A Multimodal Dynamical Variational Autoencoder for Audiovisual Speech Representation Learning

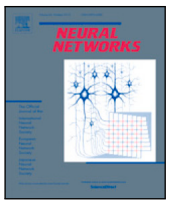
Samir Sadok, Simon Leglaive, Laurent Girin, Xavier Alameda-Pineda, Renaud Séguier

Neural Networks 172 (2024) 106120

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



Full Length Article

A multimodal dynamical variational autoencoder for audiovisual speech representation learning☆

Samir Sadok [a,*], Simon Leglaive [a], Laurent Girin [b], Xavier Alameda-Pineda [c], Renaud Séguier [a]

[a] CentraleSupélec IETR UMR CNRS 6164, France
[b] Univ. Grenoble Alpes CNRS, Grenoble-INP, GIPSA-lab, France
[c] Inria, Univ. Grenoble Alpes CNRS, LJK, France

ARTICLE INFO

Dataset link: https://samsad35.github.io/site-mdvae/

Keywords:
Deep generative modeling
Disentangled representation learning
Variational autoencoder
Multimodal and dynamical data
Audiovisual speech processing

ABSTRACT

High-dimensional data such as natural images or speech signals exhibit some form of regularity, preventing their dimensions from varying independently. This suggests that there exists a lower dimensional latent representation from which the high-dimensional observed data were generated. Uncovering the hidden explanatory features of complex data is the goal of representation learning, and deep latent variable generative models have emerged as promising unsupervised approaches. In particular, the variational autoencoder (VAE) which is equipped with both a generative and an inference model allows for the analysis, transformation, and generation of various types of data. Over the past few years, the VAE has been extended to deal with data that are either multimodal *or* dynamical (i.e., sequential). In this paper, we present a multimodal *and* dynamical VAE (MDVAE) applied to unsupervised audiovisual speech representation learning. The latent space is structured to dissociate the latent dynamical factors that are shared between the modalities from those that are specific to each modality. A static latent variable is also introduced to encode the information that is constant over time within an audiovisual speech sequence. The model is trained in an unsupervised manner on an audiovisual emotional speech dataset, in two stages. In the first stage, a vector quantized VAE (VQ-VAE) is learned independently for each modality, without temporal modeling. The second stage consists in learning the MDVAE model on the intermediate representation of the VQ-VAEs before quantization. The disentanglement between static versus dynamical and modality-specific versus modality-common information occurs during this second training stage. Extensive experiments are conducted to investigate how audiovisual speech latent factors are encoded in the latent space of MDVAE. These experiments include manipulating audiovisual speech, audiovisual facial image denoising, and audiovisual speech emotion recognition. The results show that MDVAE effectively combines the audio and visual information in its latent space. They also show that the learned static representation of audiovisual speech can be used for emotion recognition with few labeled data, and with better accuracy compared with unimodal baselines and a state-of-the-art supervised model based on an audiovisual transformer architecture.

1. Introduction and related work

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 and 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, Ahuja, & Morency, 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

☆ This work was supported by Randstad corporate research chair, by ANR-3IA MIAI (ANR-19-P3IA-0003), by ANR-JCJC ML3RI (ANR-19-CE33-0008-01), and by H2020 SPRING (funded by EC under GA #871245).

* Corresponding author.
E-mail address: samir.sadok@centralesupelec.fr (S. Sadok).

https://doi.org/10.1016/j.neunet.2024.106120
Received 26 May 2023; Received in revised form 25 October 2023; Accepted 9 January 2024
Available online 11 January 2024

digital content and communication, audiovisual speech processing has become increasingly important for a range of applications, such as speech recognition (Afouras, Chung, Senior, Vinyals, & Zisserman, 2018; Hori et al., 2019; Petridis et al., 2018), speaker identification (Roth et al., 2020), and emotion recognition (Noroozi, Marjanovic, Njegus, Escalera, & Anbarjafari, 2017; Schoneveld, Othmani, & Abdelkawy, 2021; Wu, Lin, & Wei, 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, Mohamed, & Wierstra, 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, it should capture high-level characteristics that are invariant to small and local changes in the input data, and it should be as disentangled as possible for explainability. Furthermore, hierarchical and disentangled generative models have demonstrated their efficacy to solve downstream learning tasks (Bengio et al., 2013; Van Steenkiste, Locatello, Schmidhuber, & Bachem, 2019). Variants of generative models have recently led to considerable progress in disentangled representation learning, particularly with the variational autoencoder (VAE) (Kingma & Welling, 2014; Rezende et al., 2014).

The VAE considers that an observed high-dimensional data vector $\mathbf{x}$ is generated by a low-dimensional latent vector $\mathbf{z}$. The generative process is characterized by the joint distribution $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})$, where $p(\mathbf{z})$ is the prior distribution over the latent variable and $p_\theta(\mathbf{x} \mid \mathbf{z})$ is the conditional likelihood that characterizes how the observed data is generated from the latent variable. In the VAE, the conditional likelihood is parameterized by a neural network called the decoder, whose parameters are denoted by θ. The VAE also comes with an inference model $q_\phi(\mathbf{z} \mid \mathbf{x})$, which approximates the intractable posterior distribution of the latent variable. This inference model is parameterized by a second neural network, called the encoder, whose parameters are denoted by ϕ. The inference model allows us to extract the latent variable $\mathbf{z}$ from an observed data vector $\mathbf{x}$, and the generative model allows us to generate $\mathbf{x}$ from $\mathbf{z}$. The parameters of both the inference (i.e., encoder) and generative (i.e., decoder) models are efficiently and jointly learned by maximizing a lower bound of the training data log-likelihood called the evidence lower-bound (ELBO), which is central in variational methods for inference and learning in probabilistic graphical models (Jordan, Ghahramani, Jaakkola, & Saul, 1999; Neal & Hinton, 1998).

The VAE enables deep unsupervised representation learning in a Bayesian framework. Typically, a standard Gaussian distribution is chosen for the prior distribution over the latent variable: $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$. This choice encourages the independence of the different dimensions in the learned representation, and it is considered a key factor contributing to VAE's potential for disentanglement. 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 in the model and/or in the learning algorithm. Early methods for disentanglement using VAEs focused on modifying the evidence lower bound objective function (Chen, Li, Grosse, & Duvenaud, 2018; Higgins et al., 2016; 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 et al., 2020; Sadok, Leglaive, Girin, Alameda-Pineda, & Séguier, 2023) or semi-supervised learning (Klys, Snell, & Zemel, 2018). Because of their flexibility in modeling complex data, VAEs have been extended to various types of data, including multimodal or sequential data.

VAEs have gained significant interest in modeling multimodal data due to their several advantages compared to other generative models, especially generative adversarial networks (GANs) (Goodfellow et al., 2014). VAEs are equipped with encoder and decoder models, resulting in a more stable and faster training process than GANs, which makes them well-suited for multimodal generative modeling (Suzuki & Matsuo, 2022). Several approaches have been developed to learn a joint latent space for multiple heterogeneous input data. For instance, PoE-VAE (Wu & Goodman, 2018) adopts the product of experts (PoE) (Hinton, 2002) to model the posterior distribution of multimodal data while MoE-VAE (Shi, Paige, Torr, et al., 2019) uses a mixture of experts (MoE). Another approach (Sutter, Daunhawer, & Vogt, 2021) combines these methods for improved data reconstruction. Nevertheless, limitations have been demonstrated and formalized for these methods (Daunhawer, Sutter, Chin-Cheong, Palumbo, & Vogt, 2021). For example, multimodal VAE models often produce lower-quality reconstructions compared to unimodal VAE models, particularly for complex data. To address this issue and achieve better inference, many multimodal generative models now use hierarchical approaches to disentangle joint information from modality-specific information (Hsu & Glass, 2018; Lee & Pavlovic, 2020; Sutter, Daunhawer, & Vogt, 2020).

Another area where VAE models have seen significant progress is in the modeling of sequential data, where the latent and/or observed variables evolve over time. Dynamical VAEs (DVAEs) (Girin et al., 2021) aim to tackle high-dimensional complex data exhibiting temporal or spatial correlations using deep dynamical Bayesian networks. Recurrent neural networks are often used for this purpose, and a wide range of methods have been developed that differ in their inference and generative model structures. 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 a lower bound of the log-marginal likelihood (ELBO). Of particular interest to the present paper is the disentangled sequential autoencoder (DSAE) (Li & Mandt, 2018), which separates dynamical from static latent information.

While many extensions of the VAE have been proposed to handle either multimodal or sequential data, none have been able to process both types of data simultaneously. This paper 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.

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 perspective of audiovisual speech generative modeling. Subsequently, we formalize the MDVAE generative and inference models. Finally, we introduce a two-stage training approach for unsupervised learning of the MDVAE model.

2.1. Motivation and notations

Our goal is 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

Table 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

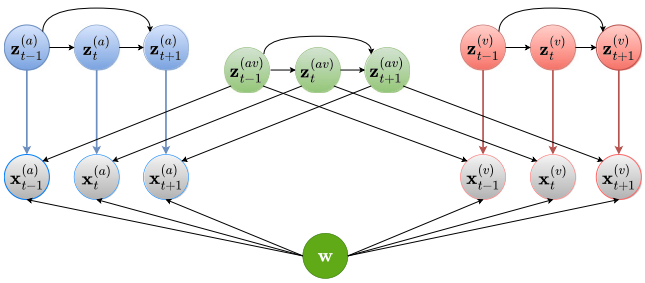


Fig. 1. MDVAE generative probabilistic graphical model.

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, we have the speaker's identity and global emotional state that 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 that is 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 $t \in \{1, 2, \ldots, T\}$ is added in 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 1. Note that for the specific modeling of audiovisual speech data, it is not particularly relevant to introduce static latent variables that are specific to each modality. Indeed, as above-described, the static information is here assumed to encode the speaker's identity and global emotional state, which are fundamentally multimodal factors. Nevertheless, for other types of data, we can readily envision using two distinct static latent variables for each modality. The methodology developed in the subsequent sections could be straightforwardly extended to this case. 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 shared between the modalities ($\mathbf{w}, \mathbf{z}^{(av)}$) from those that are specific to each modality ($\mathbf{z}^{(a)}, \mathbf{z}^{(v)}$). Note that a study by Gao and Shinkareva (Gao & 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 than can disentangle the above-mentioned latent factors in an unsupervised manner for the analysis and transformation of emotional audiovisual speech data. In the next subsections, we detail the generative and inference models of MDVAE and its two-stage training.

2.2. Generative model

The generative model of MDVAE is represented as a Bayesian network in Fig. 1, which also corresponds to the following factorization of the joint distribution of the observed and latent variables:

$$p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta\left(\mathbf{x}^{(a)} \mid \mathbf{w}, \mathbf{z}^{(av)}, \mathbf{z}^{(a)}\right) p_\theta\left(\mathbf{x}^{(v)} \mid \mathbf{w}, \mathbf{z}^{(av)}, \mathbf{z}^{(v)}\right) p_\theta(\mathbf{w})\, p_\theta\left(\mathbf{z}^{(av)}\right) \times p_\theta\left(\mathbf{z}^{(a)}\right) p_\theta\left(\mathbf{z}^{(v)}\right), \tag{1}$$

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

$$p_\theta\left(\mathbf{x}^{(a)} \mid \mathbf{w}, \mathbf{z}^{(av)}, \mathbf{z}^{(a)}\right) = \prod_{t=1}^{T} p_\theta\left(\mathbf{x}_t^{(a)} \mid \mathbf{w}, \mathbf{z}_t^{(av)}, \mathbf{z}_t^{(a)}\right); \tag{2}$$

$$p_\theta\left(\mathbf{x}^{(v)} \mid \mathbf{w}, \mathbf{z}^{(av)}, \mathbf{z}^{(v)}\right) = \prod_{t=1}^{T} p_\theta\left(\mathbf{x}_t^{(v)} \mid \mathbf{w}, \mathbf{z}_t^{(av)}, \mathbf{z}_t^{(v)}\right); \tag{3}$$

$$p(\mathbf{z}^{(av)}) = \prod_{t=1}^{T} p_\theta\left(\mathbf{z}_t^{(av)} \mid \mathbf{z}_{1:t-1}^{(av)}\right); \tag{4}$$

$$p(\mathbf{z}^{(a)}) = \prod_{t=1}^{T} p_\theta\left(\mathbf{z}_t^{(a)} \mid \mathbf{z}_{1:t-1}^{(a)}\right); \tag{5}$$

$$p(\mathbf{z}^{(v)}) = \prod_{t=1}^{T} p_\theta\left(\mathbf{z}_t^{(v)} \mid \mathbf{z}_{1:t-1}^{(v)}\right). \tag{6}$$

Eq. (2) (resp. (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 the generation of 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)–(6) is inspired by DSAE (Li & Mandt, 2018). Following standard DVAEs (Girin et al., 2021), each conditional distribution that appears in a product over the time indices in equations (2)–(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 (2)–(3), the variance coefficients are fixed to one, while for the distributions in equations (4)–(6), the variance coefficients are learned. Standard feed-forward fully-connected neural networks can be used for parametrizing the conditional distributions

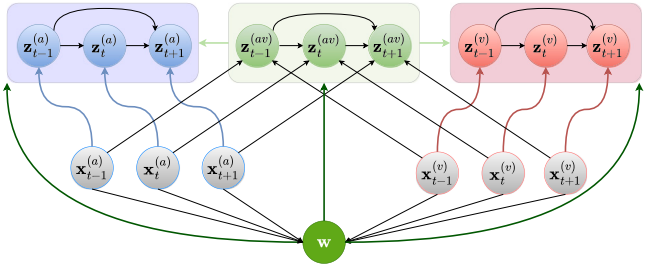


Fig. 2. MDVAE inference probabilistic graphical model.

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 Appendix A.

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 $q_\phi(\mathbf{z} \mid \mathbf{x}) \approx p_\theta(\mathbf{z} \mid \mathbf{x})$. 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, Verma, & Pearl, 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. (2021). The Bayesian network corresponding to our MDVAE model is represented in Fig. 1. For this model, it is relevant to factorize the inference model as follows:

$$q_\phi(\mathbf{z} \mid \mathbf{x}) = q_\phi\left(\mathbf{w} \mid \mathbf{x}^{(a)}, \mathbf{x}^{(v)}\right) q_\phi\left(\mathbf{z}^{(av)} \mid \mathbf{x}^{(a)}, \mathbf{x}^{(v)}, \mathbf{w}\right) q_\phi\left(\mathbf{z}^{(a)} \mid \mathbf{x}^{(a)}, \mathbf{z}^{(av)}, \mathbf{w}\right) \times q_\phi\left(\mathbf{z}^{(v)} \mid \mathbf{x}^{(v)}, \mathbf{z}^{(av)}, \mathbf{w}\right), \tag{7}$$

where

$$q_\phi\left(\mathbf{z}^{(av)} \mid \mathbf{x}^{(a)}, \mathbf{x}^{(v)}, \mathbf{w}\right) = \prod_{t=1}^{T} q_\phi\left(\mathbf{z}_t^{(av)} \mid \mathbf{z}_{1:t-1}^{(av)}, \mathbf{x}_{t:T}^{(a)}, \mathbf{x}_{t:T}^{(a)}, \mathbf{w}\right); \tag{8}$$

$$q_\phi\left(\mathbf{z}^{(a)} \mid \mathbf{x}^{(a)}, \mathbf{z}^{(av)}, \mathbf{w}\right) = \prod_{t=1}^{T} q_\phi\left(\mathbf{z}_t^{(a)} \mid \mathbf{z}_{1:t-1}^{(a)}, \mathbf{x}_{t:T}^{(a)}, \mathbf{z}_t^{(av)}, \mathbf{w}\right); \tag{9}$$

$$q_\phi\left(\mathbf{z}^{(v)} \mid \mathbf{x}^{(v)}, \mathbf{z}^{(av)}, \mathbf{w}\right) = \prod_{t=1}^{T} q_\phi\left(\mathbf{z}_t^{(v)} \mid \mathbf{z}_{1:t-1}^{(v)}, \mathbf{x}_{t:T}^{(a)}, \mathbf{z}_t^{(av)}, \mathbf{w}\right). \tag{10}$$

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 (8), (9) and (10). In these equations, we thus replace $\mathbf{x}_{t:T}^{(a)}$ by $\mathbf{x}_t^{(a)}$ and $\mathbf{x}_{t:T}^{(a)}$ by $\mathbf{x}_t^{(v)}$, and the equalities become approximations. In this inference model, $q_\phi\left(\mathbf{w} \mid \mathbf{x}^{(a)}, \mathbf{x}^{(v)}\right)$ and each conditional distribution that appears in a product over the time indices in equations (8)–(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 Appendix A.

The probabilistic graphical model of MDVAE during inference is represented in Fig. 2, corresponding to the factorization in (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 $q_\phi(\mathbf{w} \mid \mathbf{x}^{(a)}, \mathbf{x}^{(v)})$. 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 $q_\phi(\mathbf{z}^{(av)} \mid \mathbf{x}^{(a)}, \mathbf{x}^{(v)}, \mathbf{w})$. 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 $q_\phi(\mathbf{z}^{(a)} \mid \mathbf{x}^{(a)}, \mathbf{z}^{(av)}, \mathbf{w})$ (resp. $q_\phi(\mathbf{z}^{(v)} \mid \mathbf{x}^{(v)}, \mathbf{z}^{(av)}, \mathbf{w})$). 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 $\mathbf{z}^{(av)}$.

2.4. Training

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

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln p_\theta(\mathbf{x} \mid \mathbf{z})\right] - D_{\mathrm{KL}}\left(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z})\right). \tag{11}$$

where D_{KL} is the Kullback-Leibler divergence, defined by $D_{\mathrm{KL}}(q \parallel p) = \mathbb{E}_q[\ln q - \ln p]$. The first term in (11) is the reconstruction accuracy term, which aims to maximize la data log-likelihood over a training dataset. These 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. Using equations (1) and (7), the ELBO can be further developed as follows:

$$\begin{aligned}\mathcal{L}(\theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}&\left[\ln p\left(\mathbf{x}^{(a)}, \mathbf{x}^{(v)} \mid \mathbf{w}, \mathbf{z}^{(av)}, \mathbf{z}^{(a)}, \mathbf{z}^{(v)}\right)\right] \\ &- D_{KL}(q_\phi(\mathbf{w} \mid \mathbf{x}^{(a)}, \mathbf{x}^{(v)}) \parallel p_\theta(\mathbf{w})) \\ &- \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[D_{KL}\left(q_\phi\left(\mathbf{z}^{(av)} \mid \mathbf{x}^{(a)}, \mathbf{x}^{(v)}, \mathbf{w}\right) \parallel p_\theta\left(\mathbf{z}^{(av)}\right)\right)\right] \\ &- \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[D_{KL}\left(q_\phi\left(\mathbf{z}^{(a)} \mid \mathbf{x}^{(a)}, \mathbf{w}, \mathbf{z}^{(av)}\right) \parallel p_\theta\left(\mathbf{z}^{(a)}\right)\right)\right] \\ &- \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[D_{KL}\left(q_\phi\left(\mathbf{z}^{(v)} \mid \mathbf{x}^{(v)}, \mathbf{w}, \mathbf{z}^{(av)}\right) \parallel p_\theta\left(\mathbf{z}^{(v)}\right)\right)\right].\end{aligned} \tag{12}$$

2.5. Two-stage training

Unlike GANs (Goodfellow et al., 2014), VAEs often produce poor reconstructions that lack realism, and this 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, Sønderby, Larochelle, & Winther, 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,

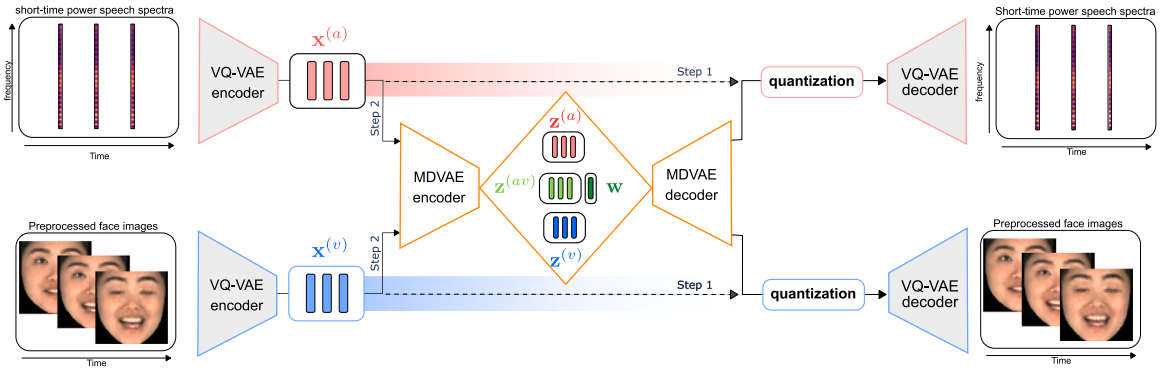


Fig. 3. The overall architecture of VQ-MDVAE. During the first step of the training process, we learn a VQ-VAE independently on each modality, without any temporal modeling. During the second step of the training process, we learn the MDVAE model on the latent representation provided by the frozen VQ-VAE encoders, before quantization.

to ensure that VAE outputs have similar deep features to their corresponding inputs (Hou, Sun, Shen, & Qiu, 2019; Pihlgren, Sandin, & Liwicki, 2020). In this work, we focus on using a vector quantized VAE (VQ-VAE) model (Van Den Oord, Vinyals, 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, Van den Oord, & Vinyals, 2019). Therefore, as illustrated in Fig. 3, we propose a two-stage training approach of 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, Bertin, & Durrieu, 2009) 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.

Beyond the reconstruction quality of the audiovisual speech data, another interest of the proposed two-stage training procedure is that it allows us to distribute the GPU memory usage between the two training stages, which is particularly useful when working with limited computational resources. During the first training stage, we train *small models* (fully convolutional VQ-VAEs) on *high-dimensional data*, corresponding to the raw audio and visual speech data. Processing modalities and time frames independently along with defining models of reasonable size allows us to efficiently manage the GPU memory to create large batches containing the raw audio or visual speech data. During the second training stage, we train a *large model* (MDVAE) on *low-dimensional data*, corresponding to the compressed audio and visual latent representations provided by the pre-trained VQ-VAEs before quantization. The effective compression of the audiovisual speech data using the VQ-VAEs allows us to increase the model capacity for the second training stage, which is required by the complexity of the learning task, and at the same time to keep a sufficiently large batch size. In summary, the two-stage training strategy decomposes the difficult problem of learning from high-dimensional multimodal and sequential data into two smaller sub-problems. The first sub-problem deals with compressing the data and ensuring a good reconstruction quality, and the second sub-problem deals with modality fusion and temporal modeling. Overall, this two-stage training leads to good reconstruction quality, efficient memory management, and accelerated training speed. Note that from a practical perspective, one could learn the VQ-VAE and MDVAE models jointly, from scratch, given sufficient computational resources.

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.

3.1. Expressive audiovisual speech dataset

The VQ-MDVAE model is trained on the multi-view emotional audiovisual dataset (MEAD) (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 25 h, 5 h, and 3 h of audiovisual speech, respectively. For the visual modality, face images in the MEAD dataset are cropped, resized to a 64×64 resolution, and aligned using Openface (Baltrušaitis, Robinson, & Morency, 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.

Table 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	2.43 ± 0.28	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	2.90 ± 0.23	3.54 ± 0.20	6.85 ± 1.15

Table 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

3.2. Training VQ-MDVAE

The architecture of the MDVAE and VQ-VAE models are described in detail in Appendix A. 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).

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 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 (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 2.5). 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, Hendriks, Heusdens, & Jensen, 2010);
- The Perceptual Evaluation of Speech Quality (PESQ) measure is an intrusive metric that evaluates the perceived quality of the resynthesized speech (Rix, Beerends, Hollier, & Hekstra, 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, Wisdom, Erdogan, & Hershey, 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 (Wang, Bovik, Sheikh, & Simoncelli, 2004).

Discussions Tables 2 and 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,

Fig. 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.

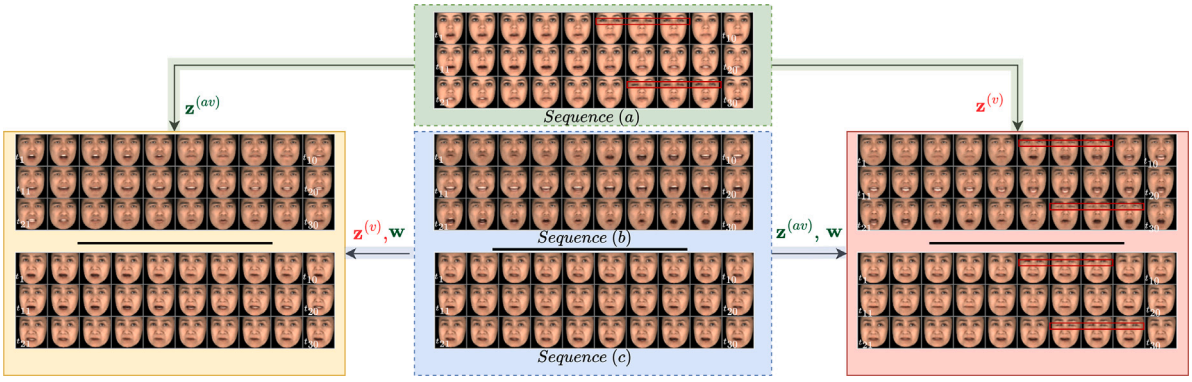


Fig. 5. This figure demonstrates the qualitative significance of each latent space for visual data using the *analysis-transformation-synthesis* experiment. The sequences in the yellow box (left) were generated using $\mathbf{z}^{(av)}$ from sequence (a) and $\mathbf{z}^{(v)}, \mathbf{w}$ from sequences (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).

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.

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.

3.4.1. Qualitative results

Visual modality Fig. 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 (AU26) and eyes closed (AU43), 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 PyFeat (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 AU43 and AU26 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 AU43 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.

Fig. 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

Fig. 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).

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.

Fig. 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.

Audio modality As for the visual modality, Fig. 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 (Kim, Salamon, Li, & Bello, 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 at https://samsad35.github.io/site-mdvae/.

3.4.2. 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 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 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, 2021) and CREPE (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 Fig. 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, Lewin, & Brown, 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

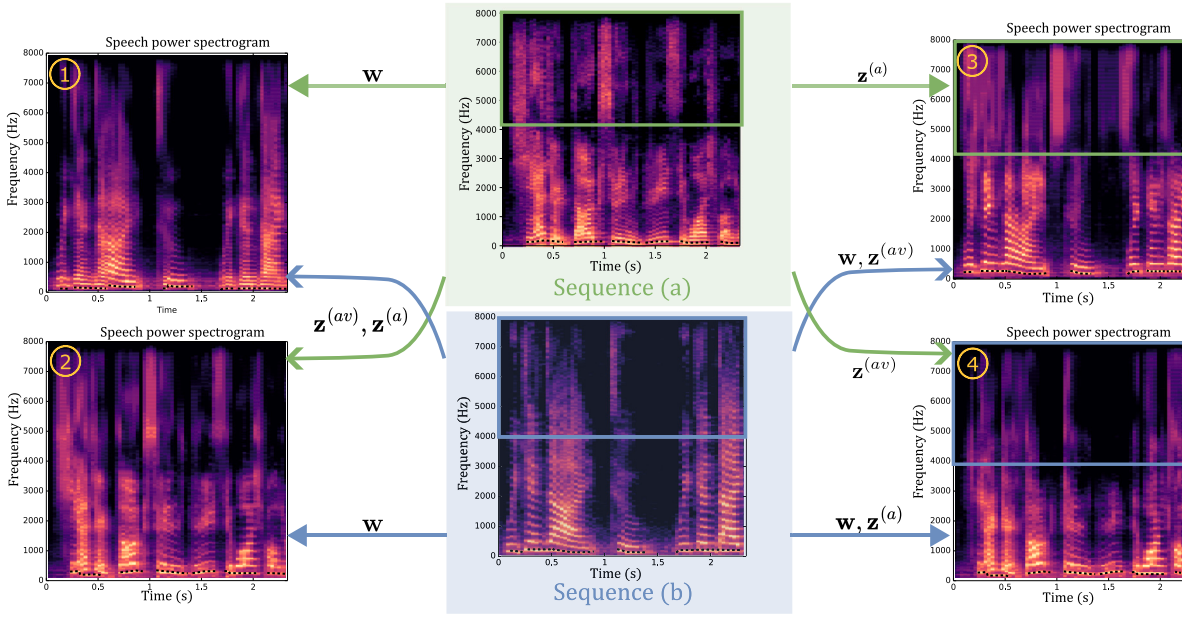


Fig. 7. Audio spectrograms generated from *analysis-transformation-synthesis* between sequence (a) in green and sequence (b) in blue. The spectrograms (1), (2), (3), and (4) are synthesized by swapping latent variables between sequence (a) and sequence (b). The black dotted line corresponds to the pitch contour.

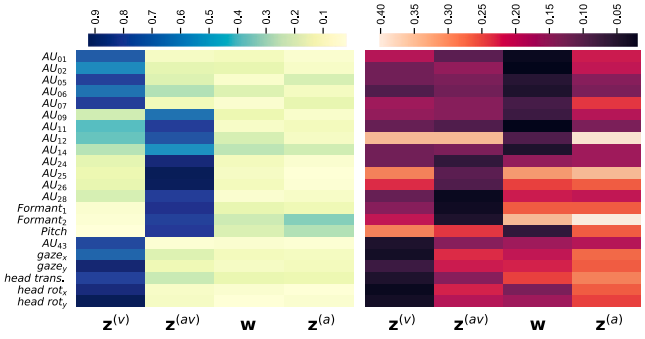
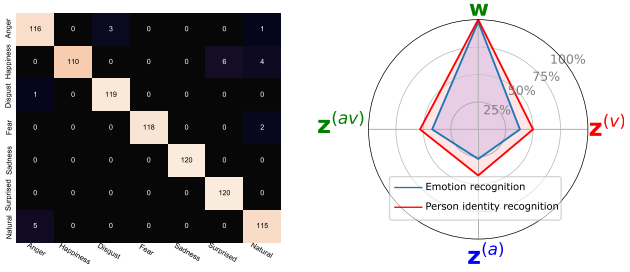


Fig. 8. Relationship between the audio/visual attributes and the latent variables of VQ-MDVAE. (left) Pearson correlation coefficient (PCC), (right) mean absolute error (MAE).



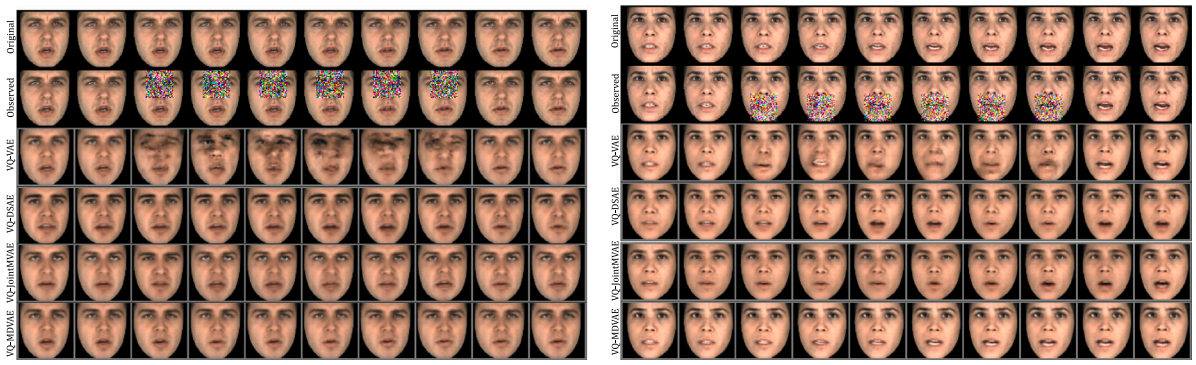
(a) Confusion matrix for emotion classification on the VQ-MDVAE output images after perturbation of the dynamical latent variables.

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

Fig. 9. Analysis of the latent variables of the VQ-MDVAE model in terms of emotion and person identity.

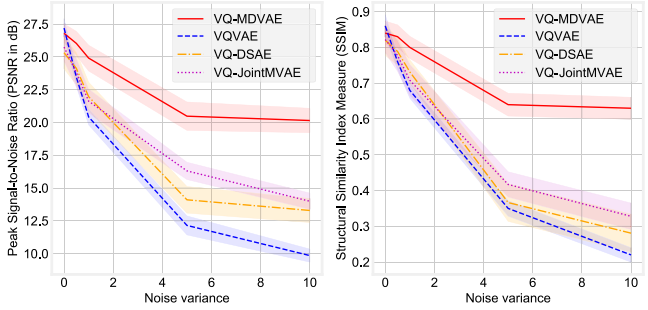
puller (AU_{12}), and dimpler (AU_{14}) on the other side, show high PCC values with respect to $\mathbf{z}^{(v)}$ and $\mathbf{z}^{(av)}$, respectively, but low MAE values with respect to $\mathbf{w}$. We argue that this result is related to the encoding of the speaker's emotional state in the latent 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 3.4.1, 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 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, Vu, & Tran, 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 Fig. 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 Fig. 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%

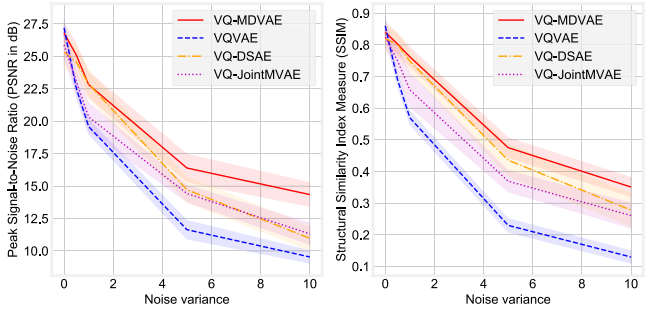


(a) Corruption of the eyes region. (b) Corruption of the mouth region.

Fig. 10. Qualitative comparison of the denoising results. From top to bottom: perturbed sequences; sequences reconstructed with VQ-VAE; sequences reconstructed with DSAE; sequences reconstructed with VQ-JointMVAE; and sequences reconstructed with VQ-MDVAE.



(a) Corruption of the mouth region.



(b) Corruption of the eyes region.

Fig. 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.

correct classification, respectively. Appendix B contains visualizations of the static latent space, while the aforementioned companion website provides qualitative results of interpolations on **w**, demonstrating how we can modify the emotion within an audiovisual speech sequence without altering the identity, and vice versa.

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 (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, Nakayama, & Matsuo, 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 Figs. 10 and 11 present the qualitative and quantitative results for 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 Figs. 11(a) and 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.

3.6. Audiovisual speech emotion recognition

This section presents emotion recognition experiments based on the static audiovisual representation **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 (Wang et al., 2020) and

RAVDESS (Livingstone & Russo, 2018). The MEAD dataset was presented in Section 3.1. The RAVDESS dataset contains 1440 audio files that were recorded by 24 professional actors, with each file 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 in the case 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 in the case where 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 several 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 (Li & Mandt, 2018) that uses the 2-stage training process proposed in the present paper. 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, Baevski, Collobert, & Auli, 2019), which is a self-supervised unimodal representation learning approach. Wav2vec is trained on the audio speech signals of the Librispeech dataset (Panayotov, Chen, Povey, & Khudanpur, 2015), which includes 960 h of unlabeled speech data, with 2338 different speakers. Finally, we also include in this experiment two state-of-the-art supervised multimodal approaches (Chumachenko, Iosifidis, & Gabbouj, 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 (Zhao, Liu, & Zhou, 2021), a model pre-trained on AffectNet (Mollahosseini, Hasani, & Mahoor, 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, Riera, and Ferrer (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, VQ-DSAE-audio and VQ-DSAE-visual, for the emotion category classification task on the MEAD dataset. In Fig. 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.

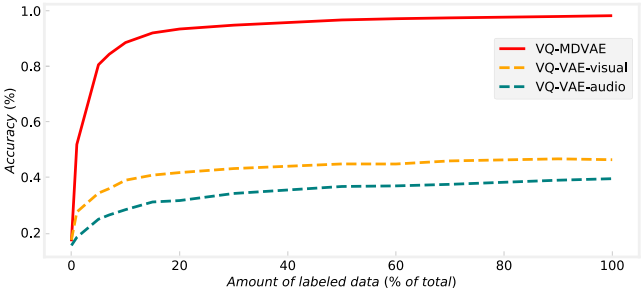


Fig. 12. Accuracy for emotion category classification as a function of the amount of labeled data used to train the MLR classification model on the MEAD dataset in the person-dependent evaluation setting.

Table 4 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 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 Fig. B.14 of Appendix B, 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 4 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 Fig. B.14 of Appendix B. 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.

Table 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 (Li & Mandt, 2018)	40.4	39.3	–	–	48.7	45.7	–	–
		VQ-DSAE-visual (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	–	–	–	–

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, Flamary, Habrard, & Rakotomamonjy, 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 (Kim & Song, 2022; Wei, Li, Sun, & Chen, 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. Conclusion

Deep generative modeling is a powerful unsupervised learning paradigm that can be applied to many different types of data. In this paper, 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 were able to 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 shown to be effective with only a few labeled data, and it 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 "subsampled training" approach proposed in Wu and Goodman (2018), or maybe using the multimodal masking strategies proposed in Bachmann, Mizrahi, Atanov, and Zamir (2022). Moreover, being able to infer 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.

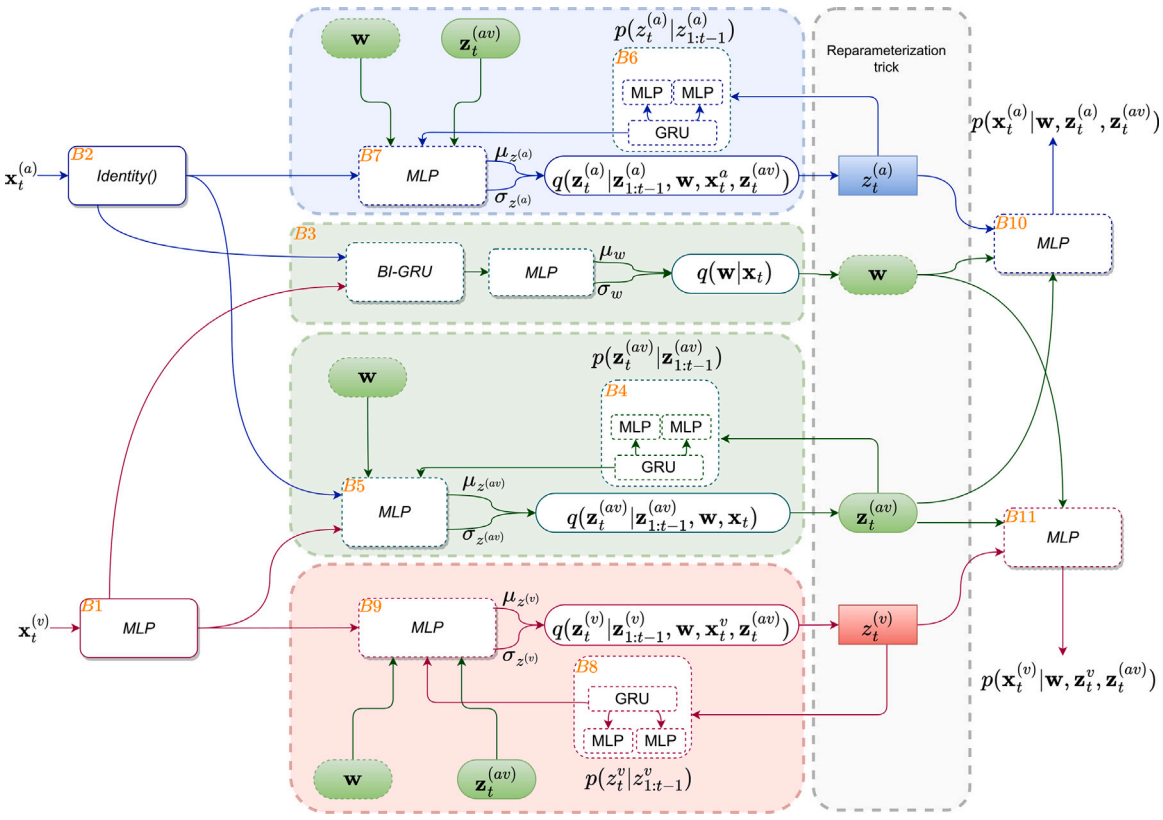


Fig. A.13. (Better zoom in) The overall architecture of the MDVAE.

Table A.5
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)}			

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The code will be shared on https://samsad35.github.io/site-mdvae/, but not the data.

Appendix A. 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.

A.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

Table A.6
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)}`

Table A.7
The architecture details of the MDVAE. The blocks from B1 to B11 are illustrated in Fig. A.13 to better understand their interactions.

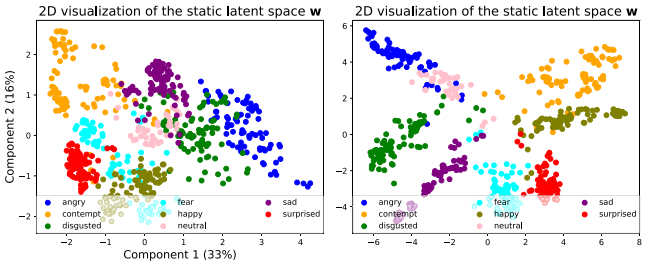
Block	Layer	Activation	Output dim.
$B1$	`Linear(`$32 \cdot 8 \cdot 8$`, 1024)`	ReLu	1024
	`Linear(1024, 512)`	ReLu	$r_v = 512$
$B2$	`Identity`	–	$r_a = 512$
$B3$	`GRU(`$r_v + r_a$`, 256, 1, True)`	–	$2 \cdot 256$
	`Linear(`$2 \cdot 256$`, 256)`	Tanh	256
	$\sigma_{\mathbf{w}}$`: Linear(256, `l_w`)`	–	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(`$l_v + l_{av} + l_w$`, 512)`	ReLu	512
	`Linear(512, 1024)`	ReLu	1024
	`Linear(1024, 2048)`	ReLu	2048
$B11$	`Linear(`$l_a + l_{av} + l_w$`, 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)`

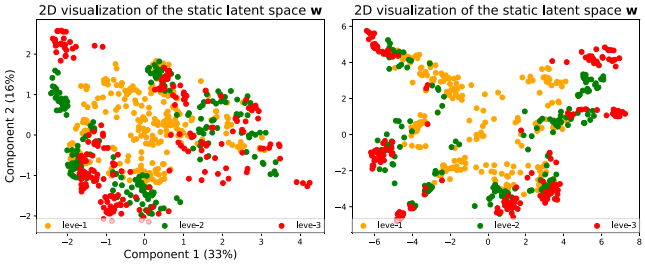
with the nearest vector from the codebook), and (iii) a decoder that maps the indices of the vectors from the codebook back to an image. The architectures of the visual and audio VQ-VAEs are described in Tables A.5 and A.6, respectively.

A.2. MDVAE

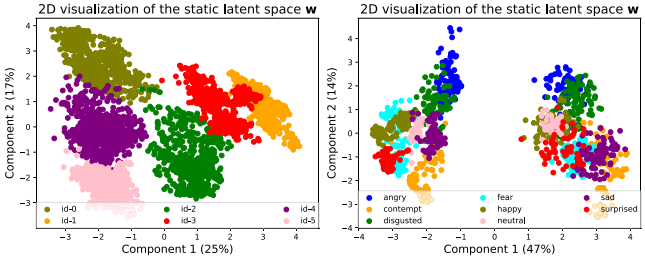
MDVAE is decomposed into two models: (i) the first is the inference model (encoder), which is further decomposed into four inferences for each latent variable, represented by Gaussian distributions whose



(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.14a 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).

Fig. B.14. 2D visualizations of the static latent space.

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 is composed of two decoders, one for the visual modality 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 A.7 and Fig. A.13 present the details of the MDVAE architecture. The figure provides an overview of 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 provide a comprehensive description of the MDVAE architecture.

Appendix B. Visualization of the MDVAE static latent space

2D visualizations of the static latent space of the MDVAE are obtained using dimension reduction methods. Fig. B.14(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. Fig. B.14(b) corresponds to the same visualization but the colors now indicate the emotion intensity levels. It can be seen that for each emotion, the intensity level increases continuously from the middle to the outside of the emotion cluster. Finally, Fig. B.14(c) shows the identity clusters for six different speakers (left figure) and the emotion clusters for two speakers (right figure), both obtained using PCA. 3D visualizations are available on the companion website, along with other dimension reduction methods.

References

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*.

Arnela, M., Blandin, R., Dabbaghchian, S., Guasch, O., Alías, F., Pelorson, X., et al. (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.

Bachmann, R., Mizrahi, D., Atanov, A., & Zamir, A. (2022). Multimae: Multi-modal multi-task masked autoencoders. In *European conference on computer vision* (pp. 348–367). Springer.

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.

Baltrušaitis, T., Robinson, P., & Morency, L.-P. (2016). Openface: An open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision* (pp. 1–10). IEEE.

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.

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.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning, vol. 4, no. 4*. Springer.

Boersma, P., & Weenink, D. (2021). Praat: Doing phonetics by computer [computer program](2011). *Version*, *5*(3), 74.

Chen, R. T., Li, X., Grosse, R. B., & Duvenaud, D. K. (2018). Isolating sources of disentanglement in variational autoencoders. In *Advances in neural information processing systems*: *vol. 31*.

Chumachenko, K., Iosifidis, A., & Gabbouj, M. (2022). Self-attention fusion for audiovisual emotion recognition with incomplete data. In *International conference on pattern recognition* (pp. 2822–2828). IEEE.

Courty, N., Flamary, R., Habrard, A., & Rakotomamonjy, A. (2017). Joint distribution optimal transportation for domain adaptation. In *Advances in neural information processing systems*: *vol. 30*.

Daunhawer, I., Sutter, T. M., Chin-Cheong, K., Palumbo, E., & Vogt, J. E. (2021). On the limitations of multimodal VAEs. In *International conference on learning representations*.

Ekman, P., & Friesen, W. V. (1978). Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

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.

Gao, C., & Shinkareva, S. V. (2021). Modality-general and modality-specific audiovisual valence processing. *Cortex*, *138*, 127–137.

Geiger, D., Verma, T., & Pearl, J. (1990). Identifying independence in Bayesian networks. *Networks*, *20*(5), 507–534.

Girin, L., Leglaive, S., Bie, X., Diard, J., Hueber, T., & Alameda-Pineda, X. (2021). Dynamical variational autoencoders: A comprehensive review. *Foundations and Trends in Machine Learning*, *15*(1–2), 1–175.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In *Advances in neural information processing systems*: *vol. 27*.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., et al. (2016). Beta-vae: Learning basic visual concepts with a constrained variational framework.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, *14*(8), 1771–1800.

Hori, C., Alamri, H., Wang, J., Wichern, G., Hori, T., Cherian, A., et al. (2019). End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *IEEE international conference on acoustics, speech and signal processing* (pp. 2352–2356). IEEE.

Hou, X., Sun, K., Shen, L., & Qiu, G. (2019). Improving variational autoencoder with deep feature consistent and generative adversarial training. *Neurocomputing*, *341*, 183–194.

Hsu, W.-N., & Glass, J. (2018). Disentangling by partitioning: A representation learning framework for multimodal sensory data. arXiv preprint arXiv:1805.11264.

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.

Kim, H., & Mnih, A. (2018). Disentangling by factorising. In *International conference on machine learning* (pp. 2649–2658). PMLR.

Kim, J. W., Salamon, J., Li, P., & Bello, J. P. (2018). Crepe: A convolutional representation for pitch estimation. In *2018 IEEE international conference on acoustics, speech and signal processing* (pp. 161–165). IEEE.

Kim, D., & Song, B. C. (2022). Optimal transport-based identity matching for identity-invariant facial expression recognition. In *Advances in neural information processing systems*.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International conference on learning representations*.

Kingma, D., & Welling, M. (2014). Auto-encoding variational bayes. In *International conference on learning representations*.

Klys, J., Snell, J., & Zemel, R. (2018). Learning latent subspaces in variational autoencoders. In *Advances in neural information processing systems*: *vol. 31*.

Larsen, A. B. L., Sønderby, S. r. K., Larochelle, H., & Winther, O. (2016). Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning* (pp. 1558–1566). PMLR.

Lazarus, A. A. (1976). Multimodal therapy. In *Handbook of psychotherapy integration* (p. 105).

Le Roux, J., Wisdom, S., Erdogan, H., & Hershey, J. R. (2019). SDR–half-baked or well done? In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 626–630). IEEE.

Lee, M., & Pavlovic, V. (2020). Private-shared disentangled multimodal vae for learning of hybrid latent representations. arXiv preprint arXiv:2012.13024.

Li, Y., & Mandt, S. (2018). Disentangled sequential autoencoder. arXiv preprint arXiv:1803.02991.

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*, *13*(5), Article e0196391.

Lo, C.-C., Fu, S.-W., Huang, W.-C., Wang, X., Yamagishi, J., Tsao, Y., et al. (2019). Mosnet: Deep learning based objective assessment for voice conversion. arXiv preprint arXiv:1904.08352.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., et al. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In *International conference on machine learning* (pp. 4114–4124). PMLR.

Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., & Tschannen, M. (2020). Weakly-supervised disentanglement without compromises. In *International conference on machine learning* (pp. 6348–6359). PMLR.

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.

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.

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.

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.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *IEEE international conference on acoustics, speech and signal processing* (pp. 5206–5210). IEEE.

Pepino, L., Riera, P., & Ferrer, L. (2021). Emotion recognition from speech using wav2vec 2.0 embeddings. *Interspeech*, 3400–3404.

Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., & Pantic, M. (2018). End-to-end audiovisual speech recognition. In *IEEE international conference on acoustics, speech and signal processing* (pp. 6548–6552). IEEE.

Pham, L., Vu, T. H., & Tran, T. A. (2021). Facial expression recognition using residual masking network. In *International conference on pattern recognition* (pp. 4513–4519). IEEE.

Pihlgren, G. G., Sandin, F., & Liwicki, M. (2020). Improving image autoencoder embeddings with perceptual loss. In *2020 international joint conference on neural networks* (pp. 1–7). IEEE.

Ramachandram, D., & Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, *34*(6), 96–108.

Razavi, A., Van den Oord, A., & Vinyals, O. (2019). Generating diverse high-fidelity images with VQ-VAE-2. In *Advances in neural information processing systems*: *vol. 32*.

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning* (pp. 1278–1286). PMLR.

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. In *IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221), vol. 2* (pp. 749–752). IEEE.

Roth, J., Chaudhuri, S., Klejch, O., Marvin, R., Gallagher, A., Kaver, L., et al. (2020). Ava active speaker: An audio-visual dataset for active speaker detection. In *IEEE international conference on acoustics, speech and signal processing* (pp. 4492–4496). IEEE.

Sadok, S., Leglaive, S., Girin, L., Alameda-Pineda, X., & Séguier, R. (2023). Learning and controlling the source-filter representation of speech with a variational autoencoder. *Speech Communication*, *148*, 53–65.

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). Wav2vec: Unsupervised pre-training for speech recognition. *Interspeech*, 3465–3469.

Schoneveld, L., Othmani, A., & Abdelkawy, H. (2021). Leveraging recent advances in deep learning for audio-visual emotion recognition. *Pattern Recognition Letters*, *146*, 1–7.

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*.

Sutter, T., Daunhawer, I., & Vogt, J. (2020). Multimodal generative learning utilizing Jensen-Shannon-divergence. *Advances in Neural Information Processing Systems*, *33*, 6100–6110.

Sutter, T. M., Daunhawer, I., & Vogt, J. E. (2021). Generalized multimodal ELBO. In *International conference on learning representations*.

Suzuki, M., & Matsuo, Y. (2022). A survey of multimodal deep generative models. *Advanced Robotics*, *36*(5–6), 261–278.

Suzuki, M., Nakayama, K., & Matsuo, Y. (2016). Joint multimodal learning with deep generative models. arXiv preprint arXiv:1611.01891.

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2010). A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *IEEE international conference on acoustics, speech and signal processing* (pp. 4214–4217). IEEE.

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. In *Proceedings of the conference. Association for computational linguistics. Meeting, vol. 2019* (p. 6558). NIH Public Access.

Vahdat, A., & Kautz, J. (2020). NVAE: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, *33*, 19667–19679.

Van Den Oord, A., Vinyals, O., et al. (2017). Neural discrete representation learning. In *Advances in neural information processing systems*: *vol. 30*.

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*.

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*, *13*(4), 600–612.

Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., et al. (2020). Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European conference on computer vision* (pp. 700–717). Springer.

Wei, X., Li, H., Sun, J., & Chen, L. (2018). Unsupervised domain adaptation with regularized optimal transport for multimodal 2D+ 3D facial expression recognition. In *IEEE international conference on automatic face & gesture recognition* (pp. 31–37). IEEE.

Wu, M., & Goodman, N. (2018). Multimodal generative models for scalable weakly-supervised learning. *Advances in Neural Information Processing Systems*, *31*.

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*, Article e12.

Zhao, Z., Liu, Q., & Zhou, F. (2021). Robust lightweight facial expression recognition network with label distribution training. In *Conference on artificial intelligence, vol. 35, no. 4* (pp. 3510–3519).