Hierarchical Social Processes

Stochastic Meta-learning of Group and Individual-level Style

Bilal El Attar

Hierarchical Social Processes

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Bilal El Attar

TUDelft

Socially Perceptive Computing Lab
Pattern Recognition & Bioinformatics Section
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
www.ewi.tudelft.nl

Hierarchical Social Processes

Author: Bilal El Attar
Student id: 4718127
Email: bilal.attar@gmail.com

Abstract

How people behave in social interactions is influenced by a multitude of factors. A large part of human communication is embedded within non-verbal communication. This type of communication is sent throughout social signals, that are embodied within low-level social cues (e.g. gaze, posture, gestures). In order for intelligent systems to seamlessly interact with humans, they need to possess some form of social intelligence. That includes expressing and recognising social signals. The field of social cue forecasting intends to predict low-level behavioral cues within social interactions, allowing systems to adapt their behavior according to the forecasted behavior of interlocutors, or synthesize human behavior on the basis of the prediction. Within social science theory, it has been established that these behavioral cues are dependent on social context, as well as individual idiosyncrasies. Under earlier work within human behavior synthesis, the latter has been mostly used, and referred to as 'style'. This work attempts to broaden the traditional view of style and proposes a model for incorporating both group, and individual-level style using a hierarchical latent variable model. To adapt to unseen groups, we incorporate this hierarchical latent structure into a meta-learning model. Introducing the hierarchical neural processes and social processes models. After testing these models on a real-world dataset containing triadic interactions, it turns out that most models fail due to posterior collapse. This prevents them from learning a useful latent representation containing semantic information with respect to forecasting future sequences of social cues. To combat this, a constant weight was assigned to a part of the loss term. However, as the issue still persists, it leaves us unable to prove whether our proposed method improves upon the baseline approach. Therefore, future work on posterior collapse in neural processes models is needed.

Thesis Committee:

Chair & Thesis Advisor:	Prof. H. Hung, Faculty EEMCS, TU Delft
Daily-supervisor:	C. Raman, Faculty EEMCS, TU Delft
Committee Member:	Prof. C. R. M. M. Oertel, Faculty EEMCS, TU Delft
Committee Member:	Prof. X. Zhang, Faculty EEMCS, TU Delft

Preface

During the first period of my master's, it was unclear to me what subject would be interesting to study for an extended period of time. However, during most courses the themes that captivated me the most were in the fields of human-computer interaction and machine learning. After orientating with Chirag about possible thesis subjects, I started reading about a variety of research fields and concepts, which got excited to work on the topic of human behavior forecasting. After months of working and learning about the topic, as well as related subjects within the research group, I am proud to say that it was a rewarding journey. Even though I was not well acquainted with the topic at first, I felt like I gained a lot of knowledge in such a short time span, both in the technical sense, as well as in the sense of designing a research project. Therefore, I would consider this to be the most fruitful experience during my entire study career.

Before you lies my thesis, as part of my final examination for the Computer Science Master's program. I hope you will enjoy reading it.

Bilal El Attar
Delft, the Netherlands
December 8, 2022

Acknowledgements

First, I would like to thank my co-daily supervisor; Chirag, for providing valuable feedback, and always nudging me in the right direction. From the start, his enthusiasm about the subject and overall positive attitude motivated me throughout the project. During numerous meetings, we were able to discuss a lot about the study in detail, and I appreciate the commitment. He was always available for asking questions and providing guidance when needed.

Next, I want to thank Hayley for being a great thesis advisor. She often challenged me to think more critically about the work, and to not forget the overarching goal of the research when being delved into the technical aspects. During the weekly meetings, we often had useful discussions about the relevant concepts and how this translates to valuable insights.

Also, I want to thank the members of the Socially Perceptive Computing Lab for the weekly lab meetings and reading groups, where I was able to learn quite a lot from a variety of interesting topics.

Last, I want to thank my family and friends for their support during this period of time.

Contents

Chapter 1

Introduction

In the last decade, there has been an increase in the interest in intelligent systems. However, these systems are often socially ignorant, meaning that they hardly possess any form of social intelligence [47]. Contemporary systems have the goal of interacting with humans. However, They often have a simplistic view of the conversation, and ”might not account for the fact that human-human communication is always socially situated” [47]. It has been established that a part of intelligence within humans, is ”the ability to express and recognise social signals and social behaviours like turn-taking, agreement, politeness, and empathy, coupled with the ability to manage them in order to get along well with others while winning their cooperation” [47]. These social signals and social behaviors are often embodied within a variety of low-level non-verbal behavioural cues (e.g. body posture, motion, facial expressions, prosodics) [47]. These low-level cues have been the center of multiple studies on generating [24, 50, 2, 50, 30, 12] and forecasting non-verbal human behavior [1, 43, 2, 38, 8]. Within social interactions, humans form their non-verbal human behavior patterns on the basis of interpersonal dynamics [47, 32, 2], amongst other factors. One example of this is turn-taking. where participants of a conversation tend to coordinate their speech by predicting if a turn-taking event is close [32, 15].

The field of non-verbal human behavior forecasting is concerned with predicting these types of low-level cues. Within this area, researchers have proposed methods to predict non-verbal behavior for multiple use cases, including autonomous vehicles, intelligent robots, and multimedia [32]. However, another direction within this field is predicting human motion within social interactions [24, 1, 38], also known as social cue forecasting. Forecasting non-verbal behavior within that setting allows us to reason about, and generate social cues for social intelligent systems while dealing with the uncertainty of future context. However, one thing to note is that human motion is a complex problem, meaning that it is influenced by a variety of factors. Some important factors that studies have shown to have an effect on gesture production, are social context [33, 19, 22], as well as certain personal, idiosyncratic factors such as age [3], and personality [22]. The concept of gestures being idiosyncratic has been studied within various synthesis tasks [50, 2] and has been put under the term ’style’. However, the proposed concepts of style seem limited and only try to capture the individual aspect of it. We argue that in addition to this individual-level style, we can also speak about group-level style, which would be inherent to the social context of the social interaction.

Having this knowledge about style within a conversation might benefit a prediction model by forecasting behaviors that are more in line with the context of the specific interaction, as well being more in line with individual manifested behaviors of a participants within the interaction (idiosyncrasies).
The goal of this thesis is to propose a method that models style within a latent variable model for the task of human motion forecasting within social interactions, on both an individual, as well as a group level. To try and incorporate this into the model, we plan on utilizing some theoretical aspects about style, together with various state-of-the-art machine learning techniques.

1.1 Research Questions

The main research question we plan on answering within this thesis is:

How can we learn group and individual-level style within a latent variable model, for the task of social cue forecasting?
The main research goal of this thesis is to design an approach that learns style in the context of social cue forecasting, using a latent variable model. A latent variable model assumes the data to be generated from unobserved/hidden variables. Modeling style within the model would help us understand the generation process for social cues, as well as might help the accuracy of the prediction model. In order to learn style, we propose the distinction between group, and individual-level style. From this main research question, we follow up with the following sub questions and hypotheses:

How does the introduction of group, and individual-level style impact the predictions of the model?
To answer this question, we plan on evaluating the outcomes both qualitatively, and quantitatively on a baseline, and proposed model. Our hypothesis is that:

The introduction of group and individual-level style benefits the accuracy and learning ability of the model.
Meaning that the model should be able to produce outputs closer to the ground truth, while also performing better with respect to the loss function used while training the model. This hypothesis is motivated by previous work, showing improved performance and log-likelihood when introducing a structure within latent variable models [40, 42, 44]. Another sub question we plan on answering is:

How is group and individual-level style represented within the latent space?
To answer this question we plan on generating a lower dimensional representation of all latent vectors and comparing them between groups, as well as between individuals within every interaction . Our hypothesis is that:

Group-level latent variables generated from the same group cluster together

This hypothesis is formed on the intuition about group-level style being different between all conversations, while latent variables generated from the same interaction would be similar or equal. Due to the fact that the social context remains similar within the same conversation.

1.2 Contributions

In this thesis, we hypothesize that situated interactions have both group dynamic and individual dynamics that can be leveraged in order to improve the performance of social cue forecasting models. We present a hierarchical latent variable model within a meta-learning framework that incorporates the use of style for the first time in a social cue forecasting setting. Therefore, the contribution of our work lies both in the area of social cue forecasting, as well as in the probabilistic machine learning field due to our proposed novel architecture.

1.3 Research Context

This study was conducted at the Delft University of Technology Socially Perceptive Computing Lab and can be interpreted as an extension of the work of Raman et al. [38]. The dataset that was used, was made publicly available by Carnegie Mellon University [24].

1.4 Outline

The thesis is organized as follows:

- Chapter 2 presents the social science background and a review of related work on non-verbal behavior synthesis, forecasting, and style.
- Chapter 3 provides an overview of the technical details and concepts that underpin our proposed method and experiments.
- In Chapter 4, we describe our methodology for learning group and individual-level style from conversations.
- Chapter 5 discusses the experiments we conducted to test our proposed method, answer the appropriate research questions, and validate our hypotheses.
- Chapter 6 presents and analyzes the results of our experiments.
- Chapter 7 discusses the limitations of our work, general phenomenons visible within the results of our model, and various ethical considerations.
- Finally, Chapter 8 concludes our work and offers directions for future research.

Chapter 2

Background & Related Work

Although a large part of human communication is embedded within verbal or textual information, it does not provide a full picture of communication [48]. Human communication is multimodal by its nature. This means that individuals make use of a combination of multiple modalities (e.g. gestures, prosodics, gaze, facial expressions, verbal language) exhibiting social signals in order to convey information to others within a social interaction. This chapter looks at types of non-verbal human behavior and style from a social science point of view, before discussing the concept and previous work in the fields of social cue forecasting and synthesis.

2.1 Social Cues and Signals

Even though a real definition of social signals is hard to define, Vinciarelli and Pentland [46] note most definitions used within the literature agree on one point: ”social signals are observable behaviours that produce, intentionally or not, tangible changes in others, whether this means modifying their inner state (e.g., to stimulate the emotions they experience), to modify their observable behaviour (e.g., to make them laugh in response to a joke) or to change their beliefs about the social setting (e.g., to make them aware of conflict or disagreement)”. This makes it evident that these cues serve multiple functions in social interactions. Within figure 2.1, the most common functions that social cues contribute to are shown [46]. Note that the lines between these cues and their assigned function are not one-to-one relationships, meaning that multiple cues or even combinations of cues can serve one, or multiple functions.

Most of these social cues are exhibited automatically without the participants being aware of showing them, leaving us with ”little insight into its critical role in interactions” [37]. And although the critical role of non-verbal communication (e.g. social cues) is undeniable [48, 37], the inner workings are still poorly understood, ”making it hard to formalize rules about how to understand and use social signals” [24]. Even though we know little about the production of social cues, studies in the past have shown that the anticipation of these social cues is important in certain areas within social interactions such as turn-taking [15, 25], where participants coordinate their speech by predicting when a turn-taking event is close, based on the behavior of their interlocutors [32]. This concept of anticipation and the overarching interpersonal factors that drive social interactions are related to the concept of adaptation.

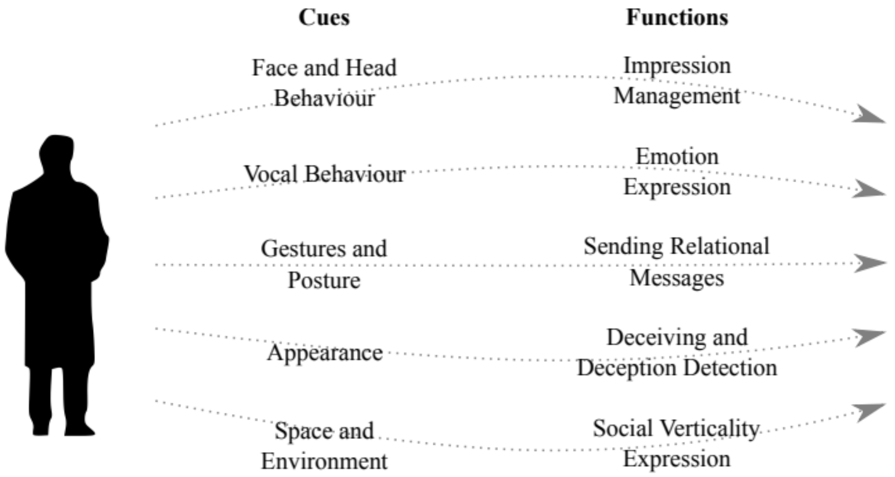


Figure 2.1: Non-verbal social cues and some functions that these contribute to. Taken from Vinciarelli and Pentland [46]

2.2 Adaptation

Adaptation undergirds social organization and is coordinated through verbal, as well as non-verbal communication channels [7]. Within social science, various forms of adaptation in social interactions have been identified and studied. These include mirroring; the phenomenon of two actors displaying identical visual signals (e.g. both sitting with their legs crossed), convergence; the concept of one actor's (non)verbal behaviour becoming more like another's over time, and synchrony; "the degree to which behaviors in an interaction are synchronized in both timing and form" [7]. The key takeaway from existing research on these forms of adaptation is that individual behavior in social interactions is influenced by the behavior of others, specifically the non-verbal social cues of each participant within a conversation.

2.3 Style

Social cues exhibited by individuals within a social interaction are dependent on social context [33, 19, 22]. Consider an individual interacting with his/her friends, as opposed to a formal setting. The intuition here is that the individual would exhibit social cues differently between the two situations. As with the use of voice, it has been shown that pitch, speech rate and voice quality help in differentiating politeness levels [9]. As the underlying concept, we can say that social cues in this case change due to the social context of the interaction (an environment where politeness may or may not be desired). Social context in this case refers to the setting in which the interaction takes place, meaning the physical environment, as well as the social environment. Although determining an exact definition of social context might be hard, it has been shown that multiple factors that are fundamental to the concept of social context are relevant for gesture production [19, 22] and language processing [33]. Some of these factors include culture [29, 17], social information (social cues exhibited by conversation partners) [33], and verbal content [20].

Additional to the notion of social cues being dependent on the social context, social cues are also idiosyncratic [35]. Meaning that each individual has their own way of emitting social signals within interactions. This variability between people has been studied within previous work on gesture production, and has been attributed to various factors such as age [3], verbal and spatial skills [21], and personality [22]. In previous work on synthesizing non-verbal human behavior, as discussed later on in this chapter, this is referred to as an individual's style. However, we reason that in addition to this individual style, one could also determine a group-level style within situated social interactions, based on the social context. The intuition is that as the social context between different groups changes (e.g. due to personal relations, or a change of subjects), the participants will exhibit different social cues. Following from the example described earlier, a person might change some of their social cues and use of voice depending on the formality of the interaction [9]. Although an exact definition of style is hard to describe, one of the goals of this thesis is to see what the style variable encompasses.

2.4 Modeling Non-verbal Human Behavior

As stated before, social cues serve an essential role in social interactions. By modeling these social cues, it is possible to both form policies within social systems that reason on future non-verbal behavior, as well as synthesize this behavior. Consider the case of a social robot that could anticipate whenever a turn-taking event is approaching within a social interaction. The robot would be able to change its behavior based on that anticipation, and instead, wait, or throw some other line of dialog depending on the prediction. An example of this is a study by Bohus and Horvitz [5], where they make use of a forecasting model in order to predict the (dis)engagement of subjects when interacting with a physically situated dialog system. They propose an engagement policy that takes the forecasted behavior into account. Figure 2.2 shows the formulated policy depending on $P_{Disengage}$; the probability of the conversation partner disengaging. According to the different thresholds on this probability, the system decides what corresponding action to take.

In addition to forming policies, a social robot could synthesize social signals, for example by using social cues in order to signal a turn change in advance. Summarizing these two tasks, the modeling of behavior is often divided into two related areas: non-verbal human behavior forecasting, and synthesis. These two fields are fairly similar, and one could argue that non-verbal human behavior forecasting encompasses non-verbal human behavior synthesis as well. However, strictly speaking, the difference between the two is that synthesis tasks generally aim to synthesize the social cues of a target person (or multiple) within a certain time window, by using some modality (e.g. speech) of the avatar itself or social cues exhibited by the other participants in the interaction within the same time window. In contrast, work on forecasting typically focuses on predicting social cues of multiple individuals (or sometimes a single individual) within a time window in the future. For this purpose, social cues of the entire group within a previous time window are used. State-of-the-art research uses deep learning techniques in order to predict, or synthesize a sequence of social cues (e.g. gestures [2], upper body orientation, location and speaking status [38]). As the two tasks are closely related, some general methods and theory could be useful for both task formulations.

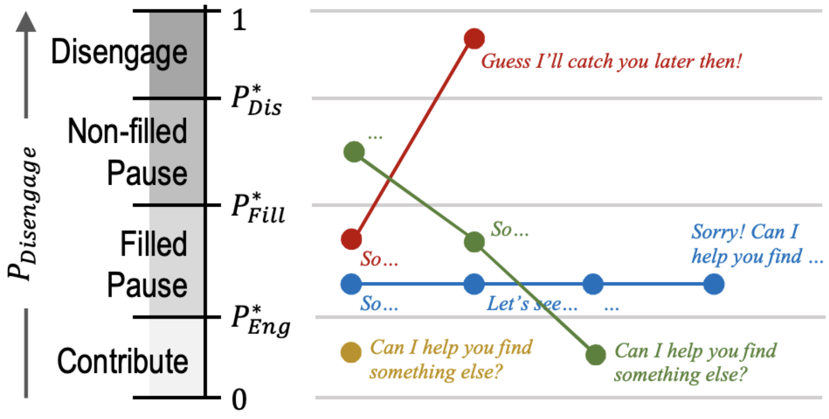

1
Disengage
P^*_{Dis}
Non-filled Pause
P^*_{Fill}
Filled Pause
P^*_{Eng}
Contribute
0
$P_{Disengage}$
Guess I'll catch you later then!
...
So...
So...
Sorry! Can I help you find ...
So...
Let's see...
...
Can I help you find something else?
Can I help you find something else?

Figure 2.2: Disengagement policy with hesitation actions, from Bohus and Horvitz [5]

2.5 Related Work

The previous section provided an overview of the background and social science concepts relevant to social cue forecasting. However, as the goal of this project is to formulate a model for learning style, it is useful to review existing approaches proposed in the literature. Within this section, different work on non-verbal human behavior and synthesis is examined. In addition, we review existing work on non-verbal human behavior synthesis that incorporate the concept of style.

2.5.1 Non-verbal Human Behavior Synthesis

As established earlier, the main difference between non-verbal human behavior synthesis and forecasting lies within the task description. While forecasting tasks generally aim to predict social cues of every individual within an interaction, given only previous sequences of that same group, synthesis tasks often focus on a single person, given some modalities within the same timeframe. For example, a study by Ferstl et al. [12] on co-speech gestures proposes a gesture generator as part of their architecture. This generator translates speech to gestures by using prosodics as input, outputting 21 upper-body keypoints. One interesting concept within this work is that they utilize the theory of gesture phases. This concept has been described in early work examining gestures [30], and is included in some form within their model. This idea of a social-science-grounded method, by including some prior knowledge of non-verbal behavior turned out to be important in generating gestures that are in line with the theory of gesture phases [12].

2.5.2 Incorporating Style

Social cues exhibited within an interaction are both dependent on social context, as well as idiosyncratic factors. The concept of gestures being idiosyncratic has been used within various works within the field of gesture synthesis by including or learning a style variable [2, 50, 16]. The motivation behind including this style variable often lies behind the idea of style transfer. In which a model would be capable of ”generating gestures for a speaking agent ‘A’ in the gesturing style of a target speaker ‘B’” [2]. Ahuja et al. [2] propose a model, named Mix-stAGE, that takes an audio signal as input and aims to perform two tasks: style preservation, as well as style transfer. The goal of style preservation is to take each individual’s style into account, while generating gestures for multiple people. Although the concept of style transfer does not seem appropriate for forecasting non-verbal behavior, the concept of style preservation could prove to be useful. A forecasting model that is able to gain knowledge of people’s individual styles within the conversation, would also be able to predict more fitting behavior that is in line with the style of that individual. This intuition forms the basis of including this in a forecasting model.

In a study by Yoon et al. [50], a gesture generation model is proposed based on the trimodal context of text, audio, and speaker identity. The speaker identity is used in this case to provide a particular style to the generation model. To provide such a style variable, a style embedding space is learned from speaker identities. By sampling from this space, it is possible to generate gestures with a different style that was not necessarily part of the dataset. The resulting embedding space is shown in Figure 2.3. Although this seems a good way to show the versatility of style and what style encompasses in this sense, it presents a ’simple’ decomposition of the concept. Therefore, we can define a few limitations. First of all, as with most work on style, the study assumes that style does not change in the short term, so the same style embedding is used throughout the entire synthesis. However, in a study by Wells [49] on gestures within a classroom environment, they found that ”student gestures grew in size and became more animated as their confidence in their utterances increased”. Suggesting that individual style can evolve over the short term. Additionally, this concept of style only considers the individual level, whereas we can also define a group-level style that is based on social context. Finally, existing work does not take the behavior of interlocutors into account for deciding this style variable, which is also due to the fact that most work focuses on monologues instead of interactions.

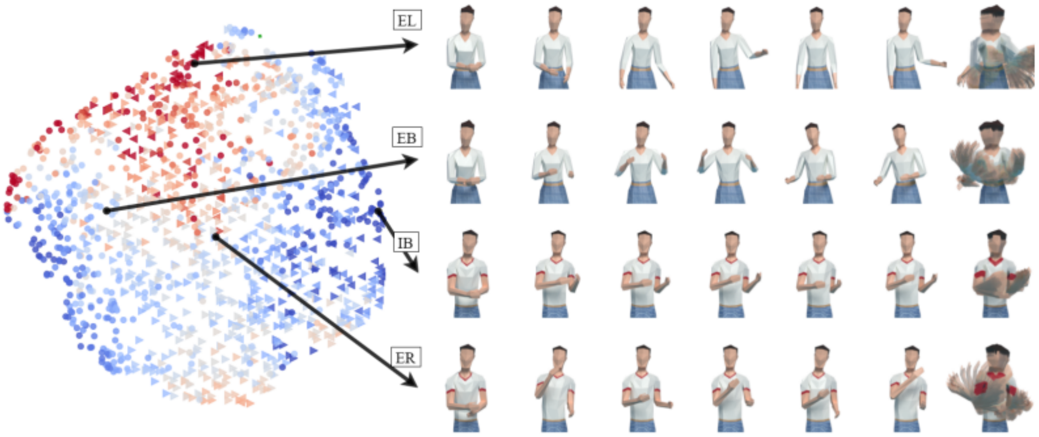


Figure 2.3: The resulting embedding space and generated samples with their associated gesturing style (generated from the same speech sample), visualized in two dimensions using UMAP [34]. E stands for extrovert, while I stands for introvert. R, L, and B stand for left, right and both respectfully. ”The points represent degrees of motion variance via color and degree of handedness by its marker types”. Image from Yoon et al. [50]

2.5.3 Non-verbal Human Behavior Forecasting

Previous work in human motion forecasting has focused on predicting human motion for multiple use cases, including autonomous vehicles, intelligent robots, and multimedia [32]. However, another direction within this field is predicting human motion within social interactions [24, 1, 38]. As told within the previous section, forecasting human motion within that setting allows us to reason about, and generate human motion for socially intelligent systems while dealing with the uncertainty of future context.

There have been a variety of models proposed within earlier work on forecasting social cues. One of these, is an interesting paper by Ahuja et al. [1] that tries to predict non-verbal behavior of an avatar in a dyadic setting. Within their task formulation, they divide the input features into two categories: interpersonal, and intrapersonal dynamics. Interpersonal dynamics is defined as the behavior of the interlocutor within the previous timesteps, while intrapersonal dynamics represents the behavior of the avatar within these previous timesteps. The model is trained on a dataset consisting of one person who interacts with 11 different participants for around 1 hour each, this one person being the same across every conversation. Referring to the social theory behind gestures being idiosyncratic, this might be a limitation. Therefore, it might perform well when predicting gestures for the individual considered within the dataset, but considering a setting where the system runs on unseen data samples, it might be less accurate.

More recently, a context-aware forecasting model for dyadic interactions has been proposed by Tuyen and Celiktutan [43]. They make use of a generative adversarial network

(GAN) [18]. This is one approach of generative modeling where the model exists of two parts; a discriminator and a generator network. The task of the generator is to generate samples that conform to the original data distribution, while the discriminator's task is to differentiate between fake, generated samples, and real samples from the data. During a min-max game of these two parts, the model tries to optimize both goals. However, as the task in this model is to predict future sequences, generating samples using a GAN has to be conditioned on a previously observed sequence of behavior. Therefore, they make use of the cGAN model [36], which is inherently a variation of the original GAN model with the use of a conditional input for the generation task. Furthermore, context-aware in the context of this research means taking interpersonal dynamics into account (features of the interlocutor). Similar to the work of Ahuja et al. [2], they found including these features to be important for generating accurate results. However, like the work of Ahuja et al. [2], the model might have problems in adhering to the idiosyncratic traits in gesturing behavior when being presented with a sample outside the dataset. Although one thing they improved upon was using a varied dataset containing multiple participants. Besides this, the model solely focuses on dyadic conversations.

In addition to some of the shortcomings described, the discussed models assume one correct deterministic future for the individual(s) within the interaction. However, observed sequences can have multiple correct futures. For example, ”a window of overlapping speech between people may and may not result in a change of speaker” [38, 11]. There are a few works that try to overcome this assumption of a single, deterministic future by considering a distribution of futures [38, 8]. A study by Raman et al. [38], which this work builds upon, predicts a distribution of futures, and also tries to tackle the issue of generalizing to unseen data samples. For the latter, they take on the task of social cue forecasting task using a meta-learning approach. The idea of meta-learning within this context is to adapt to unseen supervised tasks by learning how to learn from a dataset. Adopting a meta-learning approach leaves us with a way to infer knowledge from samples at test time (see Chapter 3: Preliminaries, for an in-depth explanation of meta-learning). In addition to this, the model is agnostic of the set of features and number of individuals within the interactions of a particular input dataset. However, as all forecasting work, it does not incorporate any notion of style. Therefore, we propose one of the first forecasting models that tries to learn both group-level, as well as individual-level style.

Chapter 3

Preliminaries

Before explaining the proposed method, we need to establish some required knowledge that forms the basis of our approach. We start off with formalizing the social cue forecasting (SCF) task and the theory of latent variables, before going into the specifics of the models used within this study: the variational autoencoder [28], the neural processes [14], and the social processes model [38].

3.1 Social Cue Forecasting

As described in work by Raman et al. [38], the goal of SCF is to predict a future sequence of behavioral cues of every person involved in a social interaction based on an observed sequence of their behavioral cues. Formally defined, a window of increasing observed timesteps is denoted as $\boldsymbol{t}_{\text{obs}} = [o1, o2, ..., oT]$, and an unobserved future timewindow can be denoted as $\boldsymbol{t}_{\text{fut}} = [f1, f2, ..., fT]$, the only condition being $f1 > oT$. This means that a delay between the observed and future time window could be possible. We use these timewindows to define the behavioral cues of all participants in the conversation over $\boldsymbol{t}_{\text{obs}}$and $\boldsymbol{t}_{\text{fut}}$:

$$Y = [\boldsymbol{b}_t^i; t \in \boldsymbol{t}_{\text{fut}}]_{i=1}^n, X = [\boldsymbol{b}_t^i; t \in \boldsymbol{t}_{\text{obs}}]_{i=1}^n \tag{3.1}$$

Where $\boldsymbol{b}_t^i$ is the set of behavioral cues of person i at timestep t. We try to predict a distribution over futures. This is motivated by the fact that an observed sequence often has multiple 'correct' or probable outputs. In addition, this is also useful for expressing uncertainty over the predicted outcomes. In short, the goal is to model the distribution $p(\boldsymbol{Y}|\boldsymbol{X})$.

3.2 Latent Variable Models

Contemporary work often makes use of latent variable models for learning a probability distribution for generative models. The goal of a generative model is to learn the distribution of the original data, in order to generate new convincing data samples. The main premise within latent variable models is that unknown hidden/unobserved/latent variables are assumed to generate the observed data. These latent variables are abstract numerical

values, that typically do not contain an easily interpretable meaning with regard to generating observed data.
Strictly speaking, the goal of a latent variable model is to model the distribution of the data $\boldsymbol{X}$: $p(\boldsymbol{X})$. Before we dive into an explanation of how a latent variable model tries to accomplish this, we define the following probability distributions:

- The prior distribution: $p(z)$. This distribution models how z, the latent variables, behave.
- The likelihood: $p(\boldsymbol{X}|z)$, which defines how every latent variable translates to a data sample $\boldsymbol{x}$ (where $\boldsymbol{X}$ is the entire set of data points).
- The posterior distribution: $p(z|\boldsymbol{X})$. This distribution describes how latent variables are generated from a data point.

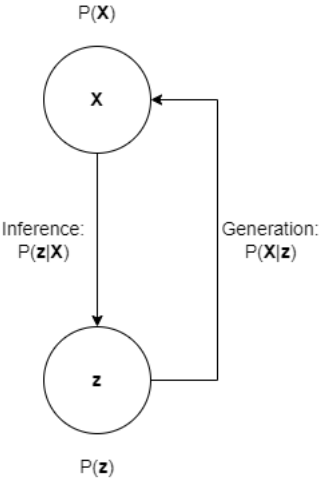


Figure 3.1: Generation and inference within a latent variable model

Within a latent variable model, we can divide the objective within two parts: inference, and generation. Inference refers to the part of finding a useful representation of the observed sequence into a latent vector z. This is formulated by the posterior $p(z|\boldsymbol{X})$. While generation in this case refers to the process of generating a future sequence from this latent vector z. This in turn is represented by the likelihood $p(\boldsymbol{X}|z)$. Figure 3.1 shows a graph of a general latent variable model and the two objectives. Inference might be a hard problem however, since it involves calculating the posterior $p(z|\boldsymbol{X})$. According to Bayes rule, we can rewrite this probability as follows:

$$p(z|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|z)p(z)}{p(\boldsymbol{X})} \tag{3.2}$$

The term in the denominator $p(\boldsymbol{X})$, is also called the evidence. By marginalizing the latent variables we rewrite the evidence as the following:

$$p(\boldsymbol{X}) = \int p(\boldsymbol{X}|z)p(z)dz \tag{3.3}$$

In order to compute the evidence $p(\boldsymbol{X})$, we need to evaluate all possible values of z, which would require exponential time, making this computation intractable. To overcome this, we approximate this posterior using a method called variational inference (VI) [4]. The general idea behind VI is to find the best approximation $q^*(z|\boldsymbol{X})$ of the true posterior $p(z|\boldsymbol{X})$ from a tractable family of distributions Q. This is done using the Kullback-Leibler (KL) divergence, which is a measure of similarity between two probability distributions. The formulation of the KL divergence is as follows:

$$KL[q(\boldsymbol{X})||p(\boldsymbol{X})] = -E_{q(\boldsymbol{X})}[log\frac{p(\boldsymbol{X})}{q(\boldsymbol{X})}] \tag{3.4}$$

We can use the KL divergence to formulate the objective of VI; finding an approximation $q^*(\boldsymbol{z}|\boldsymbol{X})$ that minimizes the KL divergence between the approximation and the true posterior. This is formulated as follows:

$$q^*(\boldsymbol{z}|\boldsymbol{X}) = \underset{q(\boldsymbol{z}|\boldsymbol{X}) \in Q}{\arg\min} KL[q(\boldsymbol{z}|\boldsymbol{X})||p(\boldsymbol{z}|\boldsymbol{X})] \tag{3.5}$$

Using this approximation $q(\boldsymbol{z}|\boldsymbol{X})$, or variational posterior as the approximation is also called, it still seems like we need the posterior $p(\boldsymbol{z}|\boldsymbol{X})$. However, we can derive an alternative objective from equation 3.5 that is easier to optimize, called the Evidence Lower Bound (ELBO):

$$ELBO(q(\boldsymbol{z}|\boldsymbol{X})) = E_{q(\boldsymbol{z}|\boldsymbol{X})}[\log p(\boldsymbol{X}|\boldsymbol{z})] - KL[q(\boldsymbol{z}|\boldsymbol{X})||p(\boldsymbol{z})] \tag{3.6}$$

Maximizing the ELBO is equivalent to minimizing the KL divergence between the approximation and true posterior. In addition to that, it also serves as a lower bound on the evidence $p(\boldsymbol{X})$, which explains the name. In short, by optimizing the ELBO, we can find an approximation closer to the true posterior, while also optimizing the log-likelihood of the evidence $\log p(\boldsymbol{X})$, thereby modeling the distribution $p(\boldsymbol{X})$.

3.3 Variational Autoencoder

One type of generative models is the variational autoencoder. As summarized in Figure 3.2, this is a neural network architecture that consists of two parts; a probabilistic encoder and a probabilistic decoder. In the case of social cue forecasting, the encoder takes a previously observed sequence as input, which it maps into a latent distribution. From this distribution, a latent vector is sampled. The decoder maps this latent vector into a future sequence of behaviors. As this is essentially a latent vector model, the model is trained using variational inference. The encoder represents the approximation of the posterior: $q_*(\boldsymbol{z}|\boldsymbol{X})$, while the conditional $p(\boldsymbol{Y}|\boldsymbol{z})$ is represented by the decoder. This type of model is trained using amortized variational inference, meaning that the model learns a single function (parameterized by a neural network) that maps the data and posteriors, in order to maximize the ELBO as formulated in Equation 3.6.

3.4 Neural Processes

A neural network-based model ”that uses latent variables $\boldsymbol{z}$ and data $\boldsymbol{X}$ as inputs; $f(\boldsymbol{X}, \boldsymbol{z})$ can be considered a neural latent variable model” [13]. Neural processes (NPs) is a class of neural latent variable models proposed by Garnelo et al. [14, 26]. This class of models makes use of the idea of meta-learning and is capable of estimating an uncertainty over predictions.

3.4.1 Meta-learning

Typical supervised learning algorithms are trained to model a function mapping an observation $\boldsymbol{x}$ to a predicted output: $f(\boldsymbol{x})$. This would be trained on a dataset $C = (\boldsymbol{X}_C, \boldsymbol{Y}_C)$, which

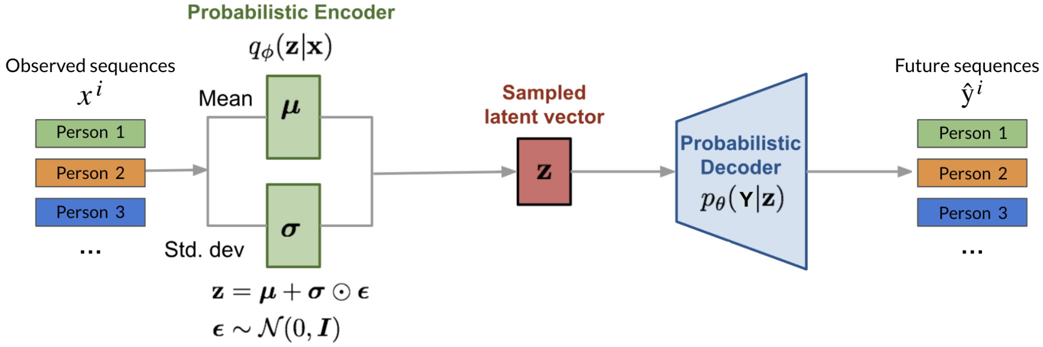


Figure 3.2: Variational autoencoder architecture. Figure is partly from Lil'Log

we can rewrite as $C = \{(\boldsymbol{x}^i, \boldsymbol{y}^i)_{i=1}^N\}$ for N individual input and output pairs. We refer to this set as the context set. At test time, the model is typically run on an unseen target set; $T = (\boldsymbol{X}_T, \boldsymbol{Y}_T) = \{(\boldsymbol{x}^i, \boldsymbol{y}^i)_{i=1}^K\}$, where we run $f(\boldsymbol{x}^t)$ for every target input in the set in order to get a prediction that is compared to the ground truth $\boldsymbol{y}^t$, also called the target output.
The concept of meta-learning is to adapt to unseen supervised tasks by learning how to learn from a dataset. This leaves us with a way to infer knowledge from samples at test time. Formally defined, instead of learning a predictor $f(\boldsymbol{x})$, we learn a predictor $f(\boldsymbol{x}, C)$. This is done by dividing the entire dataset into tasks, creating a collection of related datasets, also known as a meta-dataset. This meta-dataset is defined as $M = D_{i=1}^{N_{tasks}}$, where each task D consists of a context and target set: $D = (C, T) = \{(\boldsymbol{x}^i, \boldsymbol{y}^i)\}_{n=1}^{C+T}$.
We train the meta-learner on this meta-dataset, where the model fits each subset of target points T given the context observations C separately for each task D. At test-time, we adapt the predictor $f(\boldsymbol{x}, C)$ to an unseen task by providing a new context set to make predictions for some unseen target points.

3.4.2 Stochastic Process

We established meta-learning as learning a predictor $f(\boldsymbol{x}, C)$, however, instead of estimating a single prediction given a target input $\boldsymbol{x}_T$, NPs meta-learn a map from datasets to stochastic processes, resulting in a distribution over predictions $p(\boldsymbol{y}_T | \boldsymbol{x}_T, C)$. Within figure 3.3, an overview of the model is provided. In order to model such a stochastic process, we use neural networks that parameterize the predicted distributions. NPs encode the entire context set C to a representation $\boldsymbol{r}_C$. This representation is computed by first encoding every $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$ pair of the context set into a representation $\boldsymbol{r}_i$, before aggregating all individual representations $\boldsymbol{r}_i$ into a representation $\boldsymbol{r}_c$ using an aggregator m (often they take the mean over the individual representations). This is called the deterministic path. Additional to this, there is also a latent path, which models the distribution $p(\boldsymbol{z}|C)$. The result of the latent path is a latent variable $\boldsymbol{z}$, that represents the entire context set. This latent variable is generated by encoding the representation of every pair within the context set into a vector $\boldsymbol{s}_i$, aggregating, and then sampling a latent vector $\boldsymbol{z}$ out of the factorised Gaussian parameterised by $s_C = s(\boldsymbol{x}_C, \boldsymbol{y}_C)$. Another possibility is to compute the distribution s_C directly from the context set, instead of individually for each input-output pair in the context set. To predict a target output, we feed the sampled latent vector $\boldsymbol{z}$ from the latent path, the representation $\boldsymbol{r}_c$ from the deterministic path, and target input $\boldsymbol{x}_*$ to the decoder. The result is a prediction $\boldsymbol{y}_*$. Formally speaking, the generative process is defined as follows:

$$logp(\boldsymbol{Y}|\boldsymbol{X},C) = \int p(\boldsymbol{Y}|\boldsymbol{X},C,\boldsymbol{z})p(\boldsymbol{z},C)d\boldsymbol{z} = \int p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{r}_C,\boldsymbol{z})q(\boldsymbol{z}|\boldsymbol{s}_C)d\boldsymbol{z} \tag{3.7}$$

The parameters are learned for random subsets C and T of a task D by maximizing the ELBO, which differs from the standard latent variable model ELBO, as defined in Equation 3.6. The main difference being the log-likelihood of the conditional distribution $p(\boldsymbol{Y}|\boldsymbol{X})$, together with incorporating the meta-learning formulation:

$$logp(\boldsymbol{Y}|\boldsymbol{X}) \geq E_{q(\boldsymbol{z}|\boldsymbol{s}_T)}[\log p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{r}_C,\boldsymbol{z})] - KL[q(\boldsymbol{z}|\boldsymbol{s}_T)||q(\boldsymbol{z}|\boldsymbol{s}_C)] \tag{3.8}$$

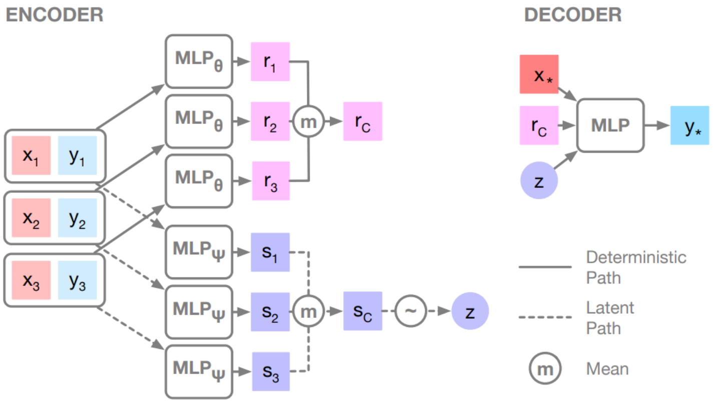


Figure 3.3: Neural processes model, taken from Kim et al. [26]. On the left side, a context set is defined, that exist of input-output pairs $\{\boldsymbol{x}_1,\boldsymbol{y}_1\}$ till $\{\boldsymbol{x}_3,\boldsymbol{y}_3\}$. These pairs are used as input for the deterministic, and latent path. Where for the former we use a neural network MLP_θ to compute the individual representations, and the latter we use a neural network MLP_Ψ for computing the individual representations

3.5 Social Processes

Social processes [38] take the ideas of meta-learning and stochasticity from the NP models and apply this to the task of social cue forecasting. The general motivation behind this is that participants are unlikely to adapt similarly across different groups, due to various factors (e.g. social context, as discussed within this chapter). Figure 3.4 provides an overview of the social processes architecture. This family of models creates a meta-dataset by dividing the dataset into tasks containing sequences from the same group, which leads to the meta-learner learning a predictor for a particular group by only conditioning on observed-future sequence pairs of that same group. This allows the model to adapt its forecasts to a group at test time, meaning it would presumably generalize better when presented with unseen groups. The goal of these models is to model the distribution $p(\boldsymbol{Y}|\boldsymbol{X},C)$, where $\boldsymbol{Y}$ is the set of target outputs, and $\boldsymbol{X}$ is the set of target inputs. The generative process is a bit different than the one formulated by the NP model in Equation 3.7. Instead of feeding an input sequence to the decoder, we encode an observed sequence $\boldsymbol{x}^i$ of every participant p_i into an encoding $\boldsymbol{e}^i$, and feed a concatenated vector $\boldsymbol{e}_*$ consisting of every individual encoding of every participant $\boldsymbol{e}^i$ into the decoder. So in essence, the decoder only accesses $\boldsymbol{x}^i$ through $\boldsymbol{e}^i$. As social behavior is interdependent (as discussed within this chapter), $\boldsymbol{e}^i$ encodes both participant p_i's own behavior, as well as the behavior of partners $p_{j,j=/i}$.

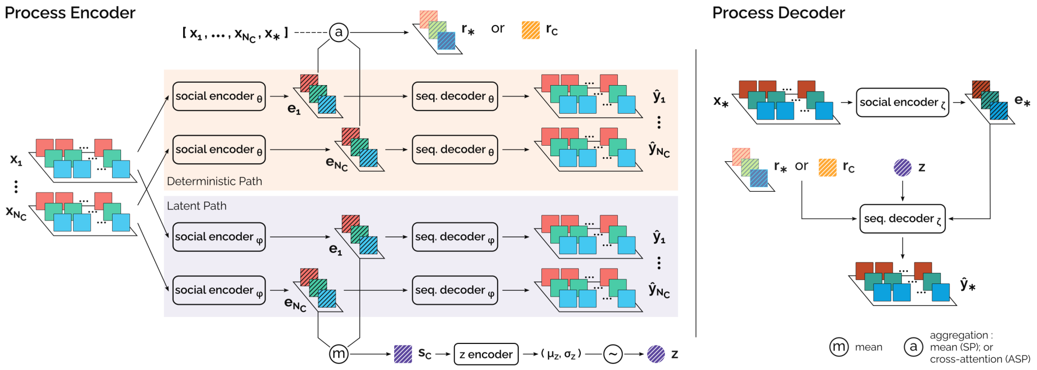


Figure 3.4: Social processes model, taken from Raman et al. [38]

Chapter 4

Methodology

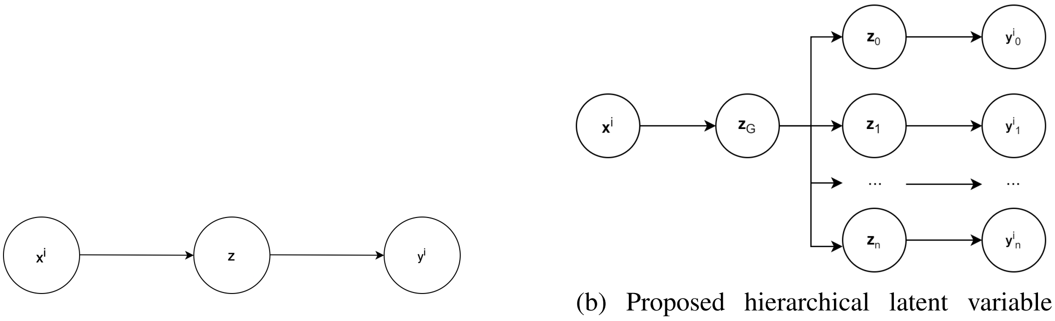
Previously, we found that non-verbal human behavior is influenced by both social context and individual factors. To account for this, we introduced the concept of style in order to model non-verbal behavior within social interactions. One important assumption that was established, is that each individual and social interaction has their own unique style. In order to create a model that is able to learn both individual-level and group-level style from unseen conversations, we propose a meta-learning hierarchical latent variable model.

4.1 Hierarchical Latent Variable Model

The task of predicting future sequences of behavioral cues is inherently a generation task conditioned on the observed sequence of behavioral cues. By introducing a latent variable model, we aim to model the probability distribution $p(\boldsymbol{Y}|\boldsymbol{X})$ with the use of a latent variable. In this specific case, this means that an observed sequence $\boldsymbol{x}_i$ is mapped into a latent variable $\boldsymbol{z}$, which is then used to generate a future sequence. Therefore we can rewrite the conditional as the following:

$$p(\boldsymbol{Y}|\boldsymbol{X}) = \int p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{z})d\boldsymbol{z} = \int p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{z})p(\boldsymbol{z}|\boldsymbol{X})d\boldsymbol{z} \tag{4.1}$$

The goal of this study is to learn style within a conversation. As established in Chapter 2, we can distinguish the style of the entire interaction from individual style, unique to each person within the interaction. This is motivated by the fact that behavioral cues are dependent on social context, and behavioral cues also being idiosyncratic. On a higher level, the group style would be based on the collective behavior of every participant, and should be distinguishable from other interactions. Because this style is based on the collective behavior of every participant, we assume the individual styles of participants also being deducible from this variable. Within previous work, the structure of the latent variable $\boldsymbol{z}$ has been altered in order to represent some semantic meaning within a generation task [42, 51]. Motivated by that idea, we plan on incorporating prior knowledge of the task in order to learn group-level, as well as individual-level style. We came up with the structure



(a) Standard latent variable model

(b) Proposed hierarchical latent variable model

Figure 4.1: The two latent variable models

as shown in figure 4.1b. From an input sequence, we encode a group-level latent vector $\boldsymbol{z}_G$, which is used to generate latent vectors $\boldsymbol{z}_0, ..., \boldsymbol{z}_n$ for all n individuals within a conversation. These individual latent vectors are in turn used to decode a single future sequence $\boldsymbol{y}_i$ for a specific individual. When incorporating this hierarchical structure in a latent variable model, we enforce the model to condition every individual latent variable $\boldsymbol{z}_1$ till $\boldsymbol{z}_n$ on the group latent variable $\boldsymbol{z}_G$, and to only take the latent vector $\boldsymbol{z}_i$ of a single person into account when predicting a future sequence for that individual p_i. Note, however, that the number of people within every conversation within the dataset should be equal to the number of individual latent vectors as specified by the architecture, and is not adaptable to different group sizes. Therefore, the number of individuals within the interactions used as input should be constant. Meaning we require some form of prior knowledge on the number of individuals within each interaction.

4.1.1 ELBO

As opposed to a traditional latent variable model, where the input is encoded into a single latent variable, a hierarchical latent variable model creates a hierarchy of latent variables. Therefore, the standard ELBO formulation changes. As our goal is to model the distribution $p(\boldsymbol{Y}|\boldsymbol{X})$, we derive the ELBO from the log-likelihood:

$$
\begin{aligned}
\log p(\boldsymbol{Y}|\boldsymbol{X}) &= \int_{\boldsymbol{z}} \log p(\boldsymbol{Y}|\boldsymbol{X}, z)dz \\
&= \int_{\boldsymbol{z}} \log p(\boldsymbol{Y}|z_{ind})p(z_{ind}|z_G)p(z_G|\boldsymbol{X})dz \\
&= \log \int_{\boldsymbol{z}} p(\boldsymbol{Y}|z_{ind})p(z_{ind}|z_G)p(z_G|\boldsymbol{X})dz \\
&= \log \int_{\boldsymbol{z}} p(\boldsymbol{Y}|z_{ind})p(z_{ind}|z_G)p(z_G|\boldsymbol{X})dz \frac{q(\boldsymbol{z}|\boldsymbol{X})}{q(\boldsymbol{z}|\boldsymbol{X})}dz \\
&= \log E_{q(\boldsymbol{z}|\boldsymbol{X})}[\frac{p(\boldsymbol{Y}|z_{ind})p(z_{ind}|z_G)p(z_G|\boldsymbol{X})}{q(\boldsymbol{z}|\boldsymbol{X})}] \\
&\geq E_{q(\boldsymbol{z}|\boldsymbol{X})}[\log \frac{p(\boldsymbol{Y}|z_{ind})p(z_{ind}|z_G)p(z_G|\boldsymbol{X})}{q(\boldsymbol{z}|\boldsymbol{X})}] && \text{(by Jensen's inequality)} \\
&= E_{q(\boldsymbol{z}|\boldsymbol{X})}[p(\boldsymbol{Y}|z_{ind}] + E_{q(\boldsymbol{z}|\boldsymbol{X})}[\frac{p(z_{ind}|z_G)p(z_G|\boldsymbol{X})}{q(\boldsymbol{z}|\boldsymbol{X})}] \\
&= E_{q(\boldsymbol{z}|\boldsymbol{X})}[p(\boldsymbol{Y}|z_{ind}] - KL[q(\boldsymbol{z}|\boldsymbol{X})||p(z_{ind}|z_G)p(z_G|\boldsymbol{X})]
\end{aligned}
\tag{4.2}
$$

Where z_{ind} is the set of individual latent vectors $\{z_1, ..., z_n\}$, and $q(\boldsymbol{z}|\boldsymbol{X}) = q(z_G|\boldsymbol{X})\prod_{i=1}^{n} q(z_i|z_G)$. As the distribution of every individual latent variable is independent of the other individual latent variables, we can further marginalize the term $p(z_{ind}|z_G)$ into $\prod_{i=1}^{n} p(z_i|z_G))$.

4.2 Meta-learning

Previously, we defined style to be different within each interaction. However, by only using a hierarchical latent variable model, it would imply that learning style across the entire dataset would not be able to adapt to unseen interactions. Removing the entire premise of our model. Therefore, it makes sense to include some form of meta-learning in the model. We define a meta-dataset by dividing the dataset into tasks containing only sequences from the same group. Just as in the social processes models [38], the meta-learner learns a predictor for a particular group by only conditioning on observed-future sequence pairs of that particular group.

Formally speaking, we consider a meta-dataset of tasks $D_{i=1}^{N_{tasks}}$, where each task only contains sequences from a particular group g_i. Within every task D, we consider a context C, and target set T. As discussed Chapter 3, both the Neural Processes and Social Processes models incorporate meta-learning. Therefore, we redefine the ELBO of those two models using the hierarchical latent variable model approach. For the NP and SP model, instead of Equation 3.8, the ELBO becomes:

$$\log p(\boldsymbol{Y}|\boldsymbol{X}) \geq E_{q(\boldsymbol{z}|\boldsymbol{s}_T)}[\log p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{r}_C, z_{ind})] - KL[q(\boldsymbol{z}|\boldsymbol{s}_T)||q(\boldsymbol{z}|\boldsymbol{s}_C)] \tag{4.3}$$

Where for $q(\boldsymbol{z}|\boldsymbol{s}_T)$ and $q(\boldsymbol{z}|\boldsymbol{s}_C)$, we can marginalize as follows: $q(\boldsymbol{z}|\boldsymbol{s}_T) = q(z_G|\boldsymbol{s}_T)\prod_{i=1}^{n} q(z_i|z_G)$, and simliar as in Equation 4.2, z_{ind} is defined as the set of individual latent vectors $\{z_1, ..., z_n\}$.

Chapter 5

Experiments

In order to validate our hypotheses, we compare our proposed hierarchical latent variable model to a few baseline standard latent variable models. The dataset used for this comes from recorded social interactions, containing a set of modalities.

5.1 Data

There is a need for authentic recorded social interactions to train our model on in order to handle unobserved social interactions at test time. To fulfill this need, we use the haggling dataset [24].

5.1.1 Haggling Dataset

The haggling dataset was proposed by Joo et al. [24], and consists of various recorded triadic interactions. Within the interactions, participants partake in a social game, named the Haggling game. The game simulates a haggling situation, where two participants play the role of sellers, and one participant plays the buyer role. The idea is that each of the two sellers promotes their product, and the buyer decides what product he/she buys between the two. Each recorded interaction lasts one minute, and the seller who sells his/her product is awarded $5 in order to give the participants that play the roles some incentive [24]. The products assigned to the two sellers are similar products, with only slight differences. In one scenario for example, one seller proposes a lightweight cellphone with medium storage, while the other seller proposes a medium-weight cellphone with large storage.
The dataset was recorded within the Panoptic Studio [23]; a geodesic sphere containing 480 VGA cameras, 31 HD cameras, 23 microphones, and 10 Kinect sensors. These were all used for recording social interactions taking place within the

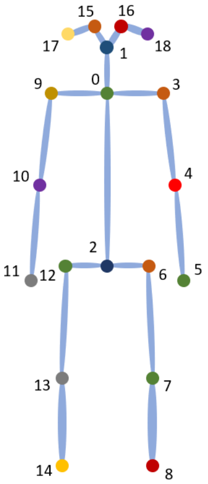


Figure 5.1: The 19 recorded joint locations / MSCOCO19 keypoints

sphere, thus removing the need of attached sensors or markers on any subject's body [24]. Allowing the participants to move without any restrictions. In total, 180 haggling sequences, spanning a total of 3 hours of interactions were recorded.

The provided features within the haggling dataset include x, y, and z coordinates of: 19 body keypoints, 42 hand keypoints (21 per hand), 70 facial keypoints, and binary speaker annotations (1 if a person is speaking, 0 otherwise). The represented joints within the 19 body keypoints are shown within Figure 5.1. As our proposed model is agnostic of the amount of input features, it is possible to use the entire set of features as provided by the haggling dataset. However, due to hardware constraints, we opted for only the 19 body keypoints (as shown in Figure 5.1) and speaking status as the set of behavioral cues $\boldsymbol{b}_t^i$ of a person p_i at time t.

5.1.2 Pre-processing

We divide the dataset into a training and validation set, and a test set respectfully. The training and validation set is used to train and validate the model during the training phase. This gives us an indication of whether the model has converged or not. Finally, the test set is used to evaluate the performance after training the models. Among the total 180 sequences, there are sequences that contain severe reconstruction errors. Therefore, we use the same train-test split as Joo et al. [24] of 79 training sets, and 28 test sets.
Keypoints within the haggling dataset are provided at 30hz. Similar to Raman et al. [38], for our prediction task we consider observed and future windows of a length of 2 seconds (60 frames). The maximum offset between the observed and future windows is 5 seconds (150 frames). As mentioned within the previous section, we only make use of the 19 body keypoints, and speaking status as modalities. Therefore, the input dimension of our data is 58 (19 x, y, and z coordinates, and binary speaking status) per frame.

5.2 Evaluation

We evaluate our proposed approach within two ways: quantitative and qualitative evaluation. As part of the quantitative evaluation, we compare the performance and model fit using the Log Likelihood (LL) and Root Mean Square Error (RMSE). For qualitative evaluation, we look at both the latent space, as well as the plotted results of the models.

5.2.1 Quantitative Evaluation

To evaluate whether our proposed approach performs better than the baseline, we compare the Log Likelihood (LL), as well as the Root Mean Square Error (RMSE) of every predicted keypoint, and speaking status within the test set.
The LL indicates how well the model fits to the ground truth. Therefore, we compute the probability density of a ground truth sample, given the probability distribution generated by the model. In other words, we compute the probability of observing a ground truth

sample, when the data is extracted from the probability distribution that resulted from the model. As this is done over the entire dataset (resulting in the joint likelihood), we use the logarithm of the likelihood. This leads to a more practical output (consider the product of small probabilities versus the sum of logarithms of small probabilities). This evaluation metric tells us how well the model is trained, or how well the produced data distribution captures the ground truth.

In addition to the LL, we also require a loss function to evaluate how close the predicted body keypoints are to the ground truth. For this purpose, we make use of the RMSE of the predicted keypoints. Within a set of n predicted sequences, we compute the RMSE of every predicted future sequence $\hat{y}_i$ and ground truth y_i as follows:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{5.1}$$

For evaluating the speaking status predictions, we calculate the accuracy as follows:

$$Accuracy = \frac{1}{n} \sum_{i=1}^{n} 1(y_i = \hat{y_i}) \tag{5.2}$$

5.2.2 Qualitative Evaluation

As part of the qualitative evaluation, we look at the plotted latent space outputs of our models, and the predicted body keypoints of our model. Here, it is important to judge whether these look in line with the ground truth, or are overall believable body movements. However, we also take a look at the plotted latent vectors.

Latent Space

To examine whether the model has captured idiosyncrasies with respect to group and individual behavior, we need to examine the produced latent vectors by the baseline model, as well as our approached model. For this, we use a technique for dimensionality reduction. Reducing the dimensionality of the resulting latent vectors allows us to visualize the data and discover certain patterns within the latent space. Therefore, we make use of the t-Distributed Stochastic Neighbor Embedding (t-SNE) [45] technique. The concept behind t-SNE is to find a projection of the original data onto a lower dimensional space so that the clustering in the original, high-dimensional data is preserved. This means that similar data points are represented as close to each other as possible within the lower dimensional representation, as opposed to other techniques that focus on plotting dissimilar points far apart (e.g. MDS [31], PCA). Since we are interested in examining clusters formed by the hierarchical latent variables, using t-SNE seems the right dimensionality reduction technique.

5.3 Model Comparisons

To compare our approach, consisting of the meta-learning hierarchical latent variable model, it is necessary to compare the proposed components of our model in isolation. Therefore, we

	Baseline	Proposed	Baseline & Proposed	Backbones
VAE	Non-meta-learning, single latent variable z	-	-	MLP, GRU
Neural Processes	Single latent variable z	Hierarchical latent variable structure	Meta-learning	MLP
Social Processes	Single latent variable z	Hierarchical latent variable structure	Meta-learning	MLP, GRU

Table 5.1: The characteristics of the baseline and proposed models

compare our hierarchical meta-learning approach against a standard meta-learning approach by training and testing the haggling dataset on two different meta-learning architectures. First of all, the Neural Processes model, which incorporates the meta-learning concept and estimates an uncertainty over predictions. Within this model, we implement the hierarchical latent variable structure, and compare this to the baseline approach on the same model using a standard latent variable z. We call this model the hierarchical neural processes model. In addition, we also compare our hierarchical approach using the social processes model in the same manner; including a baseline, as well as the hierarchical approach; the hierarchical social processes model. To allow for a fair comparison, we refrain from using the deterministic path for any of the models. This ensures that the model will not end up relying solely on the deterministic path.

Finally, to also test whether our approach improves against a standard, non-meta-learning model, we also run some additional baseline models. This consists of a standard VAE architecture.

Among all of the models, we separate the comparison into two groups, namely a group consisting of Gated Recurrent Unit Networks (GRU) [10] backbones, and a group consisting of Multi-Layer Perceptron (MLP) backbones. The latter is a standard feedforward neural network, where information flows in a single direction from input to output. A GRU on the other hand, is more suited for sequential data, and uses gating mechanisms to control the information that flows throughout the network. This gating mechanism allows the network to capture long-term dependencies that recurrent neural networks often struggle with (due to the vanishing/exploding gradients problem). In table 5.1, the models and their respective characteristics are summarized.

5.3.1 Hyperparameters and Model Specifics

For running the baseline and proposed models, we mostly use hyperparameters proposed within the work of Raman et al. [38]. For all models, we used a batch size of 128 during training. The training was done on the Delft High Performance Computing Cluster (HPC) on one of the following GPU's: NVIDIA Tesla P100, NVIDIA GeForce GTX 1080/2080 TI, or an NVIDIA Tesla V100. Training each individual model took between one hour (for the simpler VAE models), to 48 hours approximately (for the SP models).

The models are optimized using Adam [27]. This is an optimization algorithm for gradient descent. Instead of the normal case; training a neural network with a fixed learning rate, Adam uses an adaptive learning rate. It essentially computes individual learning rates for the different parameters, allowing it to converge faster. In addition, it makes use of the concept of momentum, which ensures the gradient moves in the most significant direction by taking past gradients into account.

Hyperparameter	**VAE-MLP**	**VAE-GRU**	**NP**	**SP-MLP**	**SP-GRU**
Sequence Encoder/Decoder					
# of layers	2	1	2	2	1
Hidden dim	180	320	180/460	64	320
Partner Pooler					
Number of MLP layers	-	-	-	2	2
MLP hidden dim	-	-	-	64	64
Output dim	-	-	-	64	64
Group z Encoder					
# of layers	2	2	2	2	2
Hidden dim	64	64	64	64	64
Representations					
$\boldsymbol{e}, \boldsymbol{r}, \boldsymbol{s}, z_G$ dim	64	64	64	64	64
Individual z Encoders (only in proposed)					
# of layers	-	-	2	2	2
Hidden dim	-	-	64	64	64
$\boldsymbol{z}_i$ dim			64	64	64
Number of parameters					
Baseline	2.1 M	1.3 M	3.4 M	2.9 M	2.5 M
Proposed	-	-	3.5 M	3.0 M	2.6 M

Table 5.2: Hyperparameters of all architectures

Chapter 6

Results & Analysis

In this chapter, the results of running the haggling dataset are presented and analyzed. Therefore, we divide the results into the types of evaluation. In addition, we define and run a new dataset in order to prove some intuitions about the hierarchical and baseline models.

6.1 Quantitative Results

Table 6.1 shows the obtained results over the initial run over each family on both the baseline, as well as the proposed models. Looking at the results in the table, it does become clear that the performance of the hierarchical models do not show an improvement, when compared to the baseline. As within the SP-GRU model, the proposed approach seems to improve the log-likelihood, while the NP and SP-MLP models seem to achieve a lower log-likelihood. However, th RMSE on the keypoints does not differ by a lot. Therefore, it becomes clear that there is no consistent pattern between the proposed and baseline models. Overall, the models do not seem to perform well with respect to the RMSE. From the results we can also conclude that the VAE-GRU outperforms any compared model, and is therefore the best-performing model within our experiments. A complete overview of the errors on all keypoints is provided in Appendix A.

Backbone	Model	Mean LL	Mean keypoint RMSE	Speaking Accuracy
MLP	VAE-MLP	-249.37 (39.80)	79.61 (14.77)	0.63 (0.15)
	NP Baseline	-22.67 (90.75)	15.74 (5.35)	0.65 (0.23)
	Hierarchical NP	-24.49 (102.78)	15.98 (5.46)	0.64 (0.24)
	SP-MLP Baseline	-105.58 (73.24)	25.89 (8.30)	0.64 (0.27)
	Hierarchical SP-MLP	-117.22 (91.11)	26.02 (9.71)	0.66 (0.24)
GRU	VAE-GRU	221.93 (29.63)	1.19 (0.46)	0.97 (0.02)
	SP-GRU Baseline	-23.09 (166.90)	11.25 (4.28)	0.72 (0.19)
	Hierarchical SP-GRU	-8.93 (156.17)	11.41 (4.44)	0.71 (0.22)

Table 6.1: The mean log likelihood (ll), mean RMSE (in cm) over all 19 body keypoints, and speaking accuracy along with their respective standard deviations on the test set. For the LL and speaking accuracy; higher is better, while for the RMSE; lower is better

6.2 Qualitative Results

Qualitative evaluation of the data seems to match the findings by Raman et al. [38], namely that the meta-learning models seem to produce a neutral pose that seems to minimize the overall loss with respect to every conversation. Within Figure 6.1, the ground truth future sequence, together with the prediction produced by the hierarchical NP, and the VAE-GRU model are shown. The NP proposed results represent the results of every model within our comparison, where there seems to be a neutral pose throughout the entire sequence. The VAE-GRU model, however, seems to have learned something and produces a future sequence that resembles and is close to the ground truth. Clearly proving the better RMSE and LL results.

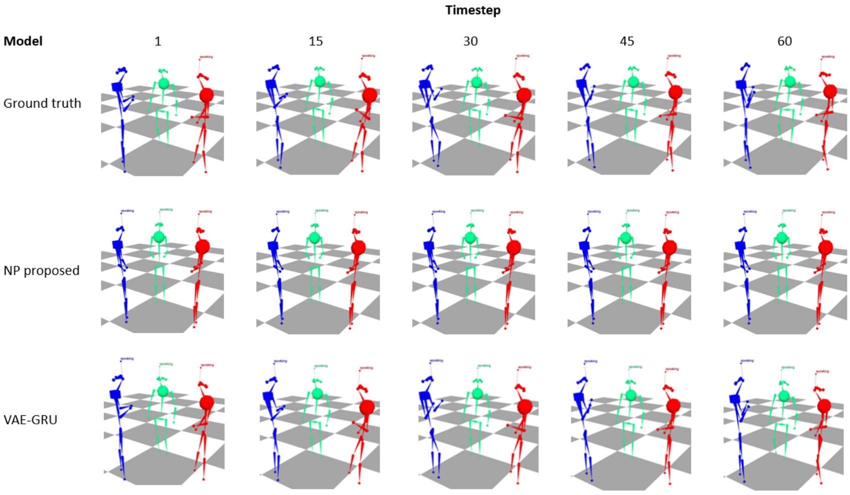


Figure 6.1: Qualitative results on haggling dataset. A random future sequence of 60 timesteps was predicted by both the VAE-GRU model, as well as the NP Style model, and compared to the ground truth. To show different frames from the predicted sequences, we took 5 frames throughout the entire future sequence.

6.2.1 Latent Space

Taking a closer look at the produced latent space by the proposed models, there does not seem to be a clear pattern within the data. Figure 6.2 shows a t-sne plot of the latent variables produced by the hierarchical NP model. However, all meta-learning models show a similar pattern. As t-SNE is used, the distance between points do not encompass any meaning, and after taking a closer look at the produced latent vectors it seems the models did not learn a representation of the data. Instead, they produce a consistent latent representation that is independent of the given target or context set. This problem could be either a training artefact, or attributed to the variance in the data being small. Meaning that all 2-second sequences throughout the dataset might be very similar. As we observed this phenomenon in all meta-learning models, we opted to create a small toy dataset that is generated using a conditional latent structure as well. Running this dataset using our proposed hierarchical latent variable model, could prove whether a hierarchical latent variable model in combination with the meta-learning concept is able to infer a hierarchical latent structure within the data.

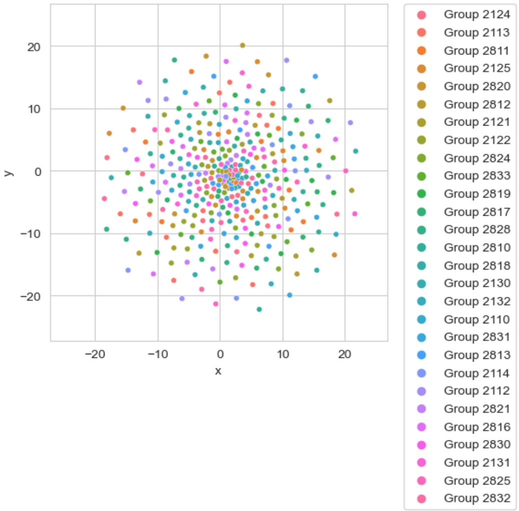


Figure 6.2: T-SNE plot of the group latent vectors using the hierarchical NP model

6.3 Toy Dataset

To prove our intuition behind the proposed hierarchical models being able to learn a hierarchical latent structure, we created a toy dataset. This toy dataset is supposed to test both hierarchical and baseline models on their ability to learn latent representations from data.

6.3.1 Data generation

This dataset consists of a series of sine functions. Within figure 6.3, the sampling procedure for these sine functions for a single group (Gaussian) is shown. For generating these sine functions, we define two Gaussian within 1D space, for which we sample 100 values for each Gaussian as value for the frequency over 700 timesteps. As a condition on which Gaussian we sample from, we define two 1D Gaussians to sample the amplitude for two sine functions. This creates a hierarchical structure for sampling a group of two sine wave sequences. In total, we generated 200 sine wave sequences (100 from each frequency gaussian), consisting of 700 timesteps. For which the model predicts a future sequence of 100 timesteps, given an observed sequence of 100 timesteps. Shortly summarized, the hierarchical latent structure for generating these sine functions consists of a Gaussian defining the frequency, and a Gaussian defining the amplitude for each individual sequence. Finally, in order to check whether the model performs well, we also generate a dataset using only the first layer of the hierarchy. Resulting in groups of two sinewaves with different frequencies, but equal amplitude values. This represents a non-hierarchical data generation process, and the baseline models should be capable of learning a useful latent representation for this.

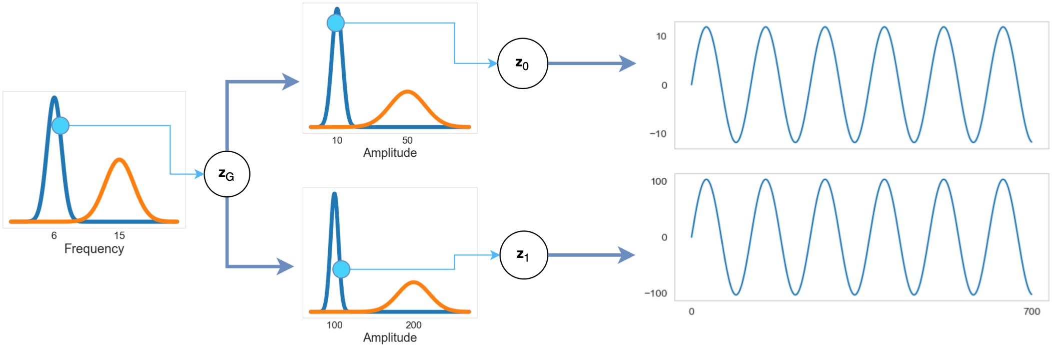


Figure 6.3: Data generation process of the toy dataset. The group latent variable z_G represents the frequency, as sampled from one of the two groups. Groups are distinguished by the color of their distribution. In this case, a group latent vector is sampled from the first Gaussian. For which we sample two individual latent variables z_0 and z_1 representing the amplitude for each respective sine wave, from Gaussians of the corresponding blue group. Resulting in two sinewaves with amplitudes +-100 and +-10, and a frequency of 6. The second-level latent variables can only be sampled from the corresponding groups (blue or orange), enforcing the data generation to contain a hierarchical latent structure

6.3.2 Posterior Collapse

Running the non-hierarchical and hierarchical dataset on all models produces the same result; a consistent latent representation. This suggests that the problem is indeed a training artefact, and not caused by characteristics of the dataset. More formally when looking at the ELBO defined for the meta-learning models in Equation 4.3, the approximate probability $q(\boldsymbol{z}|\boldsymbol{s}_C)$ is consistent across different context sets and the KL divergence between $q(\boldsymbol{z}|\boldsymbol{s}_C)$ and $q(\boldsymbol{z}|\boldsymbol{s}_T)$ converges to 0. This is presumably the reason why the latent vectors do not encompass any meaning. As this KL term is already optimized. Therefore, the model only trains the left side of the term, which is the log likelihood between the modeled distribution, and the ground truth. This problem of latent vectors becoming independent of the input is related to a similar issue often encountered in VAE's, which is known as posterior collapse. Within a VAE, the ELBO is formulated as in Equation 3.6. The difference between our meta-learning models is that the KL divergence is computed between the posterior $q(\boldsymbol{z}|\boldsymbol{X})$ and the prior $p(\boldsymbol{z})$, where the prior is often parameterized by a standard normal distribution, with 0 mean, and 1 variance. Posterior collapse occurs when the approximate posterior collapses into, or equals to, the prior. The result of this is that the first term of the ELBO is solely optimized since the KL term is equal or very close to zero. Therefore, the premise of the model of generating a meaningful latent vector is neglected, and instead, it learns to generate generic outputs that are crude representations of all seen input data.

6.3.3 Meta-learning Formulation

Within the case of both NP and SP meta-learning models, the ELBO is formulated differently as specified in Equation 4.3, and instead of the posterior $q(\boldsymbol{z}|\boldsymbol{X})$ collapsing to the prior $p(\boldsymbol{z})$, the approximation $q(\boldsymbol{z}|\boldsymbol{s}_T)$ collapses to the approximation $q(\boldsymbol{z}|\boldsymbol{s}_C)$. After examining the KL term throughout training, it seems indeed the case is that the KL divergence between the two terms converges to 0, or close to zero. The overall outcome of the problem also results in a neglected latent vector. Which could potentially explain why all meta-learning models predict a static pose for the entire future sequence. Within the VAE literature, this problem is examined in multiple works. One of the approaches to mitigate posterior collapse, is by KL annealing or initializing a fixed weight to the KL term [39, 6]. The idea of KL annealing is to start with a low value, or zero as weight of the KL term within the ELBO, and slowly increment the weight during training. Even though this is used in VAE's, where the formulation of the ELBO is different, we believe it could also hold within the meta-learning model as we do not restrict the approximate distributions $q(\boldsymbol{z}|\boldsymbol{s}_T)$ and $q(\boldsymbol{z}|\boldsymbol{s}_C)$ to be close to each other. One experiment was run using a constant weight term, however, this did not provide any change with respect to the latent representations. Therefore, a more sophisticated approach is needed.

Chapter 7

Discussion

The problem of generating a consistent latent representation leaves us unable to answer any of the research questions and hypotheses, making every compared model similar. Therefore, we consider this as a limitation of the work. Below, more limitations are described in the case of a working model. These limitations could also hinder the proposed approach from optimally learning style, both reasoned from a technical perspective, as well as more socially grounded limitations are discussed. Lastly, we describe a few ethical considerations surrounding our work.

7.1 Preventing Posterior Collapse

Due to limited time, solely a naive version of KL weighting was tested. Where the weight was initialized with a constant number. This intervention proved to be unsuccessful, therefore calling for a more sophisticated approach. Within the work of Sønderby et al. [42], the KL annealing technique proves to work on a hierarchical latent variable model. However, as established within the previous chapter, the KL term differs with respect to the ELBO formulation when compared to the meta-learning definition. Within the neural processes family, the phenomenon seems relatively unexplored. There has been one variation of the original neural processes; sequential neural processes [41], where they describe the problem of transition-collapse. However, this problem is unique to the sequential neural processes model, and therefore not of any value to our case. Shortly summarized; more work is needed for preventing the posterior collapse problem in neural processes-based models. Techniques for coping with this issue have been described in earlier work on VAE's, and could be applicable to our model.

7.2 Dataset

First of all, we can argue about the ecological validity of the experiment by using the haggling dataset. Within the interactions in the haggling dataset, individuals are assigned either a buyer, or seller role. To which degree this influences the behavior of participants is unclear. However, considering our notion of style, which is also dependent on social context, we can argue that a part of the social context is already provided by playing the haggling game. Namely, the roles of the participants, part of the verbal content (which product to sell, and typical vocabulary within haggling contexts), and the situation of the conversation. This might be a limitation as to how generalizable our system would be for unseen interactions within different contexts, as well as form a problem in differentiating between specific group styles within the dataset. Leaving us with a rather homogeneous dataset. In short, these assigned roles and context might make it more difficult to differentiate between group styles, and using a different dataset with a variety of contexts might lead to different results.

7.3 Capturing Idiosyncrasies

Another limitation within the model, which is more theoretically grounded, is our proposed modeling of style. Using our meta-learning approach, an individual's idiosyncrasies may only be captured through conversations within the same group. Therefore, the proposed model does not take into account that idiosyncrasies might be transferable over conversations containing the same individual. As discussed within the background chapter, gestures are dependent on social context and verbal content, while also being dependent on personality, and age. This hints that style is a mix of static factors (age, personality), as well as more dynamic factors (social context, verbal content) that differ between conversations. Our proposed model currently only reasons on the basis of conversation-specific factors, which might be a limitation since it provides a naive definition of style.

In addition, the premise of our approach is to divide style into group, and individual-level idiosyncrasies. However, there would be a range of possibilities to divide style into a semantic hierarchy. One of the possibilities is to model this hierarchy differently, or even have the formation of the hierarchy structure take place as part of the training.

7.4 Ethical Considerations

Even though the dataset used within this study contained personal data. The haggling dataset was collected and released with the permission of all individual participants. However, putting such a social cue forecasting system in practice might lead to some ethical concerns. First of all, for such a system to be accurate with respect to future behavioral cues, multiple modalities have to be added, which may or may not be able to be anonymized. Of which facial expressions and body movement sometimes might be able to identify individuals. Therefore, it is of the essence that users should be aware of the presence of a forecasting system. Also, as the system makes use of privacy-sensitive data (video footage of interactions), potential vulnerabilities could be exploited, where users with malicious intent can access potentially vulnerable information.

In addition, these forecasting techniques can be used for social good. Where assistive, or collaborative robots are able to help people within various fields. One important point here is that this might be implemented in social robots that function in rather vulnerable situations. Examples of this are social robots for elders with dementia, or any social robot that interacts with children. This means the system could have a large effect on the behavior of users. An example of this is a social robot using persuasion techniques via verbal, and non-verbal communication in order to convince a child of buying a product. Therefore, the implementation of such systems requires a responsible approach. In contrast, the system could also be implemented in situations that might be undesired, such as surveillance systems.

Chapter 8

Conclusion

Within this work, we delved into the theory of hierarchical latent variable models and meta-learning in order to learn group and individual-level style, unique to each conversation. Incorporating this notion of style in a non-verbal human behavior forecasting model might help us forecast behavior that is more in line with the specific conversation and individuals. Including this style variable has not been done in previous work on non-verbal human behavior forecasting. After formulating a model and running experiments on a real-world dataset, the model proved to be unsuccessful. However, as the latent representation throughout every model within the experiment (except the VAE baselines) turned out to be independent of the input, we formulated a toy dataset to test our intuition. This toy dataset proved the problem to be a training artefact. Therefore, more research is needed on how to avoid posterior collapse in neural processes models, before answering our research questions on the effectiveness and ability to learn individual and group level style of our proposed model.

8.1 Future Work

As our proposed model has not been able to prove the effectiveness of a hierarchical latent variable model for learning style, we can address certain shortcomings or experiments that lead to potential future research questions or insights.

8.1.1 Incorporating Latent Structures within Meta-learning Models

Within this study, we assumed the intuition of a meta-learning model, or more specifically, the family of neural processes models, to be able to learn a hierarchy with respect to the generated data. However, this intuition proved to be false. Therefore, an extensive study could be done on the theory of hierarchical latent variable models, and how this could be incorporated into a meta-learning model before using such a model on a specific task. Obstacles in training these models could be explored, as well as variations on the architectures.

8.1.2 Latent Structure Design

One limitation of this study was the use of a single proposed latent variable structure. However, there has been work done that shows deep hierarchical latent variable models to be successful on certain tasks. For this reason, experiments with different latent structures could be useful to examine characteristics for a successful forecasting model. In addition, the proposed model only supports a group size until the nth individual latent vector, with n being the number of individuals within a conversation. However, support for arbitrary group sizes might benefit the model by making it trainable and run across datasets.

8.1.3 Variety of Datasets

As mentioned in the discussion, the proposed model was only run for the haggling dataset, which is one of the limitations of our work. Running a proposed model on multiple datasets could provide us with a more complete view of this group-level style. As the haggling dataset is only composed of groups playing the same haggling game, it could be considered as a rather homogeneous dataset. Comparing results on different datasets, and training a model on a variety of datasets might make a proposed model more robust. In addition, testing the proposed model on multiple datasets improves the generalizability of a possible conclusion.

Bibliography

[1] Chaitanya Ahuja, Shugao Ma, Louis-Philippe Morency, and Yaser Sheikh. To react or not to react: End-to-end visual pose forecasting for personalized avatar during dyadic conversations, 2019. URL `https://arxiv.org/abs/1910.02181`.

[2] Chaitanya Ahuja, Dong Won Lee, Yukiko I. Nakano, and Louis-Philippe Morency. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. 2020. doi: 10.48550/ARXIV.2007.12553. URL `https://arxiv.org/abs/2007.12553`.

[3] Martha Alibali, Julia Evans, Autumn Hostetter, Kristin Ryan, and Elina Mainela-Arnold. Gesture-speech integration in narrative: Are children less redundant than adults? *Gesture*, 9:290–311, 12 2009. doi: 10.1075/gest.9.3.02ali.

[4] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, apr 2017. doi: 10.1080/01621459.2017.1285773. URL `https://doi.org/10.1080%2F01621459.2017.1285773`.

[5] Dan Bohus and Eric Horvitz. Managing human-robot engagement with forecasts and... ¡i¿um¡/i¿... hesitations. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, page 2–9, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450328852. doi: 10.1145/2663204.2663241. URL `https://doi.org/10.1145/2663204.2663241`.

[6] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Conference on Computational Natural Language Learning*, 2016.

[7] Judee K. Burgoon, Nadia Magnenat-Thalmann, Maja Pantic, and Alessandro Vinciarelli. *Social Signal Processing*. Cambridge University Press, USA, 1st edition, 2017. ISBN 1107161266.

[8] Judith Bütepage, Hedvig Kjellström, and Danica Kragic. Anticipating many futures: Online human motion prediction and synthesis for human-robot collaboration, 2017. URL `https://arxiv.org/abs/1702.08212`.

[9] Jonathan A. Caballero, Nikos Vergis, Xiaoming Jiang, and Marc D. Pell. The sound of im/politeness. *Speech Communication*, 102:39–53, 2018. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2018.06.004. URL `https://www.sciencedirect.com/science/article/pii/S016763931830013X`.

[10] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078, 2014. URL `http://arxiv.org/abs/1406.1078`.

[11] Starkey Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23:283–292, 1972.

[12] Ylva Ferstl, Michael Neff, and Rachel McDonnell. Multi-objective adversarial gesture generation (mig 2019). 10 2019. ISBN 978-1-4503-6994-7. doi: 10.1145/3359566.3360053.

[13] Daniel Flam-Shepherd, Yuxiang Gao, and Zhaoyu Guo. Stick-breaking neural latent variable models. 2018.

[14] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J. Rezende, S. M. Ali Eslami, and Yee Whye Teh. Neural processes, 2018. URL `https://arxiv.org/abs/1807.01622`.

[15] Simon Garrod and Martin Pickering. The use of content and timing to predict turn transitions. *Frontiers in Psychology*, 6, 06 2015. doi: 10.3389/fpsyg.2015.00751.

[16] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3492–3501, 2019.

[17] Ariel Gjaci, Carmine Recchiuto, and Antonio Sgorbissa. Towards culture-aware co-speech gestures for social robots. *International Journal of Social Robotics*, 14, 06 2022. doi: 10.1007/s12369-022-00893-y.

[18] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL `https://arxiv.org/abs/1406.2661`.

[19] Judith Holler and Geoffrey Beattie. Gesture use in social interaction: how speakers' gestures can reflect listeners' thinking. 01 2007.

[20] Autumn Hostetter. Action attenuates the effect of visibility on gesture rates. *Cognitive Science*, 38, 05 2014. doi: 10.1111/cogs.12113.

[21] Autumn Hostetter and Martha Alibali. Raise your hand if you're spatial: Relations between verbal and spatial skills and gesture production. *Gesture*, 7:73–95, 04 2007. doi: 10.1075/gest.7.1.05hos.

[22] Autumn Hostetter and Andrea Potthoff. Effects of personality and social situation on representational gesture production. *Gesture*, 12, 09 2012. doi: 10.1075/gest.12.1.04h os.

[23] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart C. Nabbe, Iain A. Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *CoRR*, abs/1612.03153, 2016. URL `http://arxiv.org/abs/1612.03153`.

[24] Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction, 2019. URL `https://arxiv.org/abs/1906.04158`.

[25] Anne Keitel and Moritz Daum. The use of intonation for turn anticipation in observed conversations without visual signals as source of information. *Frontiers in Psychology*, 6, 02 2015. doi: 10.3389/fpsyg.2015.00108.

[26] Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, S. M. Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. *CoRR*, abs/1901.05761, 2019. URL `http://arxiv.org/abs/1901.05761`.

[27] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL `https://arxiv.org/abs/1412.6980`.

[28] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013. URL `https://arxiv.org/abs/1312.6114`.

[29] Sotaro Kita. Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes*, 24(2):145–167, 2009. doi: 10.1080/01690960802586188. URL `https://doi.org/10.1080/01690960802586188`.

[30] Sotaro Kita, Ingeborg van Gijn, and Harry van der Hulst. Movement phases in signs and co-speech gestures, and their transcription by human coders. In Ipke Wachsmuth and Martin Fröhlich, editors, *Gesture and Sign Language in Human-Computer Interaction*, pages 23–35, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg. ISBN 978-3-540-69782-4.

[31] Joseph B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.

[32] Kedi Lyu, Haipeng Chen, Zhenguang Liu, Beiqi Zhang, and Ruili Wang. 3d human motion prediction: A survey, 2022. URL `https://arxiv.org/abs/2203.01593`.

[33] Katja Maquate and Pia Knoeferle. *The Interactive Mind: Effects of Social Context on Language Processing*. 06 2018.

[34] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3:861, 09 2018. doi: 10.21105/joss.00861.

[35] David Mcneill. Hand and mind: What gestures reveal about thought. *Bibliovault OAI Repository, the University of Chicago Press*, 27, 06 1994. doi: 10.2307/1576015.

[36] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. URL `http://arxiv.org/abs/1411.1784`.

[37] Miles L. Patterson. Nonverbal interpersonal communication. *Oxford Research Encyclopedia of Communication*, 2018.

[38] Chirag Raman, Hayley Hung, and Marco Loog. Social processes: Self-supervised forecasting of nonverbal cues in social conversations. *CoRR*, abs/2107.13576, 2021. URL `https://arxiv.org/abs/2107.13576`.

[39] Ali Razavi, Aäron van den Oord, Ben Poole, and Oriol Vinyals. Preventing posterior collapse with delta-vaes. *CoRR*, abs/1901.03416, 2019. URL `http://arxiv.org/abs/1901.03416`.

[40] Dinghan Shen, Asli Celikyilmaz, Yizhe Zhang, Liqun Chen, Xin Wang, Jianfeng Gao, and Lawrence Carin. Towards generating long and coherent text with multi-level latent variable models. *CoRR*, abs/1902.00154, 2019. URL `http://arxiv.org/abs/1902.00154`.

[41] Gautam Singh, Jaesik Yoon, Youngsung Son, and Sungjin Ahn. Sequential neural processes. *CoRR*, abs/1906.10264, 2019. URL `http://arxiv.org/abs/1906.10264`.

[42] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders, 2016. URL `https://arxiv.org/abs/1602.02282`.

[43] Nguyen Tan Viet Tuyen and Oya Celiktutan. Context-aware human behaviour forecasting in dyadic interactions. In Cristina Palmero, Julio C. S. Jacques Junior, Albert Clapés, Isabelle Guyon, Wei-Wei Tu, Thomas B. Moeslund, and Sergio Escalera, editors, *Understanding Social Behavior in Dyadic and Small Group Interactions*, volume 173 of *Proceedings of Machine Learning Research*, pages 88–106. PMLR, 16 Oct 2022. URL `https://proceedings.mlr.press/v173/tuyen22a.html`.

[44] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder, 2020. URL `https://arxiv.org/abs/2007.03898`.

[45] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL `http://jmlr.org/papers/v9/vandermaaten08a.html`.

[46] Alessandro Vinciarelli and Alex ’Sandy’ Pentland. New social signals in a new interaction world: The next frontier for social signal processing. *IEEE Systems, Man, and Cybernetics Magazine*, 1:10–17, 2015.

[47] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image Vis. Comput.*, 27:1743–1759, 2009.

[48] Petra Wagner, Zofia Malisz, and Stefan Kopp. Gesture and speech in interaction: An overview. *Speech Communication*, 57:209–232, 02 2014. doi: 10.1016/j.specom.2013.09.008.

[49] Kevin Wells. *Noticing Students’ Conversations and Gestures During Group Problem-Solving in Mathematics*. 05 2017. ISBN 978-3-319-46752-8. doi: 10.1007/978-3-319-46753-5_11.

[50] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics*, 39(6):1–16, dec 2020. doi: 10.1145/3414685.3417838. URL `https://doi.org/10.1145%2F3414685.3417838`.

[51] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Learning hierarchical features from generative models, 2017. URL `https://arxiv.org/abs/1702.08396`.

Appendix A

Complete Experiment Results

Backbone	**Model**	**Mean LL**	**Neck**	**Nose**	**Body center**	**Left shoulder**	**Left elbow**
			(0)	**(1)**	**(2)**	**(3)**	**(4)**
MLP	VAE-MLP	-249.37 (39.80)	85.79 (15.11)	76.94 (16.58)	83.32 (14.80)	87.59 (15.43)	88.43 (16.17)
	NP Baseline	-22.67 (90.75)	15.47 (5.35)	15.12 (5.46)	14.38 (5.03)	16.34 (5.42)	18.63 (5.82)
	NP Proposed	-24.49 (102.78)	15.87 (5.38)	15.55 (5.67)	14.67 (5.09)	16.54 (5.47)	18.70 (6.15)
	SP-MLP Baseline	-105.58 (73.24)	26.13 (8.32)	25.49 (9.12)	25.31 (8.07)	26.44 (7.99)	28.37 (9.18)
	SP-MLP Proposed	-117.22 (91.11)	26.13 (9.48)	25.74 (10.98)	25.24 (9.65)	26.55 (9.44)	28.95 (10.66)
RNN	VAE-RNN	221.93 (29.63)	0.95 (0.34)	0.90 (0.39)	0.90 (0.35)	1.35 (0.42)	1.37 (0.55)
	SP-RNN Baseline	-23.09 (166.90)	10.21 (4.36)	10.55 (4.63)	9.62 (4.17)	10.67 (4.37)	13.41 (4.67)
	SP-RNN Proposed	-8.93 (156.17)	9.99 (4.41)	10.56 (4.55)	9.53 (4.35)	11.87 (4.49)	14.65 (4.72)

Table A.1: LL and RMSE results (in cm) over the first 5 keypoints

Model	**Left wrist (5)**	**Left hip (6)**	**Left knee (7)**	**Left ankle (8)**	**Right shoulder (9)**
VAE-MLP	82.07 (17.58)	84.01 (14.99)	84.88 (15.11)	87.04 (14.88)	86.26 (14.79)
NP Baseline	23.32 (7.28)	14.65 (4.91)	15.18 (4.87)	15.75 (5.36)	16.40 (5.82)
NP Proposed	23.29 (7.33)	14.82 (5.01)	15.35 (5.12)	15.96 (5.69)	16.80 (5.71)
SP-MLP Baseline	32.02 (10.09)	25.22 (7.84)	25.52 (7.87)	25.87 (7.82)	27.99 (8.86)
SP-MLP Proposed	32.6 (11.67)	25.28 (9.44)	25.73 (9.59)	26.27 (9.51)	28 (9.91)
VAE-RNN	1.59 (1.02)	1.25 (0.43)	1.37 (0.48)	1.03 (0.39)	1.37 (0.41)
SP-RNN Baseline	19.04 (6.11)	9.74 (4.17)	10.61 (3.76)	11.40 (3.75)	11.17 (4.39)
SP-RNN Proposed	19.04 (6.25)	10.15 (4.28)	11.18 (4.26)	12.69 (4.62)	10.86 (4.60)

Table A.2: RMSE results (in cm) over 5 keypoints

Backbone	**Model**	**Right elbow (10)**	**Right wrist (11)**	**Right hip (12)**	**Right knee (13)**	**Right ankle (14)**
MLP	VAE-MLP	86.48 (15.13)	80.55 (16.85)	83.59 (14.63)	84.11 (14.84)	86.88 (14.62)
	NP Baseline	18.25 (6.42)	23.17 (6.93)	14.81 (5.30)	15.25 (5.32)	15.96 (5.64)
	NP Proposed	18.42 (6.34)	22.78 (7.11)	15.12 (5.27)	15.50 (5.52)	16.34 (5.61)
	SP-MLP Baseline	30.34 (9.45)	33.05 (10.1)	26.19 (8.43)	26.99 (8.64)	28.58 (8.77)
	SP-MLP Proposed	30.45 (10.85)	32.95 (11.67)	26.12 (9.9)	27.02 (9.91)	28.33 (10.04)
RNN	VAE-RNN	1.31 (0.49)	1.52 (0.68)	1.29 (0.41)	1.34 (0.45)	1.03 (0.43)
	SP-RNN Baseline	13.58 (4.61)	19.40 (5.74)	9.98 (4.16)	11.04 (4.14)	11.96 (4.14)
	SP-RNN Proposed	13.01 (4.85)	19.75 (5.60)	9.64 (4.36)	10.73 (4.29)	11.36 (4.62)

Table A.3: RMSE results (in cm) over 5 keypoints

Backbone	**Model**	**Right eye (15)**	**Left eye (16)**	**Right ear (17)**	**Left ear (18)**	**Speaking Accuracy**
MLP	VAE-MLP	78.40 (16.44)	83.82 (15.62)	78.20 (16.32)	83.20 (15.44)	0.63 (0.15)
	NP Baseline	15.16 (5.43)	15.54 (5.48)	15.20 (5.47)	15.62 (5.52)	0.65 (0.23)
	NP Proposed	15.60 (5.67)	15.92 (5.62)	15.69 (5.62)	16.11 (5.53)	0.64 (0.24)
	SP-MLP Baseline	25.44 (8.97)	25.89 (8.36)	25.77 (9.07)	26.53 (8.78)	0.64 (0.27)
	SP-MLP Proposed	25.72 (10.72)	26.06 (9.73)	25.98 (10.82)	26.65 (10.03)	0.66 (0.24)
RNN	VAE-RNN	0.91 (0.39)	1.20 (0.64)	0.88 (0.39)	1.22 (0.48)	0.97 (0.02)
	SP-RNN Baseline	10.39 (4.57)	10.34 (4.59)	10.58 (4.55)	10.58 (4.43)	0.72 (0.19)
	SP-RNN Proposed	10.64 (4.52)	11.06 (4.62)	10.30 (4.59)	10.38 (4.59)	0.71 (0.22)

Table A.4: RMSE results (in cm) over 4 keypoints, together with the speaking status accuracy