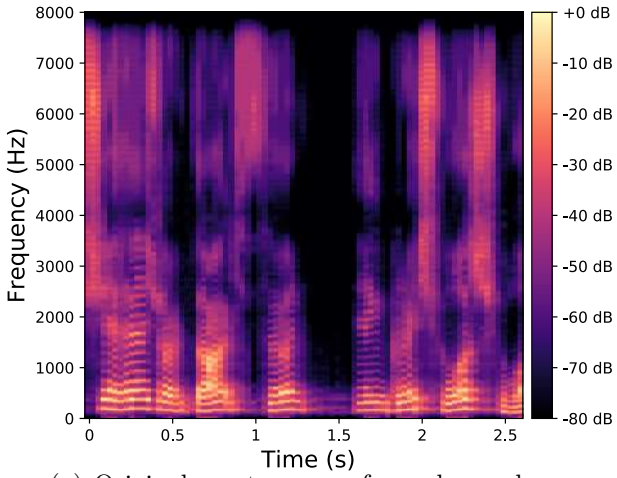
Generation of visual sequences Similarly, we illustrate the ability of the MDVAE model to generate visual frames with a qualitative example. Figure B.5(a) and B.5(b)

(a) Same identity, different emotions.

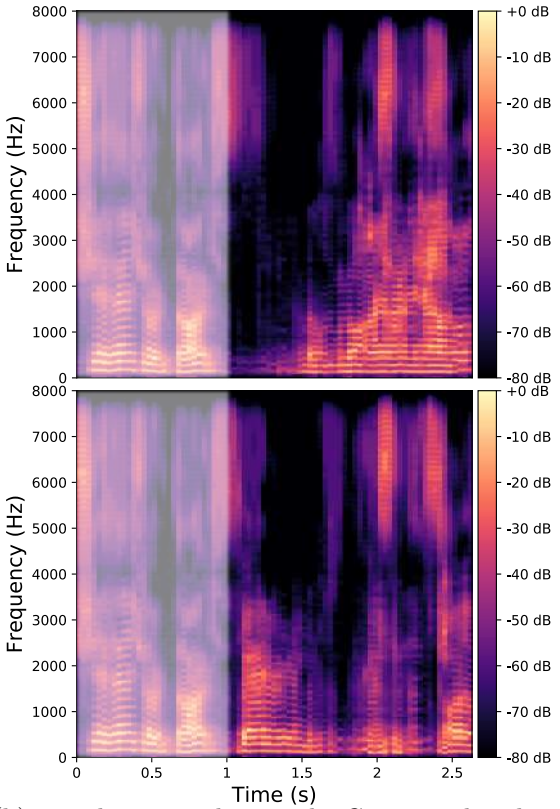
(b) Same emotion, different identities.

Figure B.3 – Qualitative results on the interpolation of $\mathbf{w}$. (top) the interpolation is done with the same person but between two different emotions; (bottom) the interpolation is done with the same emotion but between two different persons.

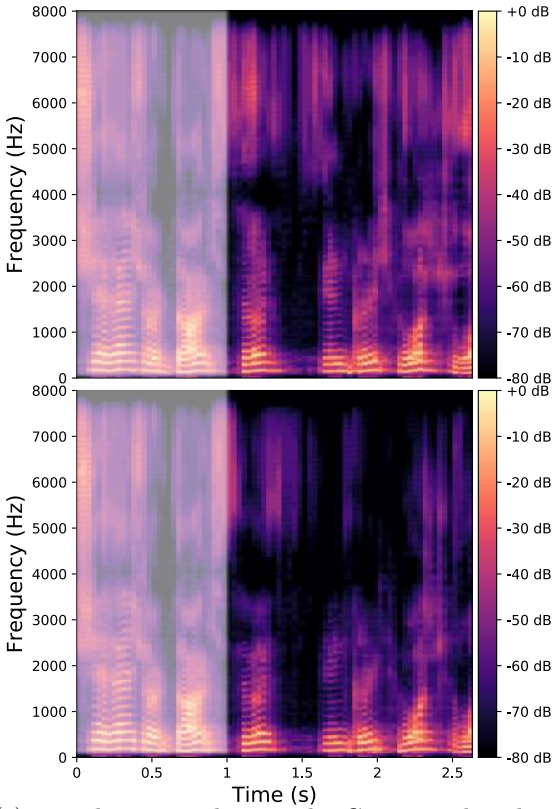
respectively illustrate this by sampling following the Gaussian prior of $\mathbf{z}^{(v)}$ and $\mathbf{z}^{(av)}$, conditioned on the analyzed-resynthesized first frame of the sequence. This conditioning strategy to the first frame allows the generation mode to have temporally coherent sequences. The first lightened lines correspond to the original visual sequences; the other lines are three different generations. As expected, we have a smooth transition from the analyzed-resynthesized frame to the generated one.



(a) Original spectrogram of a male speaker.



(b) sampling according to the Gaussian distribution prior of $\mathbf{z}^{(av)}$.



(c) sampling according to the Gaussian distribution prior of $\mathbf{z}^{(a)}$.

Figure B.4 – Example of speech power spectrogram reconstructed (0-1 s) and generated (1-2.6 s) by an MDVAE model.

(a) we sample according to the Gaussian distribution prior of $\mathbf{z}^{(v)}$.

(b) we sample according to the Gaussian distribution prior of $\mathbf{z}^{(av)}$.

Figure B.5 – We illustrate the ability of the MDVAE model to generate visual frames with a qualitative example. The figure illustrates this by sampling following the Gaussian prior of $\mathbf{z}^{(v)}$ and $\mathbf{z}^{(av)}$, conditioned on the analyzed-resynthesized first frame of the sequence. This conditioning strategy to the first frame allows the generation mode to have temporally coherent sequences. The first lightened lines correspond to the original visual sequences; the other lines are three different generations. As expected, we have a smooth transition from the analyzed-resynthesized frame to the generated one.

B.5 Generalization to other modalities

To generalize to other modalities, we trained MDVAE on two other types of modalities:

Visual Lip landmarks-MDVAE We trained VQ-MDVAE on the MEAD database, replacing the audio modality with the lip landmarks only.

When training the MDVAE on the visual part and lip landmarks, one would intuitively expect to find the visual movement of the lips and their landmarks in the shared dynamical latent space $\mathbf{z}^{(av)}$. The visual modality-specific latent space would contain other visual information, such as eyelid and head movements. Finally, emotion and identity would be found in the static latent space $\mathbf{w}$.

We conducted qualitative analyses involving the exchange of latent variables between sequences, a process outlined in detail in Section 4.3.4. Figure B.6 depicts one such experiment: the red and blue boxes contain original sequences (image sequence + landmark sequence), while the resulting four sequences within the grey boxes represent the qualitative outcomes. The two grey blocks on the left showcase sequences reconstructed by interchanging the specific dynamical variable $\mathbf{z}^{(v)}$ related to the visual modality. Notably, the lip movements and landmarks remain consistent, while the eye movements undergo changes, such as an eye blink depicted in the lower left block. The two grey blocks on the right showcase sequences reconstructed by interchanging the multimodal dynamical variable $\mathbf{z}^{(av)}$. Notably, the eye movements remain consistent, while lip movements and landmarks synchronize with the target sequence. Finally, the four blocks jointly show that $\mathbf{w}$ encodes identity and overall emotional state.

Double view-MDVAE We also trained VQ-MDVAE on the MEAD database, replacing the audio modality with the visual left view.

Intuitively, one would expect to find all visual movement, including lip and eyelid movements, in the shared dynamical latent space $\mathbf{z}^{(av)}$. The latent spaces specific to the two modalities would not encode any information. Finally, emotion and identity would be found in the static latent space $\mathbf{w}$.

As before, we have carried out a qualitative analysis involving the exchange of latent variables between sequences. Figure B.7 depicts one such experiment: the red and blue boxes contain original sequences (front view sequence + front left landmark sequence), while the resulting sequences below within the grey boxes represent the qualitative outcomes, interchanging the multimodal dynamical variable $\mathbf{z}^{(av)}$. Notably, the eye movements, lip movements, and facial movements synchronize with the target sequence.

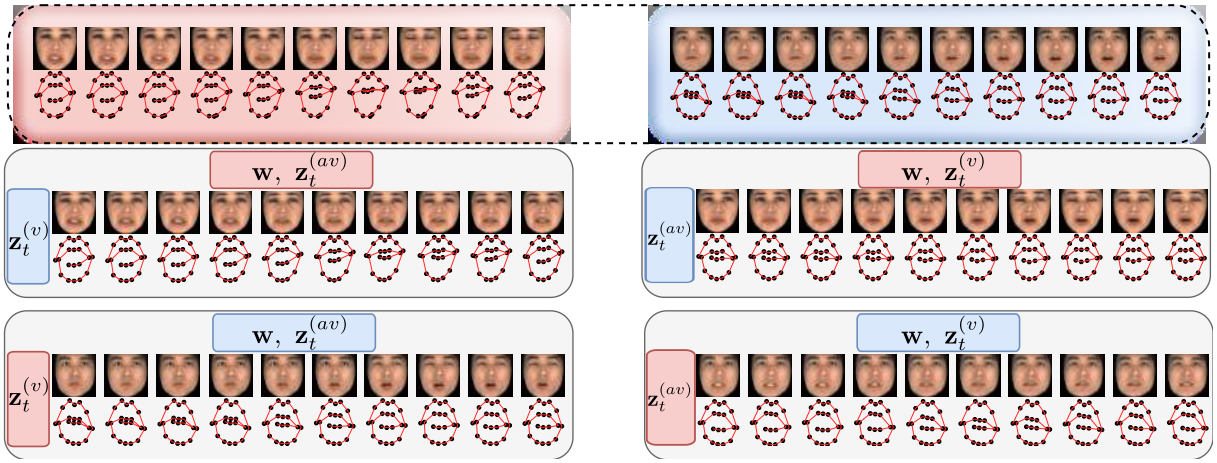


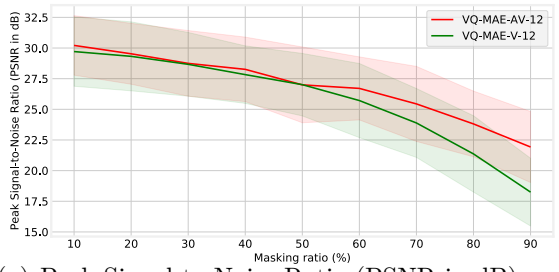
Figure B.6 – The two sequences in the red and blue boxes represent visual sequences and associated lip landmarks for two different individuals, and the sequences in the gray box represent sequences generated by MDVAE trained on the visual modality and the landmarks; each sequence is generated by interchanging the latent spaces of the two sequences on top.

Figure B.7 – The two sequences in the red and blue boxes represent visual sequences of two different views: a left view and a front view, and the sequences in the gray box represent sequences generated by MDVAE trained on the visual modality of the two different views; each sequence is generated by interchanging the latent spaces of the two sequences on top.

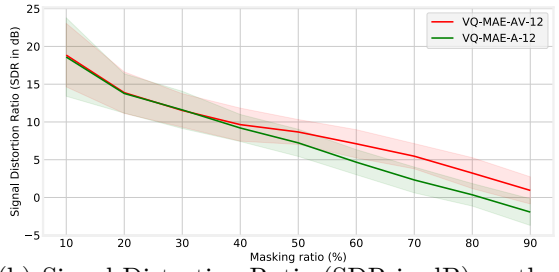
Appendix C

Appendix: VQ-MAE-AV

C.1 Audiovisual speech reconstruction quality



(a) Peak Signal-to-Noise Ratio (PSNR in dB) on the y-axis as a function of masking ratio (%) on the x-axis.



(b) Signal Distortion Ratio (SDR in dB) on the y-axis as a function of masking ratio (%) on the x-axis.

Figure C.1 – Quantitative results of the visual reconstruction (a) and audio reconstruction (b). The solid line represents the mean across the test examples of the VoxCeleb2 dataset, and the shaded area corresponds to the standard deviation.

In this experiment, we evaluate the reconstruction quality of VQ-MAE-AV when applied to masked audiovisual speech data. The model is fed with a sequence of tokens, some of which have been masked (we will study the reconstruction performance for different masking ratios), and it is used to predict the masked tokens, i.e. to reconstruct the complete audiovisual speech sequence from partially observed tokens.

For this experiment, we compare VQ-MAE-AV-*12* to VQ-MAE-A-*12* (A for audio) and VQ-MAE-V-*12* (V for visual), which are the unimodal versions of VQ-MAE-AV. The average quality performance for the speech and visual modalities is evaluated using the VoxCeleb2 test set. The Peak Signal-to-Noise Ratio (PSNR in dB) is used to assess the quality of the resynthesized visual data, the and Signal-to-Distortion Ratio (SDR in dB) is used to assess the quality of the resynthesized audio data.

Figures C.1(a) and C.1(b) present the PSNR and SDR curves, respectively, for the reconstruction quality of the visual and audio modalities as a function of the masking ratio.

Table C.1 – Performance of VQ-MAE-AV using the *joint fusion* strategy for both encoder and decoder (without contrastive learning), fine-tuned on RAVDESS for different encoder depths.

Method	Parm. (M)	Acc. (%)	f1-score (%)
VQ-MAE-AV-*6*	8.5	75.7	76.0
VQ-MAE-AV-*12*	13.5	81.5	80.1
VQ-MAE-AV-*16*	16.8	82.4	82.4
VQ-MAE-AV-*20*	20.2	81.3	81.3

Table C.2 – Performance of VQ-MAE-AV using the *cross fusion* strategy for both encoder and decoder (with contrastive and generative learning), fine-tuned using different masking strategies.

Masking strategy	Ratio (%)	Acc. (%)
Fixed	70	84.0
	80	84.2
	90	82.9
Dirichlet distribution	-	84.8

Notably, VQ-MAE-AV outperforms VQ-MAE-V for masking ratios greater than 50%. At a masking ratio of 90%, VQ-MAE-AV achieves a significant 3.68 dB gain in PSNR over VQ-MAE-V. For the audio modality, VQ-MAE-AV outperforms VQ-MAE-A for masking ratios greater than 40% and records a gain of 2.87 dB in SDR at 90% of masking. In summary, this experiment highlights the effectiveness of leveraging multimodality for improving reconstruction quality. Additional qualitative results are presented in the companion website of this paper [1]. In particular, we show that VQ-MAE-AV can reconstruct certain information in a modality even when it is completely masked (masking ratio of 100%). This is particularly interesting for the audio modality, as the model attempts to reconstruct certain phonemes and speech structures from the visual modality based solely on lip movements.

C.2 Exploring supplementary abstract study

Impact of the encoder depth Table C.1 shows the impact of encoder depth on emotion recognition performance by varying the number of attention blocks (L), where a model with n attention blocks in the encoder will be denoted by VQ-MAE-AV-n. The results show that increasing the number of blocks in the encoder leads to improved performance

1. https://samsad35.github.io/VQ-MAE-AudioVisual/

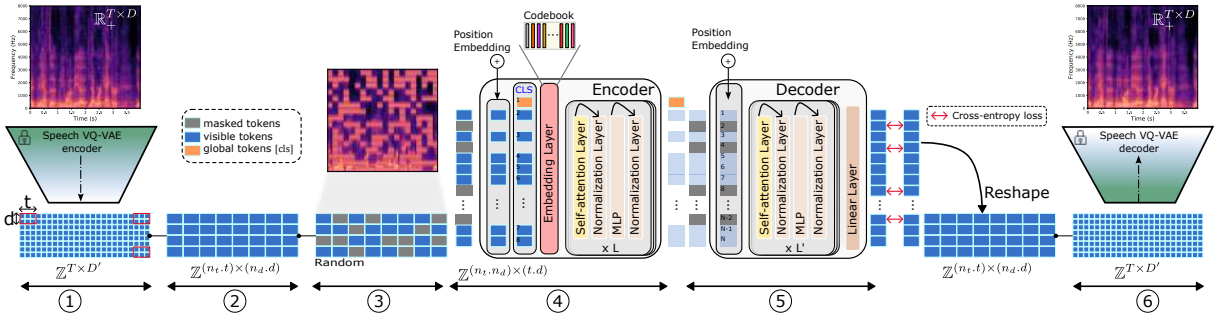


Figure C.2 – VQ-MAE-S model structure

to a certain extent. However, when the number of blocks becomes too high (as in the case of VQ-MAE-AV-*20*), there is no further improvement, and performance actually decreases by 1.1% accuracy compared to VQ-MAE-AV-*16*.

Effect of Masking Strategy Table C.2 compares the performance of two masking techniques: fixed ratio masking (70%, 80%, and 90%) for both modalities and masking based on a Dirichlet distribution sampled with parameters ($\alpha_a = \alpha_v = 1$), as discussed in the paper. Notably, masking according to the Dirichlet distribution outperforms fixed ratio masking, achieving an accuracy gain of 0.8, 0.6, and 1.9, respectively, compared to fixed ratios of 70%, 80%, and 90%.

C.3 VQ-MAE for audio speech representation

Description We introduce the vector quantized MAE for speech (VQ-MAE-S), a self-supervised model applied to emotion recognition in speech signals. VQ-MAE-S is an adapted version of the Audio-MAE model proposed in (Baade et al., 2022; Gong, Lai, et al., 2022; Xu et al., 2022). Unlike Audio-MAE, VQ-MAE-S operates on the discrete latent representation of a vector-quantized variational autoencoder (VQ-VAE) (Van den Oord et al., 2017) instead of the spectrogram representation.

Set-up and fine-tuning The speech VQ-VAE training setup and architecture closely resemble those discussed in Section 5.3.1. The notable difference is in the spectrogram pre-processing, with the overlap parameter adjusted to 75% (as opposed to 68%) for this particular case.

Regarding the architecture of VQ-MAE-S, both the encoder and decoder are comprised

Table C.3 – Overall results (accuracy (%) and f1-score (%)) on the four evaluation databases.

DATASET	RAVDESS-Speech		RAVDESS-Song		IEMOCAP		EMODB	
Metrics	Accuracy	f1-score	Accuracy	f1-score	Accuracy	f1-score	Accuracy	f1-score
Self-attention audio (Chumachenko et al., 2022)	58.3	-	-	-	-	-	-	-
SSAST (Gong, Lai, et al., 2022) (Patch-*tf*)	-	-	-	-	54.3	-	-	-
MAE-AST (Baade et al., 2022) (Patch-*tf*)	-	-	-	-	58.6	-	-	-
SpecMAE-12 (Patch-*tf*)	52.2	52.0	54.5	53.9	46.7	45.9	57.2	57.0
VQ-MAE-S-*12* (Patch-*tf*)	76.7	75.9	84.0	84.0	61.9	61.2	85.7	85.8
VQ-MAE-S-*12* (Patch-*tf*) + Query2Emo	78.2	77.5	83.7	83.4	63.1	62.5	88.4	88.3
VQ-MAE-S-*12* (Frame)	80.8	80.5	84.2	84.3	65.2	64.9	87.0	87.0
VQ-MAE-S-*12* (Frame) + Query2Emo	**84.1**	**84.4**	**85.8**	**85.7**	**66.4**	**65.8**	**90.2**	**89.1**

of successive attention blocks. Specifically, there are 12 blocks in the encoder and four blocks in the decoder. Training parameters are identical to those in Section 5.3.1.

Emotional databases for fine-tuning and evaluation We fine-tune and evaluate the proposed approaches on four emotional speech audio databases.

— **RAVDESS-Speech** (Livingstone & Russo, 2018): This English database consists of 1440 audio files recorded by 24 professional actors and labeled with eight different emotional expressions (neutral, calm, happy, sad, angry, fearful, disgust, surprised).

— **RAVDESS-Song** (Livingstone & Russo, 2018): Same as the RAVDESS-Speech database, but utterances are sung *a capella*. This database contains a total of 1012 audio files recorded by 23 actors and labeled with six emotions (neutral, calm, happy, sad, angry, and fearful).

— **IEMOCAP** (Busso et al., 2008): This database comprises approximately 12 hours of audio, annotated with several emotions, but only four emotions (neutral, happy, angry, and disgusted) have been retained to ensure a balanced distribution. It consists of dyadic sessions in which actors participate in improvisations or scripted scenarios.

— **EMODB** (Burkhardt et al., 2005): The German EMODB database comprises 535 utterances spoken by ten professional speakers. It includes seven emotions (anger, boredom, anxiety, happiness, sadness, disgust, and neutral).

Performance on speech emotion recognition Table C.3 compares the SER performance (accuracy and F1-score) of the proposed VQ-MAE-S model (with classification

from the [CLS] token), its improved version VQ-MAE-S + Query2Emo, the SpecMAE-12 baseline, and three state-of-the-art methods: SSAST (Gong, Lai, et al., 2022), MAE-AST (Baade et al., 2022), and a supervised self-attention-based approach (Chumachenko et al., 2022) on the four evaluation databases. Two configurations of masking are considered: random frame masking (Frame) and random time-frequency patch masking (Patch-*tf*). The results indicate that the proposed models outperform all other methods across all databases. For random time-frequency patch masking (Patch-*tf*), VQ-MAE-S achieves 15.2% better accuracy than SpecMAE, 7.6% better accuracy than SSAST, and 3.3% better accuracy than MAE-AST on the IEMOCAP dataset. The accuracy improvement over the supervised method on the RAVDESS-Speech database is 18.4%. Query2Emo also contributes to the SER performance, with a gain of 1.5%, 1.2%, and 2.7% compared to VQ-MAE-S (Patch-*tf*) alone on RAVDESS-Speech, IEMOCAP, and EMODB, respectively.

We conducted an ablation study to assess the importance of several hyperparameters of the proposed VQ-MAE-S model (impact of the masking ratio, the masking strategy, the continuous embedding token size, etc.), which are presented in Sadok, Leglaive, and Séguier, 2023.

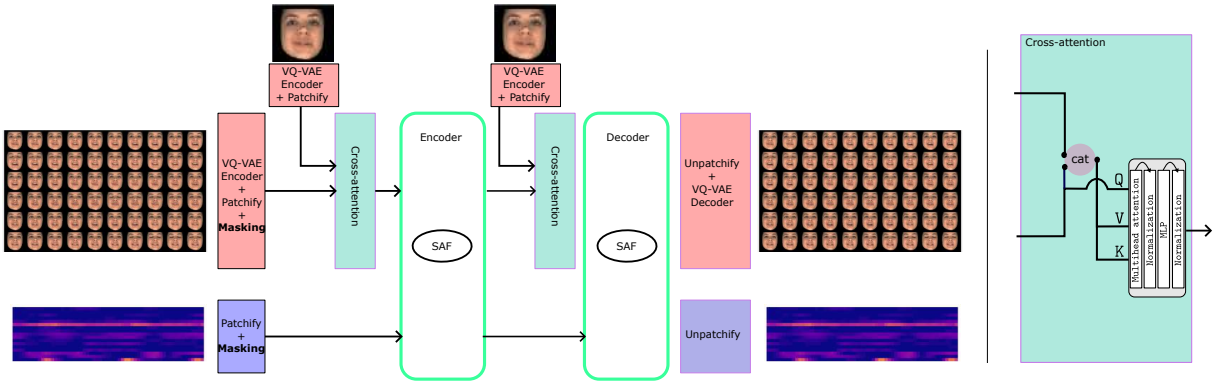


Figure C.3 – VQ-MAE-AU model structure

C.4 VQ-MAE for visual and action units representation

Description We introduce another multimodal VQ-MAE variant, denoted as VQ-MAE-AU, where the video modality (a sequence of images) and the action unit modality serve as inputs to the model. In both the encoder and decoder, we employ the "self-attention

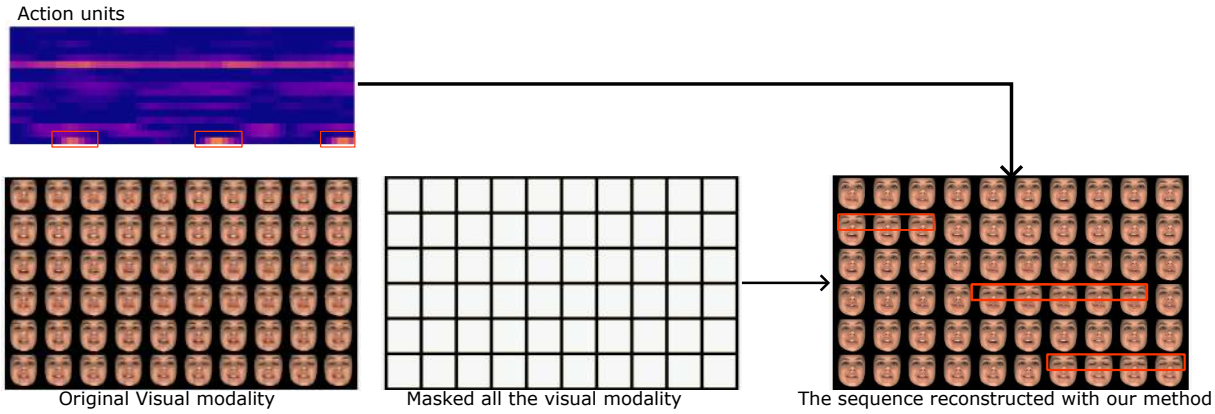


Figure C.4 – Cross-modal: from action units to images.

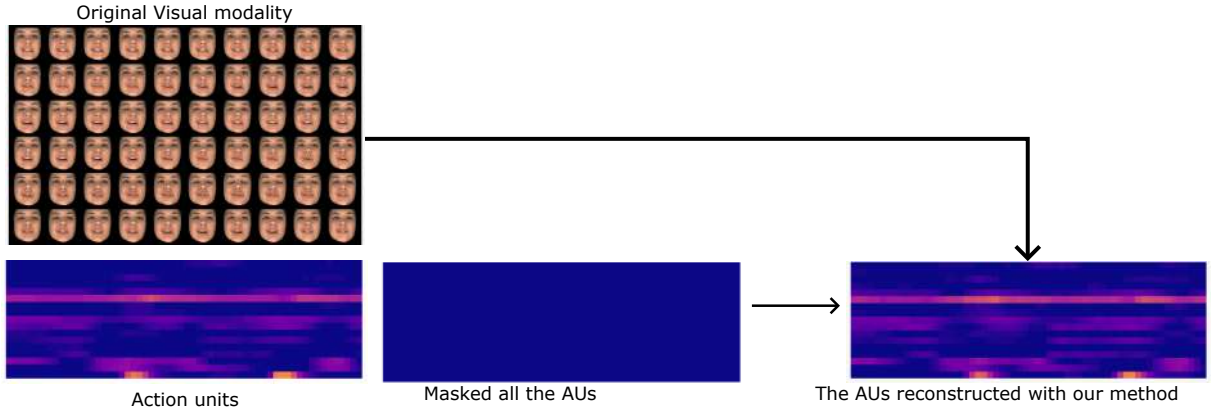


Figure C.5 – Cross-modal: from images to action units.

fusion" technique to merge these two modalities (refer to Section 5.2.5 for details). In this model, we condition the visual modality on the first frame through a cross-attention block in both the encoder and the decoder, as illustrated in Figure C.3. This block is used to learn identity information from the first frame.

Set-up The VQ-MAE-AU model is trained on the multi-view emotional audiovisual dataset (MEAD) (K. Wang et al., 2020). Face images in the MEAD dataset are cropped, resized to a 112x112 resolution, and aligned using Openface (Baltrušaitis et al., 2016). Additionally, Openface is employed for extracting action units. The action units in Openface have a scale ranging from 0 to 1, which is discretized in our study to a scale of 0 to 50 (just multiply by 50 and take the integer value).

Cross-modal and transformation Figures C.4 and C.5 demonstrate the cross-model generation process from action units to image sequence and from image sequence to action units, respectively. In Figure C.4, our model showcases the reconstruction of the entire image sequence based on unit actions and conditioned by the initial image. The reconstruction captures lip and eye movements. Specifically, the bottom row in the original AUs corresponds to AU_{43}. We have indicated three red rectangles responsible for eye blinking. The generated sequence exhibits synchronous eye blinking within temporal windows where AU_{43} activation occurs, demonstrating the faithful translation of this action unit.

Appendix: Graphical interfaces

A graphical interface was developed in Python using the Tkinter library (Lundh, 1999) for each model presented in this manuscript. These interfaces not only facilitated the generation of a large volume of qualitative results but also significantly enhanced the user-friendliness of our trained models.

D.1 Graphical interface for source-filter VAE

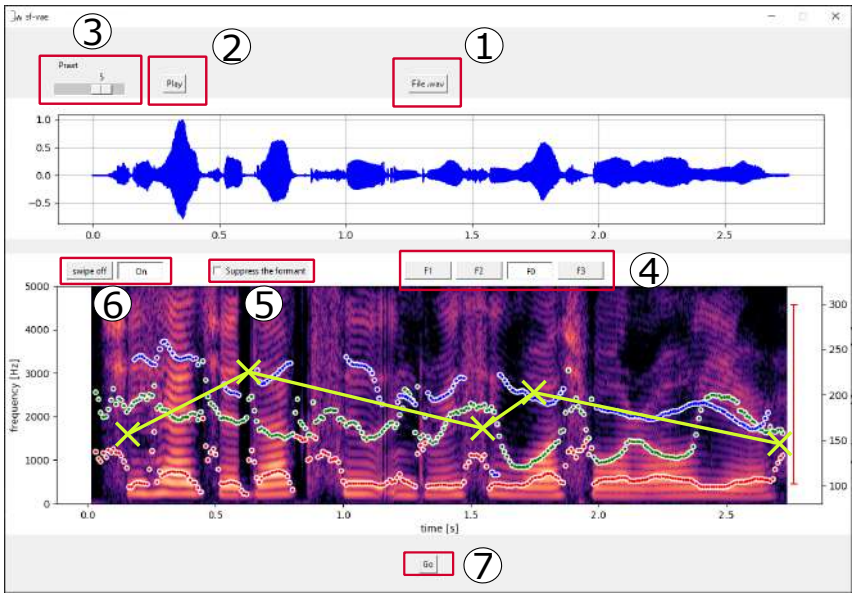


Figure D.1 – User Interface for Source-Filter VAE: ① Click the button to open a dialog box for loading an audio file onto your computer. ② Press the button to listen to the downloaded audio. ③ Set Praat to display formants. ④ Choose the factor of variation to be controlled. Simply draw a trajectory on the spectrogram with your mouse, as demonstrated in the yellow figure. ⑤ Use this button if you wish to remove the factor of variation. ⑥ Delete the trajectory of the new factor of variation. ⑦ Click here to activate the transformation.

D.2 Graphical interface for VQ-MDVAE

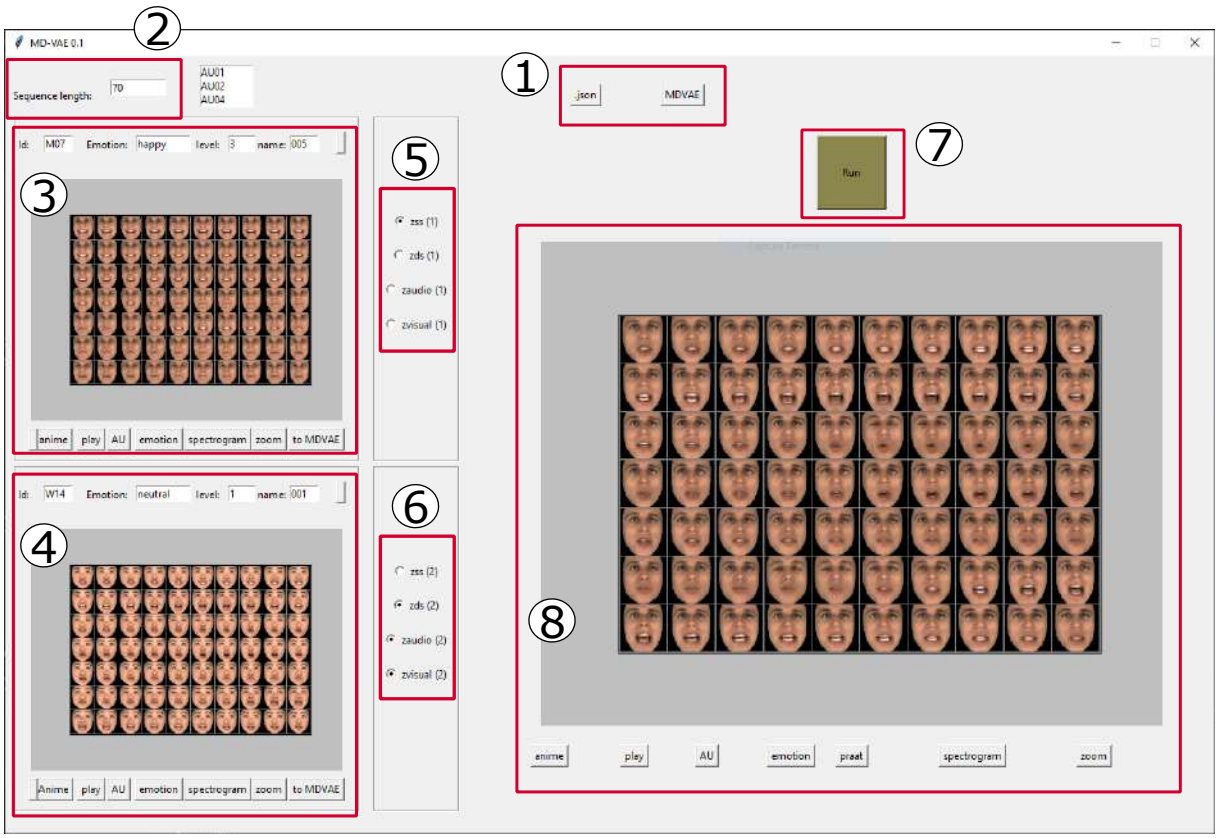


Figure D.2 – User Interface for VQ-MDVAE: ① Click the ".json" button to open a dialog box for selecting desired model parameters. Load the model by clicking the "MDVAE" button. ② Manage the number of images in the sequence. ③ and ④ Load audiovisual data; buttons at the bottom of each sequence provide options for zooming in on visual sequences, animating the sequence, tracking action units, and playing the respective audio. ⑤ and ⑥ Manage the latent vectors of the two audiovisual sequences; for example, in this example, keep the $\mathbf{w}$ of the sequence at the top (③) and the other latent vectors of the sequence at the bottom (④). ⑦ Click the "Run" button to synthesize the results. ⑧ View the audiovisual output of the VQ-MDVAE model.

D.3 Graphical interface for VQ-MAE-AV

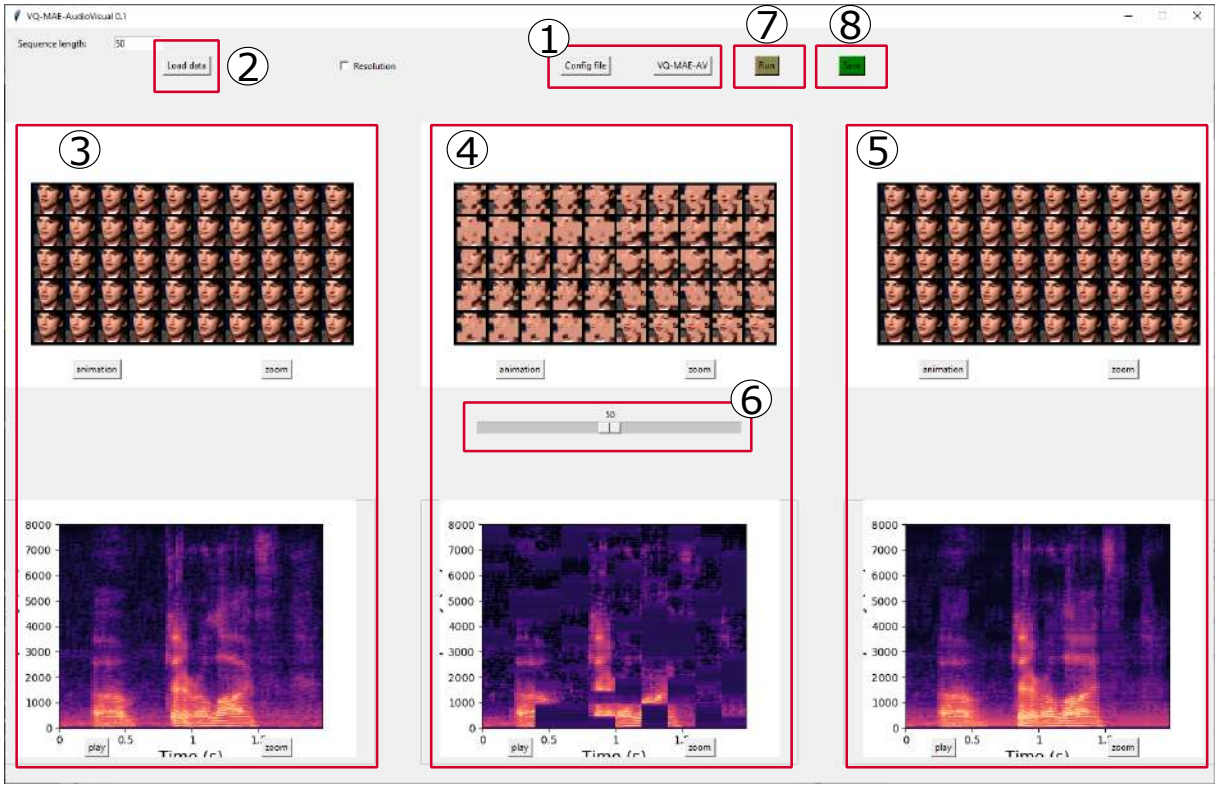


Figure D.3 – User Interface for VQ-MAE-AV: ① Click the "Config file" button to open a dialog box for selecting model parameters. Load the model by clicking the "VQ-MAE-AV" button. ② Use the "Load data" button to open a dialog box for downloading a video to your computer. The display in ③ opens automatically. ⑦ Click the "Run" button, and the displays ④ (showing masked data) and ⑤ (illustrating reconstruction using our model) will open. ⑥ Use the slider to adjust the masking percentage. ⑧ Click the "Save" button to save all the data (input, masked data, and reconstruction) to your desired folder path.

List of Figures

List of Tables

BIBLIOGRAPHY

Abate, A. F., Cimmino, L., Mocanu, B.-C., Narducci, F., & Pop, F., (2023), The limitations for expression recognition in computer vision introduced by facial masks, *Multimedia Tools and Applications*, *82 8*, 11305–11319 (cit. on p. 16).

Abbaschian, B. J., Sierra-Sosa, D., & Elmaghraby, A., (2021), Deep learning techniques for speech emotion recognition, from databases to models, *Sensors*, *21 4*, 1249 (cit. on p. 18).

Abdullah, S. M. S. A., Ameen, S. Y. A., Sadeeq, M. A., & Zeebaree, S., (2021), Multimodal emotion recognition using deep learning, *Journal of Applied Science and Technology Trends*, *2 02*, 52–58 (cit. on pp. 22, 169).

Abnar, S., & Zuidema, W., (2020), Quantifying attention flow in transformers, *Meeting of the Association for Computational Linguistics*, 4190–4197 (cit. on p. 162).

Afouras, T., Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A., (2018), Deep audio-visual speech recognition, *IEEE transactions on pattern analysis and machine intelligence* (cit. on p. 105).

Akbari, H., Yuan, L., Qian, R., Chuang, W.-H., Chang, S.-F., Cui, Y., & Gong, B., (2021), VATT: transformers for multimodal self-supervised learning from raw video, audio and text, *Advances in Neural Information Processing Systems*, *34*, 24206–24221 (cit. on pp. 139, 146).

Akuzawa, K., Iwasawa, Y., & Matsuo, Y., (2018), Expressive speech synthesis via modeling expressions with variational autoencoder, *Interspeech*, 3067–3071 (cit. on p. 80).

Alayrac, J.-B., Recasens, A., Schneider, R., Arandjelović, R., Ramapuram, J., De Fauw, J., Smaira, L., Dieleman, S., & Zisserman, A., (2020), Self-supervised multimodal versatile networks, *Advances in Neural Information Processing Systems*, *33*, 25–37 (cit. on pp. 139, 146).

Alias Partg Goyal, A. G., Sordoni, A., Côté, M.-A., Ke, N. R., & Bengio, Y., (2017), Z-forcing: training stochastic recurrent networks, *Advances in neural information processing systems*, *30* (cit. on p. 57).

Anikin, A., (2019), Soundgen: an open-source tool for synthesizing nonverbal vocalizations, *Behavior Research Methods*, *51 2*, 778–792 (cit. on pp. 84, 90, 177, 178).

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C., (2021), ViViT: a video vision transformer, *IEEE/CVF international conference on computer vision*, 6836–6846 (cit. on p. 139).

Arnela, M., Blandin, R., Dabbaghchian, S., Guasch, O., Alías, F., Pelorson, X., Van Hirtum, A., & Engwall, O., (2016), Influence of lips on the production of vowels based on finite element simulations and experiments, *The Journal of the Acoustical Society of America*, *139 5*, 2852–2859 (cit. on pp. 123, 143).

Augusma, A., Vaufreydaz, D., & Letué, F., (2023), Multimodal group emotion recognition in-the-wild using privacy-compliant features, *International Conference on Multimodal Interaction*, 750–754 (cit. on p. 105).

Baade, A., Peng, P., & Harwath, D., (2022), MAE-AST: masked autoencoding audio spectrogram transformer, *arXiv preprint arXiv:2203.16691* (cit. on pp. 139, 196–198).

Bachman, P., Hjelm, R. D., & Buchwalter, W., (2019), Learning representations by maximizing mutual information across views, *Advances in neural information processing systems*, *32* (cit. on p. 36).

Bachmann, R., Mizrahi, D., Atanov, A., & Zamir, A., (2022), MultiMAE: multi-modal multi-task masked autoencoders, *European Conference on Computer Vision*, 348–367 (cit. on pp. 136, 139, 143, 144).

Bai, K., Cheng, P., Hao, W., Henao, R., & Carin, L., (2023), Estimating total correlation with mutual information estimators, *International Conference on Artificial Intelligence and Statistics*, 2147–2164 (cit. on p. 54).

Balazs, J. A., & Velásquez, J. D., (2016), Opinion mining and information fusion: a survey, *Information Fusion*, *27*, 95–110 (cit. on p. 14).

Baltrušaitis, T., Ahuja, C., & Morency, L.-P., (2018), Multimodal machine learning: a survey and taxonomy, *IEEE transactions on pattern analysis and machine intelligence*, *41 2*, 423–443 (cit. on p. 104).

Baltrušaitis, T., Robinson, P., & Morency, L.-P., (2016), Openface: an open source facial behavior analysis toolkit, *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1–10 (cit. on pp. 115, 124, 199).

Bando, Y., Mimura, M., Itoyama, K., Yoshii, K., & Kawahara, T., (2018), Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 716–720 (cit. on pp. 76, 79, 83, 177).

Banno, H., Hata, H., Morise, M., Takahashi, T., Irino, T., & Kawahara, H., (2007), Implementation of realtime STRAIGHT speech manipulation system: report on its first implementation, *Acoustical Science and Technology*, *28 3*, 140–146 (cit. on p. 80).

Bao, H., Dong, L., Piao, S., & Wei, F., (2021), BEiT: bert pre-training of image transformers, *International Conference on Learning Representations (ICLR)* (cit. on pp. 70, 73, 139).

Bayer, J., & Osendorfer, C., (2014), Variational inference of latent state sequences using recurrent networks, *stat*, *1050*, 6 (cit. on pp. 56, 57).

Ben-Baruch, E., Ridnik, T., Zamir, N., Noy, A., Friedman, I., Protter, M., & Zelnik-Manor, L., (2020), Asymmetric loss for multi-label classification, *arXiv preprint arXiv:2009.14119*.

Bengio, Y., Courville, A., & Vincent, P., (2013), Representation learning: a review and new perspectives, *IEEE transactions on pattern analysis and machine intelligence*, *35 8*, 1798–1828 (cit. on pp. 28, 29, 37–39, 87, 104, 105, 169).

Bengio, Y., Goodfellow, I., & Courville, A., (2017), *Deep learning* (Vol. 1), MIT press Massachusetts, USA:

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H., (2006), Greedy layer-wise training of deep networks, *Advances in neural information processing systems*, *19* (cit. on p. 138).

Bengio, Y., Léonard, N., & Courville, A., (2013), Estimating or propagating gradients through stochastic neurons for conditional computation, *arXiv preprint arXiv:1308.3432* (cit. on p. 67).

Berry, M., Lewin, S., & Brown, S., (2022), Correlated expression of the body, face, and voice during character portrayal in actors, *Scientific Reports*, *12 1*, 1–13 (cit. on p. 125).

Bie, X., Leglaive, S., Alameda-Pineda, X., & Girin, L., (2022), Unsupervised speech enhancement using dynamical variational autoencoders, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 2993–3007. (Cit. on p. 79).

Bird, S., Achuthan, A., Maatallah, O. A., Hu, W., Janoyan, K., Kwasinski, A., Matthews, J., Mayhew, D., Owen, J., & Marzocca, P., (2014), Distributed (green) data centers: a new concept for energy, computing, and telecommunications, *Energy for Sustainable Development*, *19*, 83–91 (cit. on p. 165).

Bishop, C. M., (1998), Latent variable models, *Learning in graphical models*, *371* (cit. on p. 43).

Bishop, C. M., (2006a), *Pattern Recognition and Machine Learning*, Springer.

Bishop, C. M., (2006b), *Pattern Recognition and Machine Learning*, Springer, (cit. on pp. 45, 86).

Bishop, C. M., & Nasrabadi, N. M., (2006), *Pattern recognition and machine learning* (Vol. 4), Springer, (cit. on p. 110).

Blaauw, M., & Bonada, J., (2016), Modeling and transforming speech using variational autoencoders, *Interspeech*, 1770–1774 (cit. on p. 80).

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D., (2017), Variational inference: a review for statisticians, *Journal of the American Statistical Association*, *112 518*, 859–877.

Boersma, P., & Weenink, D., (2021a), Praat: doing phonetics by computer [Computer program], (cit. on p. 96).

Boersma, P., & Weenink, D., (2021b), Praat: doing phonetics by computer [computer program](2011), *Version*, *5 3*, 74 (cit. on p. 124).

Bora, A., Jalal, A., Price, E., & Dimakis, A. G., (2017), Compressed sensing using generative models, *International Conference on Machine Learning (ICML)*, 537–546 (cit. on p. 76).

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J., (2003), The theoretical status of latent variables, *Psychological review*, *110 2*, 203 (cit. on p. 43).

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., & Lerchner, A., (2018), Understanding disentangling in \beta-VAE, *arXiv preprint arXiv:1804.03599* (cit. on pp. 52, 53).

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B., et al., (2005), A database of german emotional speech., *International Speech Communication Association (Interspeech)*, *5*, 1517–1520 (cit. on p. 197).

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S., (2008), IEMOCAP: interactive emotional dyadic motion capture database, *Language resources and evaluation*, *42*, 335–359 (cit. on p. 197).

Byeon, Y.-H., & Kwak, K.-C., (2014), Facial expression recognition using 3d convolutional neural network, *International journal of advanced computer science and applications*, *5 12*.

Camacho, A., & Harris, J. G., (2008), A sawtooth waveform inspired pitch estimator for speech and music, *The Journal of the Acoustical Society of America*, *124 3*, 1638–1652 (cit. on p. 100).

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R., (2014), CREMA-d: crowd-sourced emotional multimodal actors dataset, *IEEE Transactions on affective computing*, *5 4*, 377–390 (cit. on p. 150).

Cao, S., Xu, P., & Clifton, D. A., (2022), How to understand masked autoencoders, *arXiv preprint arXiv:2202.03670* (cit. on pp. 72, 73).

Carbajal, G., Richter, J., & Gerkmann, T., (2021), Guided variational autoencoder for speech enhancement with a supervised classifier, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 681–685.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A., (2021), Emerging properties in self-supervised vision transformers, *International conference on computer vision (IEEE/CVF)*, 9650–9660 (cit. on pp. 70, 73).

Chefer, H., Gur, S., & Wolf, L., (2021), Transformer interpretability beyond attention visualization, *Conference on computer vision and pattern recognition (IEEE/CVF)*, 782–791 (cit. on p. 162).

Chelombiev, I., Houghton, C., & O'Donnell, C., (2019), Adaptive estimators show information compression in deep neural networks, *International Conference on Learning Representations (ICLR)* (cit. on p. 33).

Chen, L.-W., & Rudnicky, A., (2021), Exploring wav2vec 2.0 fine-tuning for improved speech emotion recognition, *arXiv preprint arXiv:2110.06309* (cit. on pp. 70, 138).

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., & Sutskever, I., (2020), Generative pretraining from pixels, *International conference on machine learning (ICML)*, 1691–1703 (cit. on p. 70).

Chen, R. T., Li, X., Grosse, R. B., & Duvenaud, D. K., (2018), Isolating sources of disentanglement in variational autoencoders, *Advances in neural information processing systems*, *31* (cit. on pp. 40, 52, 54, 55, 76, 77, 81, 85, 105, 170).

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G., (2020), A simple framework for contrastive learning of visual representations, *International conference on machine learning (ICML)*, 1597–1607 (cit. on pp. 69, 70, 139).

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P., (2016), InfoGAN: interpretable representation learning by information maximizing generative

adversarial nets, *Advances in neural information processing systems*, *29* (cit. on p. 40).

Chen, X., Fan, H., Girshick, R., & He, K., (2020), Improved baselines with momentum contrastive learning, *arXiv preprint arXiv:2003.04297* (cit. on p. 139).

Chen, Y., & Joo, J., (2021), Understanding and mitigating annotation bias in facial expression recognition, *International Conference on Computer Vision (IEEE/CVF)*, 14980–14991 (cit. on pp. 16, 19).

Choi, H.-S., Lee, J., Kim, W., Lee, J. H., Heo, H., & Lee, K., (2021), Neural analysis and synthesis: reconstructing speech from self-supervised representations, *Advances in Neural Information Processing Systems (NeurIPS)* (cit. on p. 80).

Chumachenko, K., Iosifidis, A., & Gabbouj, M., (2022), Self-attention fusion for audiovisual emotion recognition with incomplete data, *International Conference on Pattern Recognition (ICPR)*, 2822–2828 (cit. on pp. 131–133, 153, 154, 197, 198).

Chung, J., Nagrani, A., & Zisserman, A., (2018), VoxCeleb2: deep speaker recognition, *International Speech Communication Association (Interspeech)* (cit. on pp. 25, 140, 149, 172).

Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. C., & Bengio, Y., (2015), A recurrent latent variable model for sequential data, *Advances in neural information processing systems*, *28* (cit. on pp. 56, 57).

Courty, N., Flamary, R., Habrard, A., & Rakotomamonjy, A., (2017), Joint distribution optimal transportation for domain adaptation, *Advances in neural information processing systems*, *30* (cit. on p. 134).

Creager, E., Madras, D., Jacobsen, J.-H., Weis, M., Swersky, K., Pitassi, T., & Zemel, R., (2019), Flexibly fair representation learning by disentanglement, *International conference on machine learning (ICML)*, 1436–1445 (cit. on p. 52).

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A., (2018), Generative adversarial networks: An overview, *IEEE signal processing magazine*, *35 1*, 53–65 (cit. on p. 40).

Dahmani, S., Colotte, V., Girard, V., & Ouni, S., (2019), Conditional variational auto-encoder for text-driven expressive audiovisual speech synthesis, *Conference of the International Speech Communication Association (INTERSPEECH)*.

Dai, B., & Wipf, D., (2018), Diagnosing and enhancing VAE models, *International Conference on Learning Representations (ICLR)* (cit. on p. 85).

Daunhawer, I., Sutter, T. M., Chin-Cheong, K., Palumbo, E., & Vogt, J. E., (2021), On the limitations of multimodal vaes, *International Conference on Learning Representations (ICLR)* (cit. on p. 64).

Davis, S., & Mermelstein, P., (1980), Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *IEEE transactions on acoustics, speech, and signal processing*, *28 4*, 357–366 (cit. on p. 16).

Deisenroth, M. P., Faisal, A. A., & Ong, C. S., (2020), *Mathematics for machine learning*, Cambridge University Press, (cit. on pp. 45, 46).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L., (2009), ImageNet: a large-scale hierarchical image database, *IEEE conference on computer vision and pattern recognition*, 248–255 (cit. on pp. 70, 73).

Denning, P. J., & Denning, D. E., (2020), Dilemmas of artificial intelligence, *Communications of the ACM*, *63 3*, 22–24 (cit. on p. 161).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K., (2019), Bert: pre-training of deep bidirectional transformers for language understanding, *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186 (cit. on pp. 16, 70, 73, 139).

Dib, A., Ahn, J., Thebault, C., Gosselin, P.-H., & Chevallier, L., (2023), S2f2: self-supervised high fidelity face reconstruction from monocular image, *International Conference on Automatic Face and Gesture Recognition (FG)*, 1–8 (cit. on p. 138).

Dirac, P., (1953), The lorentz transformation and absolute time, *Physica*, *19 1--12*, 888–896, https://doi.org/10.1016/S0031-8914(53)80099-6

Dong, X., Bao, J., Zhang, T., Chen, D., Zhang, W., Yuan, L., Chen, D., Wen, F., & Yu, N., (2021), PeCo: perceptual codebook for bert pre-training of vision transformers, *arXiv preprint arXiv:2111.12710* (cit. on p. 140).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., (2020), An image is worth 16x16 words: transformers for image recognition at scale, *International Conference on Learning Representations (ICLR)* (cit. on pp. 70, 73, 139, 144).

Dubois, Y., Bloem-Reddy, B., Ullrich, K., & Maddison, C. J., (2021), Lossy compression for lossless prediction, *Advances in Neural Information Processing Systems*, *34*, 14014–14028 (cit. on p. 36).

Dupuis, K., & Pichora-Fuller, M. K., (2010), Toronto emotional speech set (TESS), (cit. on p. 99).

Ekman, P., (1973), Universal facial expressions in emotion, *Studia Psychologica*, *15 2*, 140–147 (cit. on p. 14).

Ekman, P., (1992), Are there basic emotions?

Ekman, P., & Friesen, W. V., (1978), Facial action coding system, *Environmental Psychology & Nonverbal Behavior* (cit. on p. 121).

El Ayadi, M., Kamel, M. S., & Karray, F., (2011), Survey on speech emotion recognition: features, classification schemes, and databases, *Pattern recognition*, *44 3*, 572–587.

Elbanna, G., Scheidwasser-Clow, N., Kegler, M., Beckmann, P., El Hajal, K., & Cernak, M., (2022), Byol-s: learning self-supervised speech representations by bootstrapping, *HEAR: Holistic Evaluation of Audio Representations*, 25–47.

Elman, J. L., (1990), Finding structure in time, *Cognitive science*, *14 2*, 179–211 (cit. on p. 16).

El-Nouby, A., Izacard, G., Touvron, H., Laptev, I., Jegou, H., & Grave, E., (2021), Are large-scale datasets necessary for self-supervised pre-training?, *arXiv preprint arXiv:2112.10740* (cit. on p. 73).

Ericsson, L., Gouk, H., Loy, C. C., & Hospedales, T. M., (2022), Self-supervised representation learning: introduction, advances, and challenges, *IEEE Signal Processing Magazine*, *39 3*, 42–62 (cit. on p. 20).

Eskimez, S. E., Duan, Z., & Heinzelman, W., (2018), Unsupervised learning approach to feature analysis for automatic speech emotion recognition, *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5099–5103 (cit. on p. 74).

Esser, P., Rombach, R., & Ommer, B., (2021), Taming transformers for high-resolution image synthesis, *IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883 (cit. on p. 140).

Fabius, O., & Van Amersfoort, J. R., (2014), Variational recurrent auto-encoders, *arXiv preprint arXiv:1412.6581* (cit. on pp. 56, 57).

Fang, H., Carbajal, G., Wermter, S., & Gerkmann, T., (2021), Variational autoencoder for speech enhancement with a noise-aware encoder, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 676–680.

Fant, G., (1970), *Acoustic theory of speech production*, Walter de Gruyter, (cit. on pp. 24, 77, 171).

Fefferman, C., Mitter, S., & Narayanan, H., (2016), Testing the manifold hypothesis, *Journal of the American Mathematical Society*, *29 4*, 983–1049 (cit. on p. 41).

Feichtenhofer, C., Li, Y., He, K., et al., (2022), Masked autoencoders as spatiotemporal learners, *Advances in neural information processing systems*, *35*, 35946–35958 (cit. on p. 139).

Ferradans, S., Papadakis, N., Peyré, G., & Aujol, J.-F., (2014), Regularized discrete optimal transport, *SIAM Journal on Imaging Sciences*, *7 3*, 1853–1882.

Févotte, C., Bertin, N., & Durrieu, J.-L., (2009), Nonnegative matrix factorization with the itakura-saito divergence: with application to music analysis, *Neural computation*, *21 3*, 793–830 (cit. on pp. 114, 143).

Fischer, I., (2020), The conditional entropy bottleneck, *Entropy*, *22 9*, 999 (cit. on p. 31).

Fisher, R. A., (1925), Theory of statistical estimation, *Mathematical proceedings of the Cambridge philosophical society*, *22 5*, 700–725 (cit. on p. 31).

Flanagan, J. L., & Golden, R. M., (1966), Phase vocoder, *Bell System Technical Journal*, *45 9*, 1493–1509 (cit. on p. 80).

Fleckenstein, K. S., (1991), Defining affect in relation to cognition: a response to susan mcleod, *Journal of Advanced Composition*, 447–453 (cit. on p. 13).

Floridi, L., & Cowls, J., (2022), A unified framework of five principles for ai in society, *Machine learning and the city: Applications in architecture and urban design*, 535–545 (cit. on p. 161).

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., et al., (2021), An ethical framework for a good ai society: opportunities, risks, principles, and recommendations, *Ethics, governance, and policies in artificial intelligence*, 19–39 (cit. on p. 161).

Fraccaro, M., Sønderby, S. K., Paquet, U., & Winther, O., (2016), Sequential neural models with stochastic layers, *Advances in neural information processing systems*, *29* (cit. on p. 57).

Gabbay, A., & Hoshen, Y., (2019), Demystifying inter-class disentanglement, *arXiv preprint arXiv:1906.11796* (cit. on p. 52).

Gao, C., & Shinkareva, S. V., (2021), Modality-general and modality-specific audiovisual valence processing, *Cortex*, *138*, 127–137 (cit. on p. 108).

Garofalo, J. S., Graff, D., Paul, D., & Pallett, D., (1993), CSR-I (WSJ0) Sennheiser LDC93S6B [Philadelphia: Linguistic Data Consortium], (cit. on pp. 90, 177).

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., & Zue, V., (1993), TIMIT acoustic phonetic continuous speech corpus [Philadelphia: Linguistic Data Consortium], (cit. on p. 95).

Garrido, Q., Chen, Y., Bardes, A., Najman, L., & Lecun, Y., (2022), On the duality between contrastive and non-contrastive self-supervised learning, *arXiv preprint arXiv:2206.02574* (cit. on p. 36).

Geiger, D., Verma, T., & Pearl, J., (1990), Identifying independence in bayesian networks, *Networks*, *20 5*, 507–534 (cit. on pp. 59, 110).

Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F., (2014), Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture., *Emotion*, *14 2*, 251 (cit. on p. 17).

Geng, X., Liu, H., Lee, L., Schuurams, D., Levine, S., & Abbeel, P., (2022), Multimodal masked autoencoders learn transferable representations, *arXiv preprint arXiv:2205.14204* (cit. on p. 139).

George, E. B., & Smith, M. J., (1997), Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model, *IEEE Transactions on Speech and Audio Processing*, *5 5*, 389–406 (cit. on p. 80).

Ghahramani, Z., (2003), Unsupervised learning. *In Summer school on machine learning* (pp. 72–112), Springer, (cit. on p. 20).

Ghaleb, E., Niehues, J., & Asteriadis, S., (2020), Multimodal attention-mechanism for temporal emotion recognition, *IEEE International Conference on Image Processing (ICIP)*, 251–255 (cit. on p. 153).

Ghaleb, E., Popa, M., & Asteriadis, S., (2019), Multimodal and temporal perception of audio-visual cues for emotion recognition, *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 552–558 (cit. on p. 153).

Gidaris, S., Singh, P., & Komodakis, N., (2018), Unsupervised representation learning by predicting image rotations, *International Conference on Learning Representations (ICLR)* (cit. on p. 139).

Girin, L., Leglaive, S., Bie, X., Diard, J., Hueber, T., & Alameda-Pineda, X., (2021a), Dynamical variational autoencoders: a comprehensive review, *Foundations and Trends in Machine Learning*, *15 1-2*, 1–175 (cit. on p. 112).

Girin, L., Leglaive, S., Bie, X., Diard, J., Hueber, T., & Alameda-Pineda, X., (2021b), Dynamical variational autoencoders: a comprehensive review, *Foundations and Trends in Machine Learning*, *15 1-2*, 1–175 (cit. on pp. 56, 57, 109, 110, 170).

Girin, L., Roche, F., Hueber, T., & Leglaive, S., (2019a), Notes on the use of variational autoencoders for speech and audio spectrogram modeling, *International Conference on Digital Audio Effects (DAFx)*, 1–8 (cit. on p. 114).

Girin, L., Roche, F., Hueber, T., & Leglaive, S., (2019b), Notes on the use of variational autoencoders for speech and audio spectrogram modeling, *International Conference on Digital Audio Effects (DAFx)*, 1–8 (cit. on pp. 83, 177).

Goetschalckx, L., Andonian, A., Oliva, A., & Isola, P., (2019), GANalyze: toward visual definitions of cognitive image properties, *IEEE/CVF International Conference on Computer Vision (ICCV)*, 5744–5753 (cit. on p. 82).

Goncalves, L., & Busso, C., (2022), Robust audiovisual emotion recognition: aligning modalities, capturing temporal information, and handling missing features, *IEEE Transactions on Affective Computing*, *13 4*, 2156–2170 (cit. on pp. 153, 154).

Gong, Y., Lai, C.-I., Chung, Y.-A., & Glass, J., (2022), Ssast: self-supervised audio spectrogram transformer, *AAAI Conference on Artificial Intelligence*, *36 10*, 10699–10709 (cit. on pp. 138, 139, 196–198).

Gong, Y., Rouditchenko, A., Liu, A. H., Harwath, D., Karlinsky, L., Kuehne, H., & Glass, J. R., (2022), Contrastive audio-visual masked autoencoder, *International Conference on Learning Representations (ICLR)* (cit. on p. 139).

Goodfellow, I., Bengio, Y., & Courville, A., (2016), *Deep learning*, MIT press, (cit. on pp. 21, 37).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y., (2014), Generative adversarial nets, *Advances in Neural Information Processing Systems (NeurIPS)*, 2672–2680 (cit. on pp. 40, 42, 60, 76, 82, 105, 113, 169, 170).

Gou, J., Yu, B., Maybank, S. J., & Tao, D., (2021), Knowledge distillation: a survey, *International Journal of Computer Vision*, *129*, 1789–1819 (cit. on p. 165).

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., & He, K., (2017), Accurate, large minibatch sgd: training imagenet in 1 hour, *arXiv preprint arXiv:1706.02677* (cit. on p. 151).

Graves, A., Mohamed, A.-r., & Hinton, G., (2013), Speech recognition with deep recurrent neural networks, *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6645–6649 (cit. on p. 55).

Griffin, D., & Lim, J., (1984), Signal estimation from modified short-time fourier transform, *IEEE Transactions on acoustics, speech, and signal processing*, *32 2*, 236–243 (cit. on p. 163).

Griffiths, P. E., (2002), Basic emotions, complex emotions, machiavellian emotions.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al., (2020), Bootstrap your own latent-a new approach to self-supervised learning, *Advances in neural information processing systems*, *33*, 21271–21284 (cit. on p. 70).

Gunes, H., & Schuller, B., (2013), Categorical and dimensional affect analysis in continuous input: current trends and future directions, *Image and Vision Computing*, *31 2*, 120–136 (cit. on p. 18).

Härkönen, E., Hertzmann, A., Lehtinen, J., & Paris, S., (2020), GANSpace: discovering interpretable GAN controls, *Advances in Neural Information Processing Systems (NeurIPS)*, 9841–9850 (cit. on p. 82).

Hayward, K., (2000), *Experimental phonetics*, Harlow, UK: Pearson.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R., (2022), Masked autoencoders are scalable vision learners, *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009 (cit. on pp. 70, 71, 73, 139, 140, 145, 146, 151).

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R., (2020), Momentum contrast for unsupervised visual representation learning, *Conference on computer vision and pattern recognition (IEEE/CVF)*, 9729–9738 (cit. on p. 70).

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., & Lerchner, A., (2018), Towards a definition of disentangled representations, *arXiv preprint arXiv:1812.02230* (cit. on pp. 39, 74, 82, 86).

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., & Lerchner, A., (2017a), Beta-VAE: learning basic visual concepts with a constrained variational framework, *International Conference on Learning Representations (ICLR)* (cit. on pp. 40, 53, 76, 77, 81, 105, 170).

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., & Lerchner, A., (2017b), Beta-vae: learning basic visual concepts with a constrained variational framework, *International conference on learning representations (ICLR)* (cit. on pp. 52, 53).

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K., (1995), Acoustic characteristics of American English vowels, *The Journal of the Acoustical Society of America*, *97 5*, 3099–3111 (cit. on p. 95).

Hinton, G. E., (2002), Training products of experts by minimizing contrastive divergence, *Neural computation*, *14 8*, 1771–1800 (cit. on pp. 63, 170).

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., & Bengio, Y., (2018), Learning deep representations by mutual information estimation and maximization, *arXiv preprint arXiv:1808.06670* (cit. on pp. 36, 37).

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J., (2013), Stochastic variational inference, *Journal of Machine Learning Research* (cit. on p. 49).

Hoffman, M. D., & Johnson, M. J., (2016), ELBO surgery: yet another way to carve up the variational evidence lower bound, *Workshop in Advances in Approximate Bayesian Inference, NIPS*, *1 2* (cit. on pp. 53, 54).

Honnet, P.-E., Lazaridis, A., Garner, P. N., & Yamagishi, J., (2017), *The SIWIS French speech synthesis database - Design and recording of a high quality French database for speech synthesis* (tech. rep.), Idiap, (cit. on p. 98).

Hori, C., Alamri, H., Wang, J., Wichern, G., Hori, T., Cherian, A., Marks, T. K., Cartillier, V., Lopes, R. G., Das, A., et al., (2019), End-to-end audio visual scene-aware dialog using multimodal attention-based video features, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2352–2356 (cit. on p. 105).

Hosoya, H., (2018), *Group-based learning of disentangled representations with generalizability for novel contents* [arXiv preprint arXiv:1809.02383], (cit. on p. 81).

Hotelling, H., (1957), The relations of the newer multivariate statistical methods to factor analysis, *British Journal of Statistical Psychology*, *10 2*, 69–79 (cit. on p. 87).

Hou, X., Sun, K., Shen, L., & Qiu, G., (2019), Improving variational autoencoder with deep feature consistent and generative adversarial training, *Neurocomputing*, *341*, 183–194 (cit. on p. 114).

Hsu, C.-C., Hwang, H.-T., Wu, Y.-C., Tsao, Y., & Wang, H.-M., (2016), Voice conversion from non-parallel corpora using variational auto-encoder, *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 1–6 (cit. on p. 80).

Hsu, W.-N., & Glass, J., (2018), *Disentangling by partitioning: a representation learning framework for multimodal sensory data* [arXiv preprint arXiv:1805.11264], (cit. on pp. 62, 65).

Hsu, W.-N., Zhang, Y., & Glass, J., (2017a), Learning latent representations for speech generation and transformation, *Interspeech*, 1273–1277 (cit. on pp. 52, 80, 82, 88, 97).

Hsu, W.-N., Zhang, Y., & Glass, J., (2017b), Unsupervised learning of disentangled and interpretable representations from sequential data, *Advances in Neural Information Processing Systems (NeurIPS)*, 1878–1889 (cit. on p. 80).

Huang, W., Yi, M., & Zhao, X., (2021), Towards the generalization of contrastive self-supervised learning, *arXiv preprint arXiv:2111.00743* (cit. on p. 36).

Huang, Z., Jin, X., Lu, C., Hou, Q., Cheng, M.-M., Fu, D., Shen, X., & Feng, J., (2022), Contrastive masked autoencoders are stronger vision learners, *arXiv preprint arXiv:2207.13532* (cit. on p. 139).

Ito, K., & Johnson, L., (2017), The LJ speech dataset, (cit. on p. 99).

Izard, C. E., (1971), The face of emotion (cit. on p. 13).

Jacquelin, M., Garnier, M., Girin, L., Vincent, R., & Perrotin, O., (2023), Exploring the multidimensional representation of individual speech acoustic parameters extracted by deep unsupervised models, *12th ISCA Speech Synthesis Workshop (SSW2023)*, 240–241 (cit. on p. 92).

Jadoul, Y., Thompson, B., & de Boer, B., (2018), Introducing Parselmouth: a Python interface to Praat, *Journal of Phonetics*, *71*, 1–15 (cit. on p. 96).

Jahanian, A., Chai, L., & Isola, P., (2019), On the "steerability" of generative adversarial networks, *International Conference on Learning Representations (ICLR)* (cit. on p. 82).

Jang, E., Gu, S., & Poole, B., (2016), Categorical reparameterization with gumbel-softmax, *arXiv preprint arXiv:1611.01144* (cit. on pp. 66, 68).

Jayaram, V., & Thickstun, J., (2020), Source separation with deep generative priors, *International Conference on Machine Learning (ICML)*, 4724–4735 (cit. on p. 76).

Jegorova, M., Petridis, S., & Pantic, M., (2023), SS-VAERR: self-supervised apparent emotional reaction recognition from video, *International Conference on Automatic Face and Gesture Recognition (FG)*, 1–8 (cit. on p. 138).

Jekel, C. F., & Venter, G., (2019), *PWLF: a Python library for fitting 1D continuous piecewise linear functions* [URL: https://github.com/cjekel/piecewise_linear_fit_py/raw/master/paper/pwlf_Jekel_Venter_v2.pdf], (cit. on p. 87).

Jiang, J., Xia, G. G., Carlton, D. B., Anderson, C. N., & Miyakawa, R. H., (2020), Transformer VAE: a hierarchical model for structure-aware and interpretable music

representation learning, *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 516–520 (cit. on p. 56).

Jobin, A., Ienca, M., & Vayena, E., (2019), The global landscape of ai ethics guidelines, *Nature machine intelligence*, *1 9*, 389–399 (cit. on p. 161).

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K., (1999), An introduction to variational methods for graphical models, *Machine learning*, *37*, 183–233 (cit. on p. 48).

Juvela, L., Bollepalli, B., Tsiaras, V., & Alku, P., (2019), GlotNet—a raw waveform model for the glottal excitation in statistical parametric speech synthesis, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *276*, 1019–1030 (cit. on p. 80).

Kameoka, H., Li, L., Inoue, S., & Makino, S., (2019), Supervised determined source separation with multichannel variational autoencoder, *Neural Computation*, *31 9*, 1891–1914 (cit. on p. 76).

Kawahara, H., (2006), STRAIGHT, exploitation of the other aspect of VOCODER: perceptually isomorphic decomposition of speech sounds, *Acoustical Science and Technology*, *276*, 349–353 (cit. on p. 80).

Kendall, M., (1957), *A course in multivariate analysis*, Charles Griffin, (cit. on p. 87).

Khodai-Joopari, M., & Clermont, F., (2002), A comparative study of empirical formulae for estimating vowel-formant bandwidths, *Australian International Conference on Speech, Science, and Technology*, 130–135 (cit. on p. 178).

Kim, D., & Song, B. C., (2022), Optimal transport-based identity matching for identity-invariant facial expression recognition, *Advances in Neural Information Processing Systems* (cit. on p. 134).

Kim, H., & Mnih, A., (2018), Disentangling by factorising, *International Conference on Machine Learning (ICML)*, 2649–2658 (cit. on pp. 52–54, 76, 77, 81, 105, 170).

Kim, J. W., Salamon, J., Li, P., & Bello, J. P., (2018), Crepe: a convolutional representation for pitch estimation, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 161–165 (cit. on pp. 96, 100, 122, 124).

Kim, T., & Adalı, T., (2003), Approximation by fully complex multilayer perceptrons, *Neural computation*, *15 7*, 1641–1666 (cit. on p. 29).

Kim, Y., Wiseman, S., & Rush, A. M., (2018), A tutorial on deep latent variable models of natural language, *arXiv preprint arXiv:1812.06834* (cit. on pp. 43, 49).

Kingma, D. P., & Ba, J., (2015), Adam: A method for stochastic optimization, *International Conference on Learning Representations (ICLR)* (cit. on pp. 116, 177).

Kingma, D. P., & Welling, M., (2014), Auto-encoding variational Bayes, *International Conference on Learning Representations (ICLR)* (cit. on pp. 24, 40, 42, 47, 49, 76, 83, 84, 105, 112, 169, 171).

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

Klys, J., Snell, J., & Zemel, R., (2018), Learning latent subspaces in variational autoencoders, *Advances in neural information processing systems*, *31* (cit. on pp. 105, 170).

Kong, J., Kim, J., & Bae, J., (2020), Hifi-gan: generative adversarial networks for efficient and high fidelity speech synthesis, *Advances in Neural Information Processing Systems*, *33*, 17022–17033 (cit. on p. 163).

Kong, L., Ma, M. Q., Chen, G., Xing, E. P., Chi, Y., Morency, L.-P., & Zhang, K., (2023), Understanding masked autoencoders via hierarchical latent variable models, *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7918–7928 (cit. on p. 72).

Laroche, J., (2002), Time and pitch scale modification of audio signals. *In Applications of digital signal processing to audio and acoustics* (pp. 279–309), Springer, (cit. on p. 80).

Laroche, J., & Dolson, M., (1999), Improved phase vocoder time-scale modification of audio, *IEEE Transactions on Speech and Audio processing*, *7*, 323–332 (cit. on p. 80).

Laroche, J., Stylianou, Y., & Moulines, E., (1993), HNS: speech modification based on a harmonic+noise model, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 550–553 (cit. on p. 80).

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., & Winther, O., (2016), Autoencoding beyond pixels using a learned similarity metric, *International conference on machine learning (ICML)*, 1558–1566 (cit. on p. 114).

Larsson, G., Maire, M., & Shakhnarovich, G., (2016), Learning representations for automatic colorization, *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, 577–593 (cit. on p. 70).

Latif, S., Ali, H. S., Usama, M., Rana, R., Schuller, B., & Qadir, J., (2022), AI-based emotion recognition: promise, peril, and prescriptions for prosocial path, *arXiv preprint arXiv:2211.07290* (cit. on pp. 17, 161).

Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., & Schuller, B. W., (2021), Survey of deep representation learning for speech emotion recognition, *IEEE Transactions on Affective Computing* (cit. on pp. 14, 28, 29).

Lazarus, A. A., (1976), Multimodal therapy, *Handbook of Psychotherapy Integration*, 105 (cit. on p. 104).

Le Moine, C., Obin, N., & Roebel, A., (2021), Towards end-to-end f0 voice conversion based on dual-gan with convolutional wavelet kernels, *European Signal Processing Conference (EUSIPCO)*, 36–40 (cit. on p. 76).

Le Roux, J., Wisdom, S., Erdogan, H., & Hershey, J. R., (2019), Sdr–half-baked or well done?, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 626–630 (cit. on p. 118).

LeCun, Y., et al., (1989), Generalization and network design strategies, *Connectionism in perspective*, *19143-155*, 18 (cit. on p. 16).

LeCun, Y., Bengio, Y., et al., (1995), Convolutional networks for images, speech, and time series, *The handbook of brain theory and neural networks*, *336110*, 1995 (cit. on p. 55).

Lee, J., Choi, H.-S., Jeon, C.-B., Koo, J., & Lee, K., (2019), Adversarially trained end-to-end Korean singing voice synthesis system, *Interspeech*, 2588–2592 (cit. on p. 80).

Leglaive, S., Alameda-Pineda, X., Girin, L., & Horaud, R., (2020), A recurrent variational autoencoder for speech enhancement, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 371–375 (cit. on p. 79).

Leglaive, S., Girin, L., & Horaud, R., (2018), A variance modeling framework based on variational autoencoders for speech enhancement, *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6 (cit. on pp. 76, 79, 83, 90, 177).

Leglaive, S., Girin, L., & Horaud, R., (2019a), Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 101–105 (cit. on p. 90).

Leglaive, S., Girin, L., & Horaud, R., (2019b), Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization,

ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 101–105.

Leidal, K., Harwath, D., & Glass, J., (2017), Learning modality-invariant representations for speech and images, *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 424–429.

Li, J., Niu, L., & Zhang, L., (2022a), From representation to reasoning: towards both evidence and commonsense reasoning for video question-answering, *Conference on Computer Vision and Pattern Recognition (IEEE/CVF)*, 21273–21282 (cit. on p. 163).

Li, J., et al., (2022b), Recent advances in end-to-end automatic speech recognition, *APSIPA Transactions on Signal and Information Processing*, *11*.

Li, T., Chang, H., Mishra, S. K., Zhang, H., Katabi, D., & Krishnan, D., (2022), Mage: masked generative encoder to unify representation learning and image synthesis, *arXiv preprint arXiv:2211.09117* (cit. on p. 140).

Li, Y., & Mandt, S., (2018), *Disentangled sequential autoencoder* [arXiv preprint arXiv:1803.02991], (cit. on pp. 58, 109, 117, 127, 130, 133).

Liang, T., Glossner, J., Wang, L., Shi, S., & Zhang, X., (2021), Pruning and quantization for deep neural network acceleration: a survey, *Neurocomputing*, *461*, 370–403 (cit. on p. 165).

Liu, S., Zhang, L., Yang, X., Su, H., & Zhu, J., (2021), Query2label: a simple transformer way to multi-label classification, *arXiv preprint arXiv:2107.10834* (cit. on p. 147).

Liu, S., Mallol-Ragolta, A., Parada-Cabaleiro, E., Qian, K., Jing, X., Kathan, A., Hu, B., & Schuller, B. W., (2022), Audio self-supervised learning: a survey, *Patterns*, *312*, 100616 (cit. on p. 138).

Liu, X., Sanchez, P., Thermos, S., O'Neil, A. Q., & Tsaftaris, S. A., (2022), Learning disentangled representations in the imaging domain, *Medical Image Analysis*, *80*, 102516 (cit. on pp. 39, 52).

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., & Tang, J., (2021), Self-supervised learning: generative or contrastive, *IEEE transactions on knowledge and data engineering*, *351*, 857–876 (cit. on p. 69).

Livingstone, S. R., & Russo, F. A., (2018), The ryerson audio-visual database of emotional speech and song (ravdess): a dynamic, multimodal set of facial and vocal expressions in north american english, *PloS one*, *135*, e0196391 (cit. on pp. 130, 150, 197).

Lo, C.-C., Fu, S.-W., Huang, W.-C., Wang, X., Yamagishi, J., Tsao, Y., & Wang, H.-M., (2019), Mosnet: deep learning based objective assessment for voice conversion, *Conference of the International Speech Communication Association*, 1541–1545 (cit. on p. 118).

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., & Bachem, O., (2019), Challenging common assumptions in the unsupervised learning of disentangled representations, *International Conference on Machine Learning (ICML)*, 4114–4124 (cit. on pp. 39, 40, 55, 77, 81, 85, 105).

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., & Bachem, O., (2020), A sober look at the unsupervised learning of disentangled representations and their evaluation, *Journal of Machine Learning Research*, *21*, 1–62 (cit. on pp. 77, 81).

Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., & Tschannen, M., (2020), Weakly-supervised disentanglement without compromises, *International Conference on Machine Learning (ICML)*, 6348–6359 (cit. on pp. 81, 105, 170).

Locatello, F., Tschannen, M., Bauer, S., Rätsch, G., Schölkopf, B., & Bachem, O., (2020), Disentangling factors of variation using few labels, *International Conference on Learning Representations (ICLR)* (cit. on p. 81).

Loshchilov, I., & Hutter, F., (2017), Decoupled weight decay regularization, *International Conference on Learning Representations (ICLR)* (cit. on p. 151).

Lundh, F., (1999), An introduction to tkinter, pythonware.com/library/tkinter/introduction/index.htm (cit. on p. 201).

MacDonald, E. N., Purcell, D. W., & Munhall, K. G., (2011), Probing the independence of formant control using altered auditory feedback, *The Journal of the Acoustical Society of America*, *129 2*, 955–965 (cit. on p. 77).

Makhoul, J., (1975), Linear prediction: a tutorial review, *Proceedings of the IEEE*, *63 4*, 561–580 (cit. on p. 80).

Makhzani, A., Shlens, J., Jaitly, N., & Goodfellow, I. J., (2015), Adversarial autoencoders, *ArXiv*, *abs/1511.05644* (cit. on pp. 53, 85).

Mano, L. Y., Faiçal, B. S., Nakamura, L. H., Gomes, P. H., Libralon, G. L., Meneguete, R. I., Geraldo Filho, P. R., Giancristofaro, G. T., Pessin, G., Krishnamachari, B., et al., (2016), Exploiting iot technologies for enhancing health smart homes through patient identification and emotion recognition, *Computer Communications*, *89*, 178–190 (cit. on pp. 14, 167).

Markel, J. D., & Gray, A. J., (1976), *Linear prediction of speech*, Springer-Verlag, (cit. on p. 80).

Matthey, L., Higgins, I., Hassabis, D., & Lerchner, A., (2017), Dsprites: disentanglement testing sprites dataset, (cit. on p. 52).

Mauch, M., & Dixon, S., (2014), PYIN: a fundamental frequency estimator using probabilistic threshold distributions, *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 659–663 (cit. on p. 100).

McAulay, R., & Quatieri, T., (1986), Speech analysis/synthesis based on a sinusoidal representation, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *34 4*, 744–754 (cit. on p. 80).

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O., (2015), Librosa: audio and music signal analysis in python, *Proceedings of the 14th python in science conference*, *8*, 18–25 (cit. on p. 100).

Mehrabian, A., (2017), *Nonverbal communication*, Routledge, (cit. on p. 21).

Mehrabian, A., & Ferris, S. R., (1967), Inference of attitudes from nonverbal communication in two channels, *Journal of consulting psychology*, *31 3*, 248 (cit. on p. 21).

Mehrabian, A., & Wiener, M., (1967), Decoding of inconsistent communications, *Journal of personality and social psychology*, *6 1*, 109 (cit. on p. 21).

Milner, R., Jalal, M. A., Ng, R. W., & Hain, T., (2019), A cross-corpus study on speech emotion recognition, *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 304–311 (cit. on p. 164).

Mittag, G., & Möller, S., (2020), Deep learning based assessment of synthetic speech naturalness, *Interspeech*, 1748–1752 (cit. on p. 96).

Mnih, A., & Gregor, K., (2014), Neural variational inference and learning in belief networks, *International Conference on Machine Learning (ICML)*, 1791–1799 (cit. on pp. 66, 68).

Mnih, A., & Rezende, D., (2016), Variational inference for monte carlo objectives, *International Conference on Machine Learning*, 2188–2196 (cit. on p. 68).

Moine, C. L., & Obin, N., (2020), Att-hack: an expressive speech database with social attitudes, *arXiv preprint arXiv:2004.04410* (cit. on p. 164).

Mollahosseini, A., Hasani, B., & Mahoor, M. H., (2017), Affectnet: a database for facial expression, valence, and arousal computing in the wild, *IEEE Transactions on Affective Computing*, *10 1*, 18–31 (cit. on p. 131).

Morise, M., Yokomori, F., & Ozawa, K., (2016), World: a vocoder-based high-quality speech synthesis system for real-time applications, *IEICE Transactions on Information and Systems*, *99 7*, 1877–1884 (cit. on pp. 80, 97).

Morrison, M., (2020), *Torchcrepe* [URL: https://github.com/maxrmorrison/torchcrepe], (cit. on p. 100).

Morrison, M., Jin, Z., Bryan, N. J., Caceres, J.-P., & Pardo, B., (2021), *Neural pitch-shifting and time-stretching with controllable LPCNet* [arXiv preprint arXiv:2110.02360], (cit. on p. 80).

Moulines, E., & Charpentier, F., (1990), Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones, *Speech Communication*, *9 5-6*, 453–467 (cit. on pp. 80, 97).

Muhammod, R., Ahmed, S., Md Farid, D., Shatabda, S., Sharma, A., & Dehzangi, A., (2019), Pyfeat: a python-based effective feature generation tool for dna, rna and protein sequences, *Bioinformatics*, *35 19*, 3831–3833 (cit. on pp. 121, 124).

Murphy, K. P., (2012), *Machine learning: a probabilistic perspective*, MIT press, (cit. on pp. 20, 44).

Neal, R. M., & Hinton, G. E., (1998), A view of the EM algorithm that justifies incremental, sparse, and other variants. *In Learning in graphical models* (pp. 355–368), Springer, (cit. on p. 48).

Ngiam, J., Chen, Z., Koh, P. W., & Ng, A. Y., (2011), Learning deep energy models, *International conference on machine learning (ICML)*, 1105–1112 (cit. on p. 42).

Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., & Anbarjafari, G., (2017), Audio-visual emotion recognition in video clips, *IEEE Transactions on Affective Computing*, *10 1*, 60–75 (cit. on p. 105).

Noroozi, M., & Favaro, P., (2016), Unsupervised learning of visual representations by solving jigsaw puzzles, *Computer Vision (ECCV)*, 69–84 (cit. on p. 139).

Obin, N., (2023), *From signal representation to representation learning: structured modeling of speech signals* [Doctoral dissertation, Sorbonne Université], (cit. on pp. 20, 29).

on Aging, N. I., on Aging (US), N. A. C., & of Health, U. S. D., (1980), *Our future selves: a research plan toward understanding aging, of the department of health, education and welfare*, US Department of Health, Education; Welfare, Public Health Service . . ., (cit. on p. 161).

Ong, D., Su, J., Chen, B., Luu, A. T., Narendranath, A., Li, Y., Sun, S., Lin, Y., & Wang, H., (2022), Is discourse role important for emotion recognition in conversation?, *AAAI Conference on Artificial Intelligence*, *36 10*, 11121–11129 (cit. on p. 74).

Oord, A. v. d., Li, Y., & Vinyals, O., (2018), Representation learning with contrastive predictive coding, *arXiv preprint arXiv:1807.03748* (cit. on p. 69).

Ouali, Y., (2023), *Learning with limited labeled data. (apprentissage avec peu de données étiquetées)* [Doctoral dissertation, University of Paris-Saclay, France], https://tel.archives-ouvertes.fr/tel-04127195 (cit. on pp. 17, 20).

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S., (2015), Librispeech: an asr corpus based on public domain audio books, *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5206–5210 (cit. on p. 131).

Pantic, M., Sebe, N., Cohn, J. F., & Huang, T., (2005), Affective multimodal human-computer interaction, *ACM international conference on Multimedia*, 669–676 (cit. on pp. 15, 167).

Parekh, S., Ozerov, A., Essid, S., Duong, N. Q., Pérez, P., & Richard, G., (2019), Identify, locate and separate: audio-visual object extraction in large video collections using weak supervision, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 268–272.

Pearson, K., (1901), On lines and planes of closest fit to systems of points in space, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2 11*, 559–572 (cit. on pp. 44, 86).

Peng, X., Huang, Z., Sun, X., & Saenko, K., (2019), Domain agnostic learning with disentangled representations, *International Conference on Machine Learning (ICML)*, 5102–5112 (cit. on p. 52).

Pepino, L., Riera, P., & Ferrer, L., (2021), Emotion recognition from speech using wav2vec 2.0 embeddings, *Interspeech*, 3400–3404 (cit. on pp. 131, 138).

Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., & Pantic, M., (2018), End-to-end audiovisual speech recognition, *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 6548–6552 (cit. on p. 105).

Pham, L., Vu, T. H., & Tran, T. A., (2021), Facial expression recognition using residual masking network, *International Conference on Pattern Recognition (ICPR)*, 4513–4519 (cit. on p. 126).

Piaget, J., (2000), Piaget's theory of cognitive development, *Childhood cognitive development: The essential readings*, *2*, 33–47 (cit. on p. 35).

Picard, R. W., Vyzas, E., & Healey, J., (2001), Toward machine emotional intelligence: analysis of affective physiological state, *IEEE transactions on pattern analysis and machine intelligence*, *23 10*, 1175–1191 (cit. on p. 14).

Picard, R., (1997), Affective computing cambridge, *MA: MIT Press [Google Scholar]* (cit. on pp. 13, 167).

Pihlgren, G. G., Sandin, F., & Liwicki, M., (2020), Improving image autoencoder embeddings with perceptual loss, *International Joint Conference on Neural Networks (IJCNN)*, 1–7 (cit. on p. 114).

Pirker, G., Wohlmayr, M., Petrik, S., & Pernkopf, F., (2011), A pitch tracking corpus with evaluation on multipitch tracking scenario, *Interspeech*, 1509–1512 (cit. on p. 100).

Plumerault, A., Borgne, H. L., & Hudelot, C., (2020), Controlling generative models with continuous factors of variations, *International Conference on Learning Representations (ICLR)* (cit. on p. 82).

Prenger, R., Valle, R., & Catanzaro, B., (2019), Waveglow: a flow-based generative network for speech synthesis, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3617–3621 (cit. on p. 91).

Qian, K., Zhang, Y., Chang, S., Hasegawa-Johnson, M., & Cox, D., (2020), Unsupervised speech decomposition via triple information bottleneck, *International Conference on Machine Learning (ICML)*, 7836–7846 (cit. on p. 81).

Qu, L., Li, L., Zhang, Y., & Hu, J., (2009), PPCA-based missing data imputation for traffic flow volume: a systematical approach, *IEEE Transactions on intelligent transportation systems*, *10 3*, 512–522 (cit. on p. 45).

Rabiner, L., Cheng, M., Rosenberg, A., & McGonegal, C., (1976), A comparative performance study of several pitch detection algorithms, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *24 5*, 399–418 (cit. on p. 100).

Ramachandram, D., & Taylor, G. W., (2017), Deep multimodal learning: a survey on recent advances and trends, *IEEE signal processing magazine*, *34 6*, 96–108 (cit. on pp. 104, 138).

Razavi, A., Van den Oord, A., & Vinyals, O., (2019), Generating diverse high-fidelity images with vq-vae-2, *Advances in neural information processing systems*, *32* (cit. on pp. 66, 114).

Rezende, D. J., Mohamed, S., & Wierstra, D., (2014), Stochastic backpropagation and approximate inference in deep generative models, *International Conference on*

Machine Learning (ICML), 1278–1286 (cit. on pp. 24, 40, 47, 49, 76, 83, 84, 105, 112, 169, 171).

Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., & Zelnik-Manor, L., (2021), Asymmetric loss for multi-label classification, *IEEE/CVF International Conference on Computer Vision (ICCV)*, 82–91 (cit. on p. 147).

Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P., (2001), Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, *2*, 749–752 (cit. on p. 118).

Roth, J., Chaudhuri, S., Klejch, O., Marvin, R., Gallagher, A., Kaver, L., Ramaswamy, S., Stopczynski, A., Schmid, C., Xi, Z., et al., (2020), Ava active speaker: an audio-visual dataset for active speaker detection, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4492–4496 (cit. on p. 105).

Rummel, R. J., (1988), *Applied factor analysis*, Northwestern University Press, (cit. on p. 46).

Ruthotto, L., & Haber, E., (2021), An introduction to deep generative modeling, *GAMM-Mitteilungen*, *44 2*, e202100008 (cit. on p. 42).

Sadeghi, M., Leglaive, S., Alameda-Pineda, X., Girin, L., & Horaud, R., (2020), Audio-visual speech enhancement using conditional variational auto-encoders, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 1788–1800 (cit. on p. 105).

Sadok, S., Leglaive, S., Girin, L., Alameda-Pineda, X., & Séguier, R., (2023a), Learning and controlling the source-filter representation of speech with a variational autoencoder, *Speech Communication*, *148*, 53–65 (cit. on pp. 105, 170).

Sadok, S., Leglaive, S., Girin, L., Alameda-Pineda, X., & Séguier, R., (2023b), A multimodal dynamical variational autoencoder for audiovisual speech representation learning, *arXiv preprint arXiv:2305.03582* (cit. on pp. 143, 153, 154).

Sadok, S., Leglaive, S., & Séguier, R., (2023), A vector quantized masked autoencoder for speech emotion recognition, *IEEE ICASSP 2023 Workshop on Self-Supervision in Audio, Speech and Beyond (SASB)* (cit. on pp. 140, 149, 198).

Salakhutdinov, R., (2015), Learning deep generative models, *Annual Review of Statistics and Its Application*, *2*, 361–385 (cit. on p. 42).

Scheutz, M., (2012), The affect dilemma for artificial agents: should we develop affective artificial agents?, *IEEE Transactions on Affective Computing*, *3 4*, 424–433 (cit. on p. 14).

Schneider, S., Baevski, A., Collobert, R., & Auli, M., (2019), Wav2vec: unsupervised pre-training for speech recognition, *Interspeech*, 3465–3469 (cit. on pp. 131, 133).

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y., (2021), Toward causal representation learning, *Proceedings of the IEEE*, *109 5*, 612–634 (cit. on p. 38).

Schoneveld, L., Othmani, A., & Abdelkawy, H., (2021), Leveraging recent advances in deep learning for audio-visual emotion recognition, *Pattern Recognition Letters*, *146*, 1–7 (cit. on pp. 105, 138).

Schroder, M., Devillers, L., Karpouzis, K., Martin, J.-C., Pelachaud, C., Peter, C., Pirker, H., Schuller, B., Tao, J., & Wilson, I., (2007), What should a generic emotion markup language be able to represent?, *Affective Computing and Intelligent Interaction: Second International Conference, ACII 2007 Lisbon, Portugal, September 12-14, 2007 Proceedings 2*, 440–451 (cit. on p. 18).

Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., & Rigoll, G., (2010), Cross-corpus acoustic emotion recognition: variances and strategies, *IEEE Transactions on Affective Computing*, *1 2*, 119–131 (cit. on p. 164).

Schuller, B. W., (2018), Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends, *Communications of the ACM*, *61 5*, 90–99 (cit. on p. 14).

Sebe, N., Cohen, I., & Huang, T. S., (2005), Multimodal emotion recognition. *In Handbook of pattern recognition and computer vision* (pp. 387–409), World Scientific, (cit. on pp. 17, 18, 22, 169).

Sekiguchi, K., Bando, Y., Nugraha, A. A., Yoshii, K., & Kawahara, T., (2019), Semi-supervised multichannel speech enhancement with a deep speech prior, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27 12*, 2197–2212 (cit. on p. 90).

Sekiguchi, K., Bando, Y., Yoshii, K., & Kawahara, T., (2018), Bayesian multichannel speech enhancement with a deep speech prior, *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 1233–1239.

Serra, X., & Smith, J., (1990), Spectral modeling synthesis: a sound analysis/synthesis system based on a deterministic plus stochastic decomposition, *Computer Music Journal*, *14 4*, 12–24 (cit. on p. 80).

Sethu, V., Provost, E. M., Epps, J., Busso, C., Cummins, N., & Narayanan, S., (2019), The ambiguous world of emotion representation, *arXiv preprint arXiv:1909.00360* (cit. on p. 18).

Shannon, C. E., (1948), A mathematical theory of communication, *The Bell system technical journal*, *27*, 379–423 (cit. on p. 17).

Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., & Cui, P., (2021), Towards out-of-distribution generalization: a survey, *arXiv preprint arXiv:2108.13624* (cit. on p. 19).

Shi, Y., Paige, B., Torr, P., et al., (2019), Variational mixture-of-experts autoencoders for multi-modal deep generative models, *Advances in Neural Information Processing Systems*, *32* (cit. on pp. 61, 62, 64, 170).

Shu, R., Chen, Y., Kumar, A., Ermon, S., & Poole, B., (2020), Weakly supervised disentanglement with guarantees, *International Conference on Learning Representations (ICLR)* (cit. on p. 81).

Shwartz-Ziv, R., & LeCun, Y., (2023), To compress or not to compress–self-supervised learning and information theory: a review, *arXiv preprint arXiv:2304.09355* (cit. on p. 36).

Shwartz-Ziv, R., Painsky, A., & Tishby, N., (2018), Representation compression and generalization in deep neural networks, (cit. on p. 33).

Shwartz-Ziv, R., & Tishby, N., (2017), Opening the black box of deep neural networks via information, *arXiv preprint arXiv:1703.00810* (cit. on pp. 31–33).

Sims, J. M., (2010), A brief review of the belmont report, *Dimensions of critical care nursing*, *29*, 173–174 (cit. on p. 161).

Sorrenson, P., Rother, C., & Köthe, U., (2020), Disentanglement by nonlinear ICA with general incompressible-flow networks (GIN), *International Conference on Learning Representations (ICLR)* (cit. on p. 81).

Sridharan, K., & Kakade, S. M., (2008), An information theoretic framework for multi-view learning (cit. on p. 36).

Strubell, E., Ganesh, A., & McCallum, A., (2019), Energy and policy considerations for deep learning in nlp, *arXiv preprint arXiv:1906.02243* (cit. on p. 165).

Sugiyama, M., Suzuki, T., & Kanamori, T., (2012), Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation, *Annals of the Institute of Statistical Mathematics*, *64*, 1009–1044 (cit. on p. 54).

Sun, L., Lian, Z., Liu, B., & Tao, J., (2023), MAE-DFER: efficient masked autoencoder for self-supervised dynamic facial expression recognition, *arXiv preprint arXiv:2307.02227* (cit. on pp. 153, 154).

Sutter, T., Daunhawer, I., & Vogt, J., (2020), Multimodal generative learning utilizing jensen-shannon-divergence, *Advances in Neural Information Processing Systems*, *33*, 6100–6110 (cit. on p. 64).

Sutter, T., Daunhawer, I., & Vogt, J. E., (2021), Generalized multimodal elbo, *International Conference on Learning Representations (ICLR)* (cit. on p. 170).

Suzuki, M., & Matsuo, Y., (2022), A survey of multimodal deep generative models, *Advanced Robotics*, *36 5-6*, 261–278 (cit. on pp. 60, 62, 170).

Suzuki, M., Nakayama, K., & Matsuo, Y., (2016), *Joint multimodal learning with deep generative models* [arXiv preprint arXiv:1611.01891], (cit. on pp. 62, 63, 127).

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J., (2010), A short-time objective intelligibility measure for time-frequency weighted noisy speech, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4214–4217 (cit. on p. 118).

Tamburrini, G., (2022), The ai carbon footprint and responsibilities of ai scientists, *Philosophies*, *7 1*, https://doi.org/10.3390/philosophies7010004 (cit. on pp. 165, 166).

Tappert, C., Martony, J., & Fant, G., (1963), Spectrum envelopes for synthetic vowels, *Speech Transmission Laboratory Quarterly Progress Status Report*, *4*, 2–6 (cit. on p. 178).

Tharwat, A., (2016), Principal component analysis-a tutorial, *International Journal of Applied Pattern Recognition*, *3 3*, 197–240 (cit. on p. 44).

Thickstun, J., (2020), Discrete vae's, (cit. on p. 66).

Thiemann, J., Ito, N., & Vincent, E., (2013), DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments, *International Congress on Acoustics (ICA)*, 1–6 (cit. on p. 100).

Tian, Y., Shi, J., Li, B., Duan, Z., & Xu, C., (2018), Audio-visual event localization in unconstrained videos, *European Conference on Computer Vision (ECCV)*, 247–263.

Tian, Y., Xie, L., Fang, J., Shi, M., Peng, J., Zhang, X., Jiao, J., Tian, Q., & Ye, Q., (2022), Beyond masking: demystifying token-based pre-training for vision transformers, *arXiv preprint arXiv:2203.14313* (cit. on p. 72).

Tiao, L. C., (2018), Density Ratio Estimation for KL Divergence Minimization between Implicit Distributions, *tiao.io*, https://tiao.io/post/density-ratio-estimation-for-kl-divergence-minimization-between-implicit-distributions/ (cit. on p. 54).

Tipping, M. E., & Bishop, C. M., (1999), Mixtures of probabilistic principal component analyzers, *Neural computation*, *11 2*, 443–482 (cit. on p. 45).

Tishby, N., & Zaslavsky, N., (2015), Deep learning and the information bottleneck principle, *IEEE information theory workshop (itw)*, 1–5 (cit. on pp. 31, 33).

Tong, Z., Song, Y., Wang, J., & Wang, L., (2022), VideoMAE: masked autoencoders are data-efficient learners for self-supervised video pre-training, *Advances in neural information processing systems*, *35*, 10078–10093 (cit. on p. 139).

Touvron, H., Cord, M., El-Nouby, A., Bojanowski, P., Joulin, A., Synnaeve, G., & Jégou, H., (2021), Augmenting convolutional networks with attention-based aggregation, *arXiv preprint arXiv:2112.13692* (cit. on p. 145).

Tran, H., Brelet, L., Falih, I., Goblet, X., & Nguifo, E. M., (2022), L'ambiguite dans la representation des emotions: etat de l'art des bases de donnees multimodales, *Conference Extraction et Gestion de Connaissances* (cit. on pp. 17, 168).

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., & Salakhutdinov, R., (2019), Multimodal transformer for unaligned multimodal language sequences, *Conference. Association for Computational Linguistics. Meeting*, *2019*, 6558 (cit. on pp. 131, 133, 138, 153).

Tsai, Y.-H. H., Liang, P. P., Zadeh, A., Morency, L.-P., & Salakhutdinov, R., (2018), Learning factorized multimodal representations, *International Conference on Learning Representations (ICLR)* (cit. on p. 37).

Vahdat, A., & Kautz, J., (2020), NVAE: a deep hierarchical variational autoencoder, *Advances in neural information processing systems*, *33*, 19667–19679 (cit. on p. 114).

Valin, J.-M., & Skoglund, J., (2019), LPCNet: improving neural speech synthesis through linear prediction, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5891–5895 (cit. on p. 80).

Vallet, F., Essid, S., & Carrive, J., (2012), A multimodal approach to speaker diarization on tv talk-shows, *IEEE transactions on multimedia*, *15 3*, 509–520 (cit. on p. 105).

Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al., (2016), Conditional image generation with pixelcnn decoders, *Advances in neural information processing systems*, *29* (cit. on p. 68).

Van den Oord, A., Vinyals, O., & Kavukcuoglu, K., (2017), Neural discrete representation learning, *Advances in neural information processing systems*, *30* (cit. on pp. 25, 66, 67, 114, 116, 117, 127, 140, 142, 172, 196).

Van der Pligt, J., Zeelenberg, M., van Dijk, W. W., de Vries, N. K., & Richard, R., (1997), Affect, attitudes and decisions: let's be more specific, *European review of social psychology*, *8 1*, 33–66 (cit. on p. 14).

Van Steenkiste, S., Locatello, F., Schmidhuber, J., & Bachem, O., (2019), Are disentangled representations helpful for abstract visual reasoning?, *Advances in neural information processing systems*, *32* (cit. on pp. 52, 74, 105, 169).

Vapnik, V. N., (1999), An overview of statistical learning theory, *IEEE transactions on neural networks*, *10 5*, 988–999 (cit. on p. 19).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I., (2017), Attention is all you need, *Advances in neural information processing systems*, *30* (cit. on pp. 16, 72, 144).

Ververidis, D., Kotropoulos, C., & Pitas, I., (2004), Automatic emotional speech classification, *2004 IEEE international conference on acoustics, speech, and signal processing*, *1*, I–593 (cit. on p. 19).

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A., (2008), Extracting and composing robust features with denoising autoencoders, *International conference on Machine learning (ICML)*, 1096–1103 (cit. on p. 70).

Voloshynovskiy, S., Taran, O., Kondah, M., Holotyak, T., & Rezende, D., (2020), Variational information bottleneck for semi-supervised classification, *Entropy*, *22 9*, 943 (cit. on p. 34).

Wang, C., & Blei, D. M., (2013), Variational inference in nonconjugate models, *Journal of Machine Learning Research*, *14 Apr*, 1005–1031.

Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., & Loy, C. C., (2020), Mead: a large-scale audio-visual dataset for emotional talking-face generation, *European Conference on Computer Vision (ECCV)*, 700–717 (cit. on pp. 115, 130, 199).

Wang, X., Chen, H., Tang, S., Wu, Z., & Zhu, W., (2022), Disentangled representation learning, *arXiv preprint arXiv:2211.11695* (cit. on p. 51).

Wang, X., Takaki, S., & Yamagishi, J., (2019), Neural source-filter waveform models for statistical parametric speech synthesis, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 402–415 (cit. on p. 80).

Wang, Y., Boumadane, A., & Heba, A., (2021), A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding, *arXiv preprint arXiv:2111.02735* (cit. on p. 138).

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P., (2004), Image quality assessment: from error visibility to structural similarity, *IEEE transactions on image processing*, *134*, 600–612 (cit. on p. 118).

Wara, M., & Victor, D. G., (2008), A realistic policy on international carbon offsets, *Program on Energy and Sustainable Development Working Paper*, *74*, 1–24 (cit. on p. 166).

Webber, J. J., Perrotin, O., & King, S., (2020), Hider-finder-combiner: an adversarial architecture for general speech signal modification., *Interspeech*, 3206–3210 (cit. on p. 81).

Wei, X., Li, H., Sun, J., & Chen, L., (2018), Unsupervised domain adaptation with regularized optimal transport for multimodal 2d+ 3d facial expression recognition, *IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 31–37 (cit. on p. 134).

Wu, C.-H., Huang, Y.-M., & Hwang, J.-P., (2016), Review of affective computing in education/learning: trends and challenges, *British Journal of Educational Technology*, *476*, 1304–1323 (cit. on pp. 14, 167).

Wu, C.-H., Lin, J.-C., & Wei, W.-L., (2014), Survey on audiovisual emotion recognition: databases, features, and data fusion strategies, *APSIPA transactions on signal and information processing*, *3*, e12 (cit. on p. 105).

Wu, M., & Goodman, N., (2018), Multimodal generative models for scalable weakly-supervised learning, *Advances in neural information processing systems (NeurIPS)*, *31* (cit. on pp. 61–63, 136, 170).

Xie, B., & Park, C. H., (2023), Multi-modal correlated network with emotional reasoning knowledge for social intelligence question-answering, *International Conference on Computer Vision (IEEE/CVF)*, 3075–3081 (cit. on p. 163).

Xie, J., Lu, Y., Zhu, S.-C., & Wu, Y., (2016), A theory of generative convnet, *International Conference on Machine Learning (ICML)*, 2635–2644 (cit. on p. 42).

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., & Hu, H., (2022), SimMiM: a simple framework for masked image modeling, *International conference on computer vision (IEEE/CVF)*, 9653–9663 (cit. on pp. 70, 139).

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Wei, Y., Dai, Q., & Hu, H., (2023), On data scaling in masked image modeling, *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10365–10374 (cit. on p. 73).

Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., Metze, F., Feichtenhofer, C., et al., (2022), Masked autoencoders that listen, *arXiv preprint arXiv:2207.06405* (cit. on pp. 139, 196).

Yamamoto, R., Felipe, J., & Blaauw, M., (2019), *Pysptk* [URL: https://github.com/r9y9/pysptk], (cit. on p. 100).

Yan, S., (2023), *Personalizing facial expressions by exploring emotional mental prototypes* [Doctoral dissertation, CentraleSupélec, Université Paris-Saclay], (cit. on p. 162).

Yang, H.-C., & Lee, C.-C., (2019), An attribute-invariant variational learning for emotion recognition using physiology, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1184–1188 (cit. on p. 74).

Yang, K., Zhang, T., & Ananiadou, S., (2023), Disentangled variational autoencoder for emotion recognition in conversations, *IEEE Transactions on Affective Computing* (cit. on p. 74).

Yang, Z., Chen, M., Saad, W., Hong, C. S., & Shikh-Bahaei, M., (2020), Energy efficient federated learning over wireless communication networks, *IEEE Transactions on Wireless Communications*, *20 3*, 1935–1949 (cit. on p. 165).

Yin, H., Melo, F. S., Billard, A., & Paiva, A., (2017), Associate latent encodings in learning from demonstrations, *AAAI conference on artificial intelligence*.

Yu, J., Li, X., Koh, J. Y., Zhang, H., Pang, R., Qin, J., Ku, A., Xu, Y., Baldridge, J., & Wu, Y., (n.d.), Vector-quantized image modeling with improved VQGAN, *International Conference on Learning Representations* (cit. on p. 151).

Zeiler, M. D., & Fergus, R., (2014), Visualizing and understanding convolutional networks, *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, 818–833 (cit. on p. 162).

Zhang, C., Zhang, C., Song, J., Yi, J. S. K., Zhang, K., & Kweon, I. S., (2022), A survey on masked autoencoder for self-supervised learning in vision and beyond, *arXiv preprint arXiv:2208.00173* (cit. on pp. 70, 138).

Zhang, R., Isola, P., & Efros, A. A., (2016), Colorful image colorization, *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, 649–666 (cit. on p. 70).

Zhang, Y.-J., Pan, S., He, L., & Ling, Z.-H., (2019), Learning latent representations for style control and transfer in end-to-end speech synthesis, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6945–6949.

Zhao, H., Gan, C., Ma, W.-C., & Torralba, A., (2019), The sound of motions, *IEEE/CVF International Conference on Computer Vision*, 1735–1744.

Zhao, Z., Liu, Q., & Zhou, F., (2021), Robust lightweight facial expression recognition network with label distribution training, *Conference on artificial intelligence (AAAI)*, *35 4*, 3510–3519 (cit. on p. 131).

Zhu, J.-Y., Zhang, Z., Zhang, C., Wu, J., Torralba, A., Tenenbaum, J., & Freeman, B., (2018), Visual object networks: image generation with disentangled 3d representations, *Advances in neural information processing systems*, *31* (cit. on p. 52).

COLLEGE DOCTORAL BRETAGNE | MATHS, TELECOMS INFORMATIQUE, SIGNAL SYSTEMES, ELECTRONIQUE

CentraleSupélec

Titre : Apprentissage de représentation de la parole audiovisuelle pour la reconnaissance des émotions.

Mot clés : Apprentissage profond des représentations, modèle génératif profond, reconaissance des émotions, parole audiovisuelle.

Résumé : La rareté des données étiquetées constitue un défi majeur dans la reconnaissance des émotions audiovisuelles. Pour relever ce défi, des méthodes récentes d'apprentissage non supervisé et auto-supervisé ont émergé, visant à réduire la dépendance aux données étiquetées en apprenant des représentations robustes applicables à diverses tâches. Les représentations apprises doivent répondre aux critères d'informativité, de généralisabilité, d'interprétabilité et de contrôlabilité. Pour cela, les modèles génératifs profonds ont gagné en importance dans l'apprentissage à partir de données complexes et de grande dimension telles que les images, l'audio et le texte. Cette thèse vise trois objectifs principaux : Premièrement, *développer* des modèles génératifs pour l'apprentissage non-supervisé ou auto-supervisé de représentations de la parole audiovisuelle ; Deuxièmement, *structurer* ces modèles génératifs afin d'apprendre des représentations désentrelacées pour améliorer l'interprétabilité de nos modèles ; Enfin, *analyser* les performances et l'efficacité de ces modèles pré-entraînés pour la tâche de reconnaissance des émotions.

Title: Audiovisual speech representation learning applied to emotion recognition

Keywords: Deep representation learning, deep generative modeling, emotion recognition, audiovisual speech processing.

Abstract: The scarcity of labeled data presents a major challenge in audiovisual speech emotion recognition. Furthermore, the complexity and subjectivity of emotions introduce ambiguity in their representation, which is consequently reflected in data and methods relying on supervised learning. To address this challenge, recent unsupervised and self-supervised learning methods have emerged, aiming to minimize the reliance on labeled data by learning robust representations applicable to various tasks. Effective representations should meet informativeness, generalizability, interpretability, and controllability criteria, where deep generative models have gained prominence for their success in learning from complex and high-dimensional data like images, audio, and text, fulfilling the above criteria. This thesis pursues three primary objectives: First, *developing* and expanding generative models for unsupervised or self-supervised learning of audiovisual speech representation learning; Second, *structuring* the generative model in order to learn disentangled representations to improve the interpretability of our models; And finally, *analyzing* the performance and efficiency of these models applied for the emotion recognition task.