Multimodal Expressive Gesturing With Style

Mireille Fares

▶ **To cite this version:**

Mireille Fares. Multimodal Expressive Gesturing With Style. Human-Computer Interaction [cs.HC]. Sorbonne Université, 2023. English. NNT : 2023SORUS017 . tel-04100511

HAL Id: tel-04100511
https://theses.hal.science/tel-04100511

Submitted on 17 May 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Multimodal Expressive Gesturing With Style

École Doctorale Informatique, Télécommunications et Électronique

THÈSE DE DOCTORAT
DE SORBONNE UNIVERSITÉ

présentée et soutenue publiquement par

Mireille Fares

le 15 Février 2023

Directrice de thèse: **Catherine Pelachaud**
Co-encadrant: **Nicolas Obin**

devant le jury composé de :

M. Thierry ARTIÈRES, Professeur, École Centrale Marseille,	Rapporteur
Mme Chloé CLAVEL, Professeure, Institut Polytechnique de Paris,	Examinatrice
M. Michael NEFF, Professeur, University of California,	Rapporteur
M. Nicolas OBIN, Maître de Conférences, Sorbonne Université,	Examinateur
Mme Catherine PELACHAUD, Professeure, Sorbonne Université,	Examinatrice
M. Brian RAVENET, Maître de Conférences, Université Paris-Saclay,	Examinateur
Mme Laure SOULIER, Maîtresse de Conférences, Sorbonne Université,	Examinatrice

... And Men created an intelligence at their own image

To my parents

and

To the memory of Yolla

Abstract

Human communication is essentially and inherently *multimodal*, it encompasses a gestalt of multimodal signals that involve much more than the speech production system. Primarily, the *verbal* and *non-verbal* communication modes are inextricably and jointly intertwined to deliver the semantic and pragmatic content of the message and tailor the communication process. These exchanged multimodal signals involve both vocal and visual channels which, when combined, render the communication more *expressive*. The vocal mode is characterized by *acoustic features* - namely *prosody* - while the visual mode involves *facial expressions*, *hand* gestures and *body* gestures. The evolving virtual and online communication created the need for generating expressive communication for human-like embodied agents, including *Embodied Conversational Agents* (ECA) and social robots. One crucial communicative signal for ECAs, that can convey a wide range of messages is *visual* (facial and body) motion that accompanies speech and its semantic content. The generation of *appropriate* and *coherent* gestures allows ECAs to articulate the speech intent and content in a human-like expressive fashion.

The central theme of the present manuscript is to leverage and control the ECAs' *behavioral expressivity* by modelling the complex *multimodal* behavior that humans employ during communication. Concretely, the driving forces of this thesis are twofold: (1) to exploit *speech prosody*, *visual prosody* and *language* with the aim of synthesizing expressive and human-like behaviors for ECAs; (2) to control the style of the synthesized gestures such that we can generate them with the style of any speaker. With these motivations in mind, we first propose a semantically-aware and speech-driven facial and head gesture synthesis model trained on a corpus that we collected from TEDx talks. Then we propose *ZS-MSTM 1.0*, an approach that allows the synthesis of stylized upper-body gestures, driven by the content of a source speaker's speech (audio and text) and corresponding to the style of *any* target speakers, *seen* or *unseen* by our model. *ZS-MSTM 1.0* is trained on *PATS* corpus which includes multimodal data of speakers having different behavioral style, however our model is not limited to *PATS* speakers, and can generate gestures in the style of any newly coming speaker without further training or fine-tuning, rendering our approach *zero-shot*. More specifically, *behavioral style* is modelled based on multimodal speakers' data - language, body gestures, and speech -, and independent from the speaker's identity ("ID"). We additionally extend this model and propose *ZS-MSTM 2.0*, which generates stylized facial gestures in addition to the upper-body gestures. We train *ZS-MSTM 2.0* on *PATS* corpus, which we extended to include dialog acts and 2D facial landmarks aligned with the other multimodal features of this dataset (2D body poses, language, and speech).

Keywords: Human Behavior Modelling, Gesture Synthesis, Multimodality, Embodied Conversational Agents, Zero-Shot Style Transfer, Visual Prosody

Acknowledgment

I want to thank the *Sorbonne Center for Artificial Intelligence (SCAI)* and the Labex SMART for the three-years full scholarship that made this research possible. I would like to especially thank Professor Xavier Fresquet, and Professor Gérard Biau.

Most of all, I would like to express my deep and sincere gratitude to my research supervisors Professor Catherine Pelachaud and Professor Nicolas Obin for giving me the opportunity to conduct this project and providing invaluable guidance throughout this research. It was a great privilege and honor to work under their supervision. Catherine, you have been an amazing mentor through all the highs and lows. Your patience and encouragement are unequalled. Your effort to celebrate the successes are also unequalled. I hope to mirror some of these inspiring qualities in my research which will continue to guide me well beyond my PhD. Nicolas, thank you for your intense supervision especially in the final stages of this work. This journey would not have been possible without your invaluable advises and motivation. I really appreciate your scientific and insightful comments and questions which gave me additional perspectives of my work and positively impacted my PhD thesis.

I would also like to thank *Institut des Systèmes Intelligents et de Robotique (ISIR)* and *Sciences et Technologies de la Musique et du Son (STMS)* at *Institut de recherche et coordination acoustique/musique (IRCAM)* for providing offices, equipment and a creative research environment.

Some individuals deserve special thanks. Thanks to Professor Axel Roebel who provided continuous bug reports, suggestions, and technical support. Special thanks to my colleagues at ISIR and IRCAM. I have been really lucky to have crossed paths with many talented collaborators who have taught me a lot and made this project a lot of fun - Clement Le Moine Veillon, Léane Salais, Yann Teytaut, Michele Grimaldi, Jieyeon Woo, Liu Yang, Fajrian Yunus, Fabien Boucaud, Sooraj Krishna and Reshmashree Kantharaju. I would like to thank them for their great spirit, inspiration and motivation.

I am especially indebted to Professor Joe Tekli who was my first research supervisor during my undergraduate studies in computer engineering, from whom I have learned a lot. The research I have conducted under his supervision has added a lot of value to my thesis.

And my biggest thanks goes to my parents - my mother Jeanette, my father Dany - and my brother Georges for all the support they have shown me through this research, who always motivated me to reach for the stars.

Contents

List of Figures

List of Tables

Nomenclature

General

$(.)$	set
[.]	sequence
ϵ	epsilon

Math Operators

$\cdot$	euclidean dot product operator
x	scalar
$\|x\|$	magnitude of x
X	matrix
$X^\intercal$	transpose of X
$\widehat{X}$	estimate of X
$sa(.)$	self attention
$A(.)$	cross attention

Corpora Features

AU	action unit
$AU1$	inner brow raiser
$AU2$	outer brow raiser
$AU4$	brow lowerer
$AU5$	upper lid raiser
$AU6$	cheek raiser
$AU7$	lid tightener

R_X roll euler angle

R_Y pitch euler angle

R_Z yaw euler angle

f_0 fundamental frequency

$Jitt$ jitter

$Shimm$ shimmer

HNR harmonic-to-noise ratio

$Hamm$ hammarberg index

$BERT\ Embeddings$ bert embeddings

Segmental Units

W word unit

IPU Inter-Pausal Unit

S segment defined by a number of frames

t duration of a segment in terms of frames

Speakers Data Sets

SD speaker dependent

SI speaker independent

Networks Components, Layers, and Units

E encoder

D decoder

C classifier

G generator

Dis discriminator

N_{enc} number of encoding layers

N_{dec} number of decoding layers

$\mathbf{N}_{hid}$ number of hidden layers

d_{model} embedding dimension

d_{att} attention layer embedding size

N_{lay} number of layers

K_{size} kernel size for convolutional neural networks

N_{filt} number of filters for convolutional neural networks

$Drop$ dropout

L^{BI} bi-directional lstm layer

General Training Hyperparameters

α alpha in LeakyReLU activation function

BS batch size

β_1 adam beta 1 parameter

β_2 adam beta 2 parameter

N_{ep} number of training epochs

N_h number of attention heads

N_{it} number of training iterations

Lr learning rate

Lr_{init} initial learning rate

Lr_e end learning rate

St_{size} step size

W_{steps} number of warmup steps

LSTM and Transformer-based Gesture Synthesis Models

X_{face} input facial gestures sequence

X_{speech} input fundamental frequency sequence

N_W the number of f_0-values corresponding to the spoken word $\boldsymbol{W}$

X_{text} input bert text embeddings

$X^{(1)}_{speech}$ encoded fundamental frequency sequence

$\widehat{Z}_{AU}$ output action unit sequence

$\widehat{Z}_R$ output head rotation sequence

$E^{L^{BI}}_{speech}$ bi-directional lstm-based encoder for encoding input $X^{(1)}_{speech}$

$E^{L^{BI}}_{text}$ bi-directional lstm-based encoder for encoding input X_{text}

E_{face} action units and head rotation encoder

E_{speech} fundamental frequency encoder

$E_{\textbf{\textit{speech, text}}}$ fundamental frequency and text encoder

D_{face} action units and head rotation decoder

D_{speech} fundamental frequency decoder

h_{speech} the latent representation of input speech X_{speech}

h_{face} the latent representation of input facial gestures X_{face}

AE_{speech} fundamental frequency auto-encoder

AE_{face} facial features auto-encoder

ZS-MSTM 1.0 and 2.0 Models

h_{style} style embedding

$h_{content}$ content embedding

$E_{content}$ content encoder

E_{style} style encoder

E^{face}_{style}, 2D facial landmarks encoder

E^{pose}_{style} 2D pose encoder

E^{style}_{tags} dialog tags encoder

E^{style}_{speech} speech Mel spectrogram encoder

X_{face} input 2D facial landmarks

X_{pose} input 2D pose

X_{tags} input dialog tags

X_{text} input bert text embeddings

X_{speech} input speech Mel spectrograms

$\widehat{Z}_{face}$ output 2D facial landmarks

$\widehat{Z}_{pose}$ output 2D pose

X_{source} source speaker gestures

$\widehat{Z}_{\textit{gestures}}$ output gestures predictions

Chapter 1

Introduction

Contents

Man with all his noble qualities, with sympathy which feels for the most debased, with benevolence which extends not only to other men but to the humblest living creature, with his God-like intellect which has penetrated into the movements and constitution of the solar system—with all these exalted powers—Man still bears in his bodily frame the indelible stamp of his lowly origin.

Charles Darwin (1871/1898)

Human beings, in contrast to other species, do have the ability to think systematically and creatively about techniques. By virtue of humanity's nature as a toolmaker, humans have been technologists from the beginning of humanlike life (Bogin and Varea [2020]). The history of technology encompasses the whole evolution of humankind. The evolution of technology has introduced the notion of "*digital humanism*" that leverages on the connection between human and humanoids (Davies [2016], Wagner et al. [2020]).

The field of research in *Embodied Conversational Agents (ECAs)* has emerged as new interface between humans and machines. ECAs behaviors are often modeled from human communicative behaviors. They are endowed with the capacities to recognize and generate verbal and non-verbal cues (Lugrin [2021]). ECAs are envisioned to support humans in their daily lives.

This thesis revolves around modeling multimodal data and learning the complex correlations between the different modalities employed in human communication to leverage the *behavioral expressivity* of ECAs and control their *behavioral style*. More specifically, the objectives of this thesis are:

1. to exploit *speech prosody*, *visual prosody* and *language* with the aim of synthesizing expressive and human-like behaviors for ECAs;

2. to control the style of the synthesized gestures such that we can generate them with the style of any speaker.

This Chapter first introduces Embodied Conversational Agents (ECAs), which are the core animated agents used in this thesis. Then, we dive into the different challenges of the ECA research. We then discuss the existing generative models that were developed for *gesture synthesis*, and *style transfer for gesture animation*, as well as the limitations of these works. Next, we explain the thesis scope, research questions, and our contributions. The structure of the Manuscript is then outlined. Finally, we list all the publications that were published in the context of this thesis.

1.1 Embodied Conversational Agents

Socially Intelligent Agents (SIA) (see Figure 1.1) are virtually or physically embodied agents that are capable of autonomously communicating with people in a socially intelligent manner using multimodal behaviors (Lugrin [2021]). They have been evolved under different names such as *Embodied Conversational Agents*, or *Social Robotics*. They include both physical and virtual embodied agents.

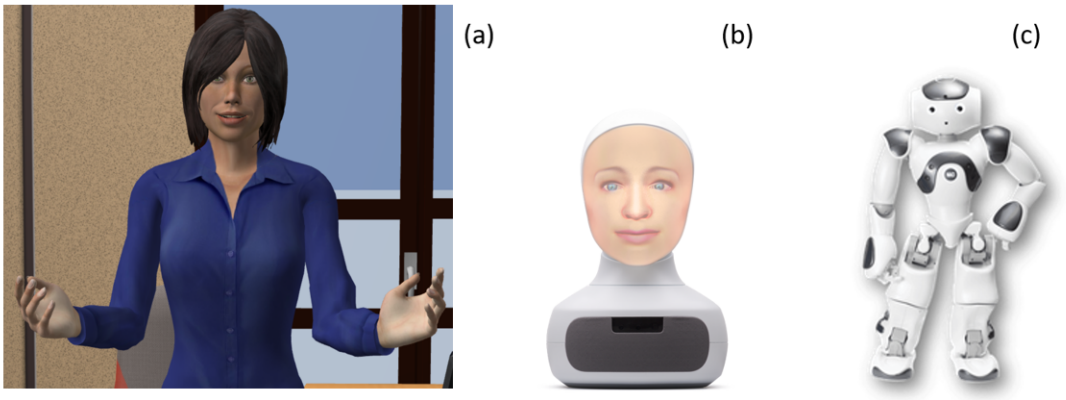


Figure 1.1 Illustration of (a) Greta, an embodied conversational agent (Pelachaud [2015]), (b) Furhat social robot (Al Moubayed et al. [2013]), and (c) a humanoid NAO robot (Shamsuddin et al. [2011])

The virtual Embodied Conversational Agents (ECA) (Figure 1.1 (a)) originated from the idea of simulating multimodal human communication where the modalities are those of human communication. These agents are autonomous animated characters that emulate human behavior and communication (Ruttkay and Pelachaud [2004]). They display similar properties as humans employ during conversation. These properties include the capacity to generate and respond *verbally* and *non-verbally* (Ball and Breese [2000]). Embodied agents form a type of software agents that can interact with humans as computers. With the advances in technology, these agents are nowadays used in myriad applications. They are being developed for many applications such as assistance, education, entertainment and health. For instance, ECAs can be employed in educational and training systems to bring knowledge and skills.

Communication beyond words. Communication involves spoken language and multimodal non-verbal behavior (Giles [2016]). The non-verbal behavior of virtual agents is crucial to render them lifelike, and allow a broader and efficient human-like communication and expressivity. Gestures can visually illustrate many aspects of the spoken message, and can convey additional information to the message; for instance indicating the size of a box with a hand gesture (McNeill et al. [2005]). An ECA exhibiting human-like appropriate facial expression, gaze, posture and body gestures is more expressive and engaging (Lugrin [2021]).

1.1.1 Research Aims

Embodied agents constitute potentially a beneficial interface between humans and computers. However, for humans to perceive ECAs natural and engaging, ECAs must exhibit a human-like behavior. A major goal in ECA research is to render agents *believable* and *human-like*. The ECA research field aims to simulate the human multimodal behavior and reproduce human-like expressive gestural and visual prosody in ECAs. Many advances in this field have been made, however this outgrowing field still has many challenges. To enable a smooth and engaging interaction between humans and ECAs, it is crucial to consider different major challenges:

- Humans communicate *verbally* and *non-verbally*. Speech, voice *prosody*, facial and body gestures are continuously employed during communication. For the ECAs' behavior to be *human-like*, they should display *multimodal behaviors* when communicating. More specifically, the computational models that drive their behaviors should learn the complex relationships that exist between *speech prosody*, *visual prosody* and *text semantics*, and learn the mappings between these modalities. This is achieved through *multimodal modeling* meaning modeling the combination of the different modalities. Therefore, multimodal channels of communication should be considered thoroughly when developing generative models for gesture synthesis.

- Researchers currently face the issues of how to build models that compute gestures given their communicative intentions or emotions, then perform these gestures in coordination with speech in order to convey their communicative meaning. In addition to that, the gestures produced by the agent should be *coherent* with what is being said.

- Besides producing natural and coherent gestures, modeling the temporal relationship between speech and gestures is a great challenge. ECAs should be able to produce gestures in conjunction with speech. That is, gestures should be *aligned* and *synchronized* with speech and their shapes should be linked to the speech content. More specifically, gestures and speech should be *temporally* aligned.

- Movements and gestures are person-specific and *idiosyncratic* in nature (McNeill et al. [2005]), and each speaker has his or her own *behavioral style* that is linked to his/her personality, role, culture, etc. Different speakers may gesture differently when saying the same utterance. The same speaker may have different gesturing styles depending on the *situation* he or she is in, his/her role, the interlocutors, etc. Behavior generation models should learn each speaker's unique behavioral style. Modeling human behavioral style should take into account the multimodal aspect of *behavioral style* that is found within and across the multiple modalities - visual gesture prosody, speech prosody, and text context - while learning the diversity of gestures of each speaker.

1.2 Existing Works and Limitations

Below we discuss the existing generative models that were developed for *gesture synthesis*, and *stylized gesture animation*. We specifically pay attention to the limitations of these works.

1.2.1 Gesture Synthesis Models

Various gesture generation approaches were previously proposed.

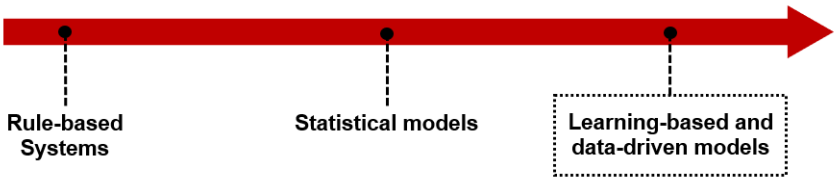


Figure 1.2 Timeline of gesture generation approaches

The earliest approaches are based on a set of rules that refer to existing correspondences between patterns produced by human communication and behavior. Such approaches are known as *rule-based*. The main limitation of *rule-based* approaches is that they require considerable human effort to determine the rules. The produced gestures are limited, not diverse and lack variability. Moreover, the collection and implementation of the rules is time-consuming and requires lot of resources. Later on, to overcome the limitations of *rule-based* approaches, researchers turned to developing *statistical* approaches that can synthesize gestures based on the statistics in a given corpus of human non-verbal behavior. *Statistical* models did overcome some of the *rule-based* limitations, however the produced gestures still suffered from a lack of diversity and variability.

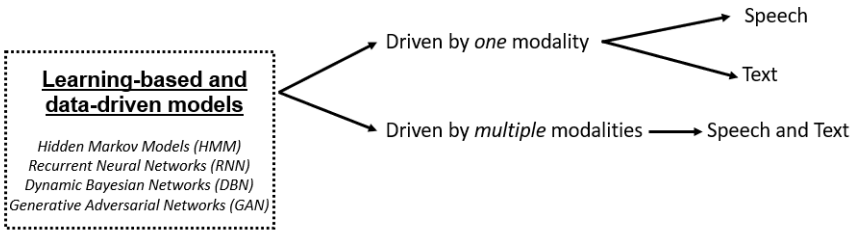


Figure 1.3 Learning-based approaches

In the last few years, *learning-based* models were proposed. These models are learned from large amounts of data (they're often referred to as *data-driven*), and are based on machine learning algorithms. A large number of gesture generative models have been proposed, principally based on sequential generative parametric models such as *Hidden Markov Models* (HMM) and gradually moving towards *Deep Neural Networks* (DNN) enabling spectacular advances over the last few years. A variety of generative statistical models aimed to predict the multimodal behavior of an ECA. For instance, HMMs (Hofer and Shimodaira [2007]), Recurrent Neural Networks (RNN) (Wang et al. [2021], Haag and Shimodaira [2016]), and Dynamic Bayesian Networks (DBN) (Marioryad and Busso [2012], Sadoughi and Busso [2019]) have been used to generate ***head motion*** from speech; Generative Adversarial Networks (GAN) (Karras et al. [2017], Vougioukas et al. [2019]) have been proposed to produce ***facial gestures*** from speech. The main limitation of most of these works is that they exploit as input only **one** modality of human communication which, in most cases, is *speech*. Some works (Ishi et al. [2018], Yoon et al. [2019]) use text transcription of language to synthesize gestures. However, gestures are

a function of both *speech prosody* and *language*, especially *facial*, *hand* and *body* gestures. Other works (Kucherenko et al. [2020], Yoon et al. [2020] and Ahuja et al. [2020a]) fused both modalities - speech prosody and language - for synthesizing gestures. However, these works have some limitations:

1. First, they train their models on datasets containing *one* speaker data.
2. Second, their work is only focused on synthesizing hand gestures, without taking into account *facial* gestures.
3. Third, most of them model one type of gesturing, without considering the correlation between modalities. For instance, *facial expressions* and *head movements* are highly correlated to prosody (Yehia et al. [2002]) - more specifically, the fundamental frequency f_0. Modeling them together allows capturing the multimodal correlation.

1.2.2 Style Transfer for Gesture Animation Models

A large number of other generative models were proposed in the past few years for synthesizing gestures in the behavioral style of specific speakers. They assume that *behavioral style* is encoded in the *body gesturing*. Some of these works generate full body gesture animation driven by text in the style of one specific speaker (Neff et al. [2008]). Other approaches (Alexanderson et al. [2020], Karras et al. [2017], Cudeiro et al. [2019], Ginosar et al. [2019]) are speech-driven. For some of these approaches, the behavioral style of the synthesized gestures is changed by exerting direct control over the synthesized gestures' velocity and force (Alexanderson et al. [2020]). For others (Cudeiro et al. [2019], Karras et al. [2017], Ginosar et al. [2019]), they produce the gestures in the style of a *single speaker* by training their generative models on one *single speaker*'s data, and synthesizing the gestures corresponding to this specific speaker's audio.

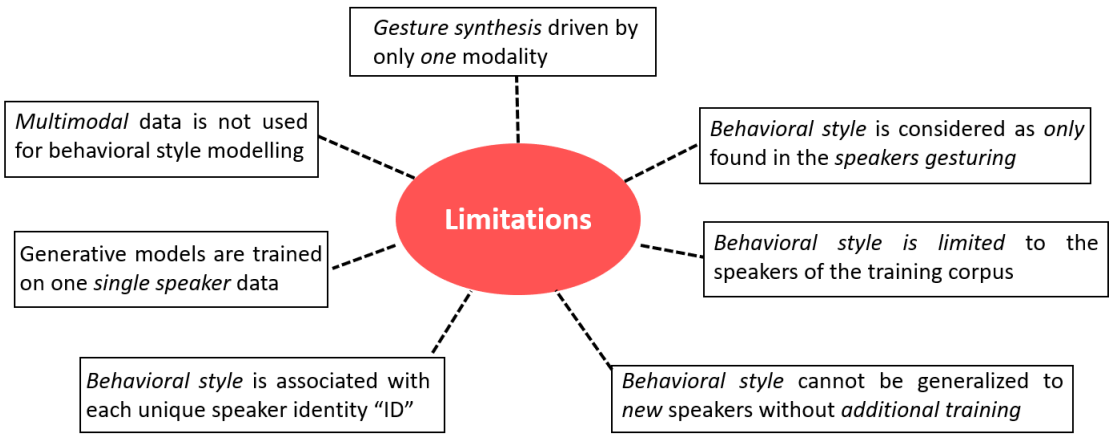


Figure 1.4 Limitations of previous generative models for synthesizing gestures in the behavioral style of specific speakers. The limitations include the challenge of generalizing behavioral style to new speakers without additional training, reliance solely on speakers' gesturing as the primary source of behavioral style, the neglect of multimodal data for modeling behavioral style, the use of generative models trained on data from a single speaker, and the association of behavioral style with each unique speaker identity.

The main limitation of these works is that they have focused on generating gestures (facial and head gestures in particular) that are aligned with either speech or text. They did not exploit *multimodal* data neither for modeling gesture synthesis, nor for modeling the *behavioral style* of speakers. Moreover, their generative models are trained on one *single speaker* data. The only attempts to model and transfer the style from a *multi-speakers* database (Ahuja et al. [2020b] and Ahuja et al. [2022]) are only *speech-driven*. Their approaches do not exploit language (text) information for synthesizing the gestures. To capture the *behavioral style*, their models considered only upper-body motion. In their approaches, *behavioral style* is associated with a unique speaker identity "ID". These approaches can only simulate *behavioral style* of speakers that were seen during training. They cannot generalize *behavioral style* to new speakers without training and fine-tuning their model (Ahuja et al. [2022]). They perform "neural domain adaptation" between a *source* speaker style and a specific *target* speaker style. This adaptation requires additional training.

1.3 Thesis Scope

The central theme of this thesis is to model the relationships existing between human visual prosody, speech prosody, and spoken verbal language, to build ECAs that can communicate expressively with humans. Our objective is to capture the relationship between communicative modalities and to develop generative models that can synthesize expressive visual prosody for ECAs. Another objective is to model human multimodal behavioral style to control and adapt the ECA's synthesized gestures to any behavioral style, by transferring the behavioral style from one target speaker to another source one.

Below we dive deeper into the different objectives of this thesis and the related research questions.

1.3.1 Semantically-Aware and Speech Driven Multimodal Gesture Synthesis

How can we model relationships between visual prosody, speech prosody, language, and the speech context (text semantics) so that we can synthesize human-like multimodal gestures? It is challenging to answer this question and understand these relationships due to the complex relationships present in human verbal and non-verbal signals that are emitted during speech. Expressive visual prosody is indeed an issue in current generative models. The *first* objective of this thesis aims to develop a model that can synthesize human-like expressive visual prosody, driven by *speech prosody*, and *language (text)*. The expressivity and human-likeness of the produced gestures are leveraged by effectively exploiting the human multimodal behavior data given as input to the model. The goal is to render the generated gestures human-like, synchronized with the given input speech and aligned with the text content.

Research Questions

Concretely, we study the following research questions:

1. *Modeling multimodal behavior* - How can we exploit human multimodal behavior - speech prosody, visual prosody, and language - to generate expressive and human-like *facial* and *body* visual prosody in Embodied Conversational Agents? How to

design computational models that can capture the relationship between these different modalities?

2. *Generalization* - How can we generalize the learned gestural latent space to new speakers data, that are unseen during the training phase of our generative model?

1.3.2 Style Transfer for Gesture Animation

Gestures have unique semiotic properties, they are *idiosyncratic*, and are chosen on the spot by speakers as they speak (McNeill et al. [2005]). Each speaker has a different gesturing and behavioral style when communicating. Speakers sharing the same role and settings (i.e. TV hosts, lecturers, televangelists, etc.) may have some common gestural style. Modeling behavioral style is a key challenge and a complex problem, since it is multimodal (Knapp et al. [2013]), and found in verbal and non-verbal behavior (Campbell-Kibler et al. [2006], Moon et al. [2022], Obin [2011], Obermeier et al. [2015], Wagner et al. [2014]). The *second* objective of this thesis is to model human behavioral style. The goal is to be able to synthesize gestures that accompany spoken language within a certain context, while learning the uniqueness of each speaker's style.

Style modeling and control in gesture is receiving more attention in order to propose more expressive behaviors that could possibly be adapted to a specific audience (Neff et al. [2008], Karras et al. [2017], Cudeiro et al. [2019], Ahuja et al. [2020b], Ginosar et al. [2019], Alexanderson et al. [2020], Ahuja et al. [2022]). Another goal is to be able to control the style of the synthesized gestures, and to perform *style transfer* amongst different speakers that are seen by our model during the training phase, and speakers that have not been seen during the training phase (*zero-shot style transfer*). For instance, let's consider two different speakers Alice and Bob. Alice is speaking with her own gesturing style. Bob is gesturing with his own. Both of them are likely to gesticulate differently. Our goal is to transfer the gestural style from Alice to Bob, so that Bob can follow the same gesturing style as Alice while speaking any utterance. Modeling *behavioral style* requires learning a style space based on multimodal speakers data. The challenge is to build this style space, that is independent from speakers' identity (which in most previous works is defined by their "ID"), and is only dependent on *seen* speakers multimodal data - that our model has *seen* during training. It will enable us to generalize *behavioral style* to new *unseen* speakers, and hence it will allow us to perform zero-shot style transfer from unseen target speakers to other speakers. To formalize this problem, we aim to learn style from multimodal data - speech prosody, visual prosody, speech semantics and dialog acts - without using speakers' identifiers. Moreover, we want the model to be able to generalize behavioral style to speakers that have not been seen during training, **without any further training or fine-tuning**, thus allowing us to perform *zero-shot* style transfer.

Research Questions

Concretely, we study the following research questions:

3. *Multimodal style modeling.* How can we learn style latent space of given speakers, given their multimodal data, and independently from their identity?

4. *Style transfer.* How can we synthesize facial and body gestures of a source speaker, given the source speaker multimodal data, but with the style of another speaker?

5. *Generalization of style to unseen speakers - Zero-shot style transfer.* How can we render our approach able to perform zero-shot style transfer on new unseen speakers, without the need of any further training or fine-tuning?

1.4 Thesis Contributions

The main theme of this thesis is to model the existing complex relationships in *human multimodal behavior*, so that embodied agents can have similar *behavioral expressivity*. The objectives are twofold: (1) to synthesize expressive multimodal gestures for ECAs, by learning the relations between visual, speech prosody and speech content, and (2) to model *human multimodal behavioral style* and be able to adapt the ECA's synthesized gestures to any *human behavioral style*. For the purpose of tackling these objectives and addressing the different previously mentioned limitations and technical challenges, we propose different models and datasets that are discussed below in detail.

1.4.1 Developing TEDx Corpus and Extending PATS Corpus (Chapter 4)

The first contribution of this thesis is the development of the *TEDx Corpus*, as well as the extension of the *PATS Corpus* (Ahuja et al. [2020b]) for the purpose of using them in our different studies related to modeling multimodal gesture synthesis and human behavioral style. These corpora are presented and discussed in Chapter 4.

- The first corpus is the *TEDx Corpus*, a corpus that we collected to train, test, and validate our models that were developed to generate speech-driven and semantically-aware facial gestures (research questions Q1, and Q2). More specifically, the *TEDx Corpus* provides a large amount of TEDx aligned data related to three modalities: *speech audio*, *text semantics*, and *facial cues* - more precisely *eyebrow gestures* and *head motion*.

- The second corpus is the *PATS corpus* (Ahuja et al. [2020b]), a corpus that was previously built by Ahuja et al. [2020b] and which we extended to include additional features. This corpus was used in this thesis to train, test, and validate our models that were developed to tackle the research questions related to *body visual prosody expressivity* (research questions Q1, and Q2) and *multimodal style modeling and transfer* (research questions Q3, Q4, and Q5). The *PATS Corpus* was originally proposed to study the correlations between multimodal features related to *speech audio*, *text semantics* and *body pose gestures*. We extended it to include additional multimodal features related to *2D facial landmarks*, and *dialog tags* - tags of the dialog acts of the spoken utterances.

Part of Chapter 4 appeared in the proceedings of ACM International Conference on Multimodal Interaction 2020 (Fares [2020]) and "Workshop sur les Affects, Compagnons Artificiels et Interactions" 2021 (Fares et al. [2021]).

1.4.2 Semantically-aware and speech-driven facial gestures (Chapter 5)

Generating coherent human-like facial expressivity in ECAs is crucial. To address the research questions Q1, and Q2 discussed earlier, we propose two learning-based models for synthesizing facial gestures including eyebrows motion and head rotations. Eyebrow

motions are represented by facial Action Units, and their intensities as described by the FACS (HAGER [2002]), which are described in details in Chapter 2. We exploit both speech and text modalities to generate co-expressive human-like eyebrows motion and head gestures (Chapter 5). More specifically, we propose two different networks:

- As a starting point, we propose an *end-to-end LSTM neural network architecture* (presented in Chapter 5) that predicts upper-face gestures, based on both *speech prosody* and *text semantics*. It generates sequences of Action Units related to eyebrows and eyelids movements. The first LSTM-based model that was developed is used as a *baseline* when evaluating our second model.
- The second model is a novel approach that makes use of *Transformers* and *Convolutions* to synthesize upper-facial gestures as well as head rotations (presented in Chapter 5). Head and eyebrow gestures are both correlated to speech prosody (Yehia et al. [2002]). We additionally model the correlation between head motion and upper facial gestures to allow the generation of a more coherent and natural behavior of the agent.

For both models, the synthesized gestures are based on different modalities, specifically *audio data*, *text data*, and *facial data*. More specifically, the synthesized facial gestures are generated based on *speech prosody* and *text semantics*. To the best of our knowledge, predicting facial movements based on both speech prosody and text semantics have not yet been investigated. We used the *TEDx Corpus* that we have gathered to train our model, and which is discussed in Chapter 4. We conduct several objective and subjective evaluations to validate our approach. Objective evaluations aimed to assess the errors generated by our models, the correlation of the synthesized gestures w.r.t the ground truth, and the activation / non-activation of the Action Units. Subjective evaluations aimed to assess the appropriateness, coherence, naturalness, synchronization and alignment of the synthesized gestures with the given input modalities data. We show that using both modalities leverages the quality of the results.

Part of Chapter 5 appeared in the proceedings of *ACM International Conference on Multimodal Interaction 2020* (Fares [2020]) and *the European Signal Processing Conference 2022 (Fares et al. [2022])*.

1.4.3 Zero-shot style transfer for *body pose* and *facial* gestures synthesis (Chapter 6 and 7)

Another contribution made in this thesis is the development of ***ZS-MSTM 1.0*** and ***ZS-MSTM 2.0*** for modeling ECAs with *behavioral style*. Both approaches allow zero-shot multimodal style transfer for *2D body pose* (in ***ZS-MSTM 1.0 and 2.0***) and *2D facial landmarks* (in ***ZS-MSTM 2.0***) synthesis. To address the research questions Q3, Q4, and Q5, we propose an efficient yet effective machine learning approach to synthesize gestures driven by *prosodic features* and *text* in the style of different speakers including those unseen during training.

ZS-MSTM 1.0 produces stylized upper-body gestures, driven by the *content* of a source speaker's speech - text semantics embeddings and audio Mel spectrogram - and conditioned on a target speaker's *multimodal style embedding*. The stylized generated gestures

correspond to the style of target speakers that can have been *seen* or *unseen* during training. This model allows us to directly infer an embedding style vector from multimodal data (text semantics, speech and pose) of any speaker, by simple projection into the embedding style space. The style transfer performed by our model allows the transfer of style from any unseen speakers, without requiring any further training or fine-tuning. In contrast to previous works, the learned style space is independent from speaker's identity "ID", which allows our model to generalize on new unseen speakers' data without any further fine-tuning, rendering our approach *zero-shot*. The proposed architectures in previous works are based on the disentangling of *content* and *style* information, which is based on the assumption that *style* is only encoded in gestures. However, both *text* and *speech* also convey *style* information, and the embedding of style must take into account all the modalities of human behavior. The style space is learnt from 16 *PATS* speakers but can extrapolate to new speakers that were not present during training, and therefore unseen by our model. More specifically, in this latent space, the learned distribution of the style vectors allow to have good representations of the behavioral style of the speakers that were seen during training. When testing on speakers that were never seen during training, the model is still capable of generating good behavioral style vectors for these unseen speakers.Objective and subjective evaluations are conducted to evaluate our approach and validate it.

ZS-MSTM 2.0 is an extension of ***ZS-MSTM 1.0***: it includes the synthesis of *2D facial landmarks*, leading to a model that can synthesize 2D body poses and facial animation. ***ZS-MSTM 2.0*** is the first model allowing the synthesis of 2D body poses and facial motion. We conduct an objective evaluation to assess ***ZS-MSTM 2.0***.

ZS-MSTM 1.0 is presented in Chapter 6 and ***ZS-MSTM 2.0*** is presented in Chapter 7. Part of Chapter 6 will appear in *SIVA'23, workshop of the IEEE International Conference on Automatic Face and Gesture Recognition 2023*.

1.5 Thesis Outline

This thesis is organized in 3 parts:

1. The first part consisting of Chapter 2 that establishes the necessary background knowledge for *multimodal communication*, and Chapter 3 which gives an overview of the existing gesture generation approaches, discusses and analyzes the approaches, their underlying principles, and their limitations.

2. The second part consists of Chapter 4 and Chapter 5, which are related to *Multimodal Gesture Synthesis*. We present in Chapter 4 two corpora: (1) TEDx Corpus which we have built and (2) *PATS Corpus* which we have extended. Moving forward, Chapter 5 presents our LSTM-based model, which is our baseline for synthesizing speech-driven and semantically-aware facial gestures; and a second novel approach for synthesizing facial gestures that makes use of *Transformers* and *Convolutions*.

3. Part three consists of two chapters related to synthesizing stylized facial and body gestures. Chapter 6 presents **ZS-MSTM 1.0**, a model that synthesizes the upper-body gestures of a speaker in the style of a second target speaker that could be *seen* or *unseen* during training. Chapter 7 presents our final model **ZS-MSTM 2.0**, which

is an extension of **ZS-MSTM 1.0**, and it generates *facial* and *body* gestures of a source speaker in the style of another target one in a zero-shot fashion.

1.6 Publications and Submissions

- Fares, M., Pelachaud, C., and Obin, N. (2022, August). Transformer Network for Semantically-Aware and Speech-Driven Upper-Face Generation. In EUSIPCO 2022.
- Fares, M. (2020, October). Towards multimodal human-like characteristics and expressive visual prosody in virtual agents. In Proceedings of the 2020 International Conference on Multimodal Interaction (pp. 743-747).
- Fares, M., Pelachaud, C., and Obin, N. (2020, June). Multimodal modeling of expressiveness for human-machine interaction. In Workshop sur les Affects, Compagnons artificiels et Interactions.
- Fares, M., Pelachaud, C., and Obin, N. (2021, October). Multimodal-Based Upper Facial Gestures SynThesis for Engaging Virtual Agents. In Workshop sur les Affects, Compagnons Artificiels et Interactions.
- Fares, M., Pelachaud, C. and Obin, N. Zero-Shot Style Transfer for Multimodal Data-Driven Gesture SynThesis. In SIVA'23.
- Fares, M., Pelachaud, C. and Obin, N., I-Brow: Hierarchical and Multi-Modal Transformer Model for Eyebrows Animation Synthesis. In the 25th International Conference on Human-Computer Interaction
- Fares, M., Grimaldi, M., Pelachaud, C., and Obin, N. (2022). Zero-Shot Style Transfer for Gesture Animation driven by Text and Speech using Adversarial Disentanglement of Multimodal Style Encoding. arXiv preprint arXiv:2208.01917.
- *Under Review for Publication (as of Feb 2023)* - Fares, M., Pelachaud, C. and Obin, N., Zero-Shot Style Transfer for Body Gesture Synthesis

The key points of this Chapter:

Goal of this thesis

- The central theme of this thesis is to model the relationship between human visual prosody, speech prosody, and spoken verbal language to build ECAs that can naturally communicate verbally and non-verbally with humans.
- Our objective is to exploit these mechanisms to render visual prosody in ECAs.
- Another objective is to model human multimodal behavioral style and to control and adapt the ECA's synthesized gestures to any behavioral style, by transferring the behavioral style from one target speaker to another source one.

Thesis Research Questions

- *Modeling multimodal behavior* - How can we exploit human multimodal behavior - speech prosody, visual prosody, and language - to generate expressive and human-like facial and body visual prosody in ECAs? How to design computational models that can capture the relationship between these different modalities?
- *Generalization* - How can we generalize the learned gestural latent space to new speakers data, that are unseen by our generative model?
- *Multimodal style modeling*. How can we learn style latent space of given speakers, given their multimodal data, and independently from speakers' identity?
- *Style transfer*. How can we synthesize facial and body gestures of a source speaker, given the source speaker multimodal data, but in the style of another speaker?
- *Generalization of style to unseen speakers - Zero-shot style transfer*. How can we render our approach able to perform zero-shot style transfer on new unseen speakers, without the need of any further training or fine-tuning?

Part I

Background and Related Work

Chapter 2

Introduction to Multimodal Communication and Behavior Style

Contents

> *Gesture and speech are available as two separate modes of representation and are coordinated because both are being guided by the same overall aim. That aim is to produce a pattern of action that will accomplish the representation of a meaning.*
>
> Adam Kendon - 1983

Human communication is a complex system that involves verbal and non-verbal channels of communication, where the burden of information is conveyed through multiple channels. Non-verbal vocalizations, hand gestures, body gestures, head motion and facial expressions are emitted during speech, they are closely integrated with the speaker's words, and may emphasize and disambiguate them. Before diving into the core study of this research, in this Chapter we turn to introduce the different modalities of communication used in human communication to understand the relationships between them, the different mechanisms governing them and how they cooperate with each-other for the purpose of expressive communication. We first discuss the origins of human behavior, then we dive into the different modalities employed during human communication: *speech prosody* which carries lexical information, *visual prosody* expressed by *facial* and *body* gestures, and *language*, which is the *verbal* channel of communication. Finally, we turn to discuss *human behavioral style* and its presence in the *multimodal* channels.

2.1 Phylogenetic Origins of Human Behavior

From an evolutionary point of view, humans are primates whose communication system has evolved during a long and shared phylogenetic history. Human communication is a complex *system of systems*, that is strikingly different from any other known natural communication system (McNeill et al. [2005], Levinson and Holler [2014], Argyle [2013]). It is the process of sending and receiving messages through multiple modalities. Multimodal human communication involves a series of verbal and non-verbal cues that are tightly related to speech content (Argyle [2013], Feyereisen et al. [1991], Armstrong et al. [1995]). Visual and speech prosody are important channels of communication, encompassing non-verbal vocalizations, and non-verbal cues, which are emitted during communication, some consciously and some unconsciously, to convey speech meaning.

Human beings are primates, as are our cousins the great apes - chimpanzees, gorillas, orangutans - and monkeys. Human behavior has phylogenetic origins (Knapp et al. [2013], Argyle [2013]), it is similar to our nonhuman primate relatives manifesting behaviors. An important support for the theory of evolution of Charles Darwin (Darwin [1998]) was the evidence of similarities in expressive behavior across various species. The process of evolutionary advancement was demonstrated by the increasing use of face, voice, and body for the purpose of communication and for expressing emotions. *Expressivity* was considered by Darwin as a key link in the argument of evolution. For Darwin, a rich repertoire of expressive and signaling behaviors is linked to the complexity of a species' social organization. The common biological and social problems that human and non-human primates encounter explain the similarities in their behavior. For instance, chimpanzees, like humans, show empathy to those who are suffering, and reconcile after a fight by the mean of a touch or an embrace (Ladygina-Kohts et al. [2002]). Research on primate communication has shown that great apes use gestures to communicate different intentions

(Byrne et al. [2017]).

Non-verbal communication - humans V. animals. The behavioral similarities between humans and non-human primates gave birth to the research in *non-verbal communication* (Argyle [2013]). The evolutionary origins of animal behavior can be traced, which explain the behavioral similarities with humans. However, humans are very different, and the main difference is in the usage of *language*. *Language* is a complex expressive system built on speech, and its presence or absence is the primary difference between animal and human communication systems (Levinson and Holler [2014]). Animal communication is mainly about their internal intentions and states, whereas humans conversations are about people, events, or the past and the future (Argyle [2013], Knapp et al. [2013]). A new set of non-verbal cues and signals arose with the use of *language* - they are produced in conjunction with speech, to accompany it, provide feedbacks, and cope with the alignment of utterances. Humans used non-verbal communication for communicating emotions and controlling interpersonal relations, and interestingly, they maintained the uses of non-verbal communication throughout their evolution (Argyle [2013]). Humans also differ from animals in the unrivalled complexity and expressivity on the one hand (Levinson and Holler [2014]), as well as in planning their social behavior which comprises of *social acts* - behavior that is planned, with a certain goal in mind, often with words (Argyle [2013]). There is a hierarchical structure that encompasses basic non-verbal cues into these social acts.

The origins of human language. The core ecology for human language use is in face to face interactions (Levinson and Holler [2014]). This is where languages are learnt, and in this niche, language not only involves vocal tract and lungs, but also the speaker's face, eyes, trunk, gaze, and hands (Feyereisen et al. [1991], Levinson and Holler [2014], Argyle [2013]). The non-verbal signals that are produced during communication may be grouped into "*channels*" or "*modalities*" of communication (Feyereisen et al. [1991]). The speaker generates a *multimodal* display that is partly semiotic and partly entrained by the vocal synthesis.

Gesture and human language co-evolution. Contrary to the "*gesture-first*" theory that claims that language started as a gesture language and was gradually supplanted by speech, David McNeill (McNeill et al. [2005]) argues that gesture and language belong to a single system of verbalized thinking and communication, and they both cannot be considered as the twin of the other. In this system both gesture and language are two different modalities that are crucial and constitute together adult human language, as one system. Despite the different roles each modality has in human communication, the whole ensemble should be considered as a "*system of systems*" (McNeill et al. [2005], Levinson and Holler [2014], Argyle [2013]), that have gathered over the two and half million years that humans have been cognitively progressive and tool-using species. During their co-evolution, gesture and speech have mutually adjusted to one another (Levinson and Holler [2014]).

2.2 Multimodal Human Communication - A "*System of Systems*"

Human language happens embedded within an interchange of multimodal signals. The different modalities play different roles in human communication, but function as one integrated system. As indicated above, human communication is a *system of systems*, and the burden of information can be switched from one system to another, and can be conveyed through multiple systems. For instance, human language changes the *channel* or *modality* of communication transferring lexical material from mouth to hands, as in deaf sign languages (Liddell and Metzger [1998]).

A series of verbal and non-verbal signals are emitted during communication, and they are closely linked to the underlying intention to be communicated (Argyle [2013], Feyereisen et al. [1991], Armstrong et al. [1995]).

Verbal communication. Verbal signals are expressed in form of spoken language that is a constituent to convey the speaker's meaning and intent. *Language* is a communication tool used by humans to express their ideas and convey their emotions. Linguistic communication is primarily achieved by combining language units such as words, sounds, and utterances. *Text* is the term used to refer to these combinations of language units. It is a sequence of words that have meanings that are produced and interpreted in a given *context*. Communication does not take place in a vacuum, but in a certain context. According to Levinson et al. [1983], context is understood to cover the identities of participants, the temporal and spatial parameters of the speech event, and the beliefs, knowledge and intentions of the participants in that speech event; other elements related to the spatial context, the role and relationship of the participants, and their culture are also part of the context. In pragmatics, context contributes a lot to discourse production and interpretation. It can be classified into co-text, situational context and cultural context. Text and context are two key elements in human communication, and from the viewpoint of semantics, they are complementary (Shen [2012]). In the linguistic communication, context determines text and text reflects it (Shaozeng [1995]).

Non-verbal communication. Non-verbal communication is the first form of communication in the lifespan of humans. Before humans evolved their ability to speak and use language, they were able to communicate using their visual body gestures - their non-verbal channels of communication (Knapp et al. [2013]). Human-Human Interaction (HHI) involves all non-verbal cues such as body, face, voice, appearance, touch, distancing, and other physical cues. Non-verbal behaviors convey tremendous information to the interlocutors. One important channel of communication in HHI is the human face. A variety of verbal, emotional, and conversational cues are displayed on the face while interacting. Humans use their gaze to convey their desire to switch speaking turns, and their hands movements to express their thoughts (Burgoon et al. [2021]). Non-verbal vocalizations, hand gestures, bodily movements, facial expressions, and gazes are emitted during speech, they are closely integrated with the speaker's speech, and may emphasize or disambiguate them. Non-verbal communication involves two main aspects: "*perception*" and "*production*". On the one hand, non-verbal communication is used by the interlocutors to "perceive" information about the speaker. On the other hand, it is used by the speaker to "produce" and convey his/her intention.

Visual and speech prosody. During speech, humans continually employ various gestures, known as "*visual prosody*", which is a form of facial (Graf et al. [2002]), head (Ding et al. [2013]), hands (Biau et al. [2016]) or body (Brentari et al. [2011]) movement generated in conjunction with *verbal communication*. These gestures involve different head movements, blinks, eyebrow gestures, gaze, frowning, nose wrinkling or lips moistening (Wang et al. [2018]). They are associated with prosody and para-linguistic information. *Speech prosody* refers to various speech characteristics like *intonation*, *rhythm*, and *stress*. Paralinguistic information refers to the cues, which can be used to convey emotion, such as *pitch*, *volume* and *intonation*. Facial gestures are consciously or unconsciously used to accentuate words, or mark speech pauses. Many facial expressions and head nods are tied to the speech's syntactic and prosodic structure. For instance, a stressed word is often accompanied by a head nod. A rising voice at the end of an utterance may be accompanied with an upper movement of the head, probably accompanied with a raise eyebrows. Hand gestures are also employed during speech, and they are connected to its semantic content, as well as its prosodic structure. For instance, a "*beat*" gesture is a type of hand gesturing that is performed by the movement of hands along with the rhythmical pulsation of speech. It is tightly synchronized with the prosodic contours of the speaker's discourse. Both speech and hand gestures arise from the same underlying cognitive process.

2.3 What is Prosody?

2.3.1 Prosody, the "music" of language.

The term "prosody" refers to all suprasegmental aspects of speech (Xu [2019]). It provides important information beyond an utterance's lexical meaning. It reflects *expressiveness* in both *speech* and *body gesturing*. It gives additional meaning to the spoken words, and keeps listeners engaged. Prosody involves highlighting the right words, using voice pitch, voice loudness, intonation, voice modulation, and voice timbre. The rise and fall of a speaker's voice can add meaning to an utterance. Vocal timbre is important for conveying different emotions in the content like sadness, happiness, anger, or excitement. Prosody also involves taking appropriate pauses. Knowing when to pause, and how long, is essential when speaking. More specifically, pauses can add anticipation, and allow the message to sink in. Prosody is considered as a parallel channel of communication, transferring messages that cannot be deduced from the lexical channel. When used with speech, it is known as the intonation, rhythm, or "music" of language, and is considered as an important aspect of all natural languages. Prosodic structures, similar to all other language characteristics, are generated not only by the vocal cords of spoken language users, but also by hand gestures, body gestures, eyebrows and face motion, heads, and bodies of sign language users (Shih and Kochanski [2002], Esteve-Gibert and Guellaï [2018], Munhall et al. [2004], Nespor and Sandler [1999], Dachkovsky and Sandler [2009], Wilbur [2000], Van der Kooij et al. [2006]). They carry additional semantic information which is used to create a strong and enduring memory representation. For instance, speakers spontaneously raise their voice to mark pitch accent that coincide with raised eyebrow.

2.3.2 Speech prosody.

Speech prosody gives evidences to several channels of linguistic and paralinguistic information. Linguistic functions like stress and tone are conveyed as local movement of pitch. Speech prosody involves multi-channel signals to convey lexical meaning (stress, accentual and tone languages), non-lexical information (intonation type; i.e: question V. declarative sentences), discourse functions (focus, prominence, discourse segments, etc.), and paralinguistic parameters (i.e. excitement expressed by high pitch and fast speed; sadness expressed by low pitch and slow speed.) Moreover, prosody is connected to the physical system. During the course of statement utterances, there is a tendency for pitch to decline (Cohen et al. [1973], Shih [2000]), as shown in figure 2.1.

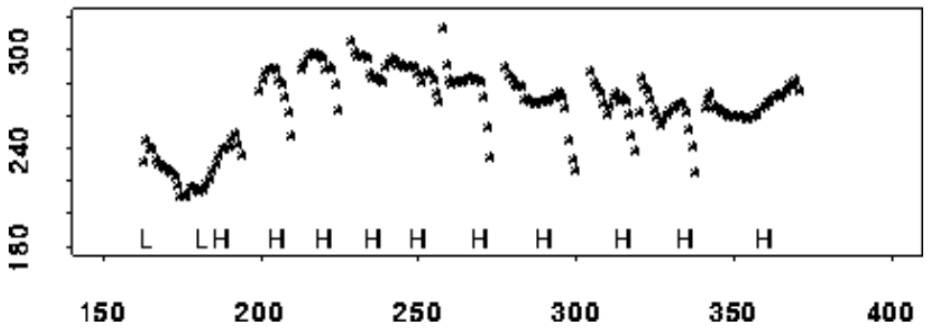


Figure 2.1 An example of a pitch declination in a sequence of high level tones, which are marked by "H".

2.3.3 Visual prosody.

Gestures are very connected to speech, to the point that people gesture even when they are alone (Corballis [1999]), and blind people often gesture when speaking (Iverson and Goldin-Meadow [1998]). People around the world produce spontaneous facial and hand gestures while speaking, known as "*visual prosody*". They are produced in conjunction with speech and include eyebrow motion, head nods, and some hand gestures (i.e beat gestures). Visual prosody is the visual aspect of speech and has parallels with the prosodic characteristic of speech. It aids expressiveness, and helps conveying additional information.

2.3.4 "*Multimodal*" expression of prosody.

As previously discussed, prosody is mainly composed of the speech and gestures dimensions. Both dimensions are deeply intertwined at the temporal, semantic, and pragmatic levels. Speakers' body motions are temporally aligned with the prosodic structure found in speech, pitch accents and boundary tones serving as anchoring points for important phases in body motions (Hadar et al. [1983], Ruiter [1998], Esteve-Gibert and Prieto [2013]). Speech prosody and gestures can both have a deictic element through which speakers emphasize some components in speech (Levelt et al. [1985], Roustan and Dohen [2010]), they can clarify syntactic components (Guellaï et al. [2014], Krivokapić et al. [2017]), and convey the speaker's emotions, beliefs, and attitudes (Ekman and Oster [1979], Kendon [2004], Esteve-Gibert et al. [2017]).

2.4 Multimodal Human Communication

Language, speech prosody and visual prosody are both involved as main channels of communication during human communication. In this section we dive deeper into the different components of human communication that involve language, speech and visual prosody, and which are considered main features in this thesis.

2.4.1 Prosodic Features

Vocalizations are sounds of different frequencies and intensities. When decoded, some of them can convey meaningful speech, while others can communicate emotions or interpersonal attitudes (Argyle [2013]). When we speak, we can vary the *vocalizations* in our voice. We can speak with a voice that can be *high* or *low* (pitch), *loud* or *soft* (loudness), and *fast* or *slow* (speech-rate). This variation is what we call *speech prosody*. It is represented by the *rhythm*, the *stress*, and *intonation* of speech, and it can be thought as adding *musicality* to speech. Speech prosody also carries many information such as the speaker's emotional state, or a certain emphasis. The acoustic prosodic cues emitted during speech are considered as part of language. For instance, speakers have a *rising pitch* when asking questions. They *pause* to show syntax, and use *loudness* to emphasize a point. Prosodic signals also convey emotional information. These prosodic signals are able to convey information about speakers. They can also alter the signification of the message. How a speaker speaks can reflect information about his or her personality, age, social class, and who they are. These various signals can be classified as shown in Figure 2.2. Employing accents or special *intonations* can add up extra information to the transmitted message, in many ways such as question intonation for statements. You can tell a lot about the meaning behind a speaker's words by evaluating his/her prosodic cues. The same utterance can hold a very different meaning in different contexts, and the prosodic features used will heavily influence this meaning.

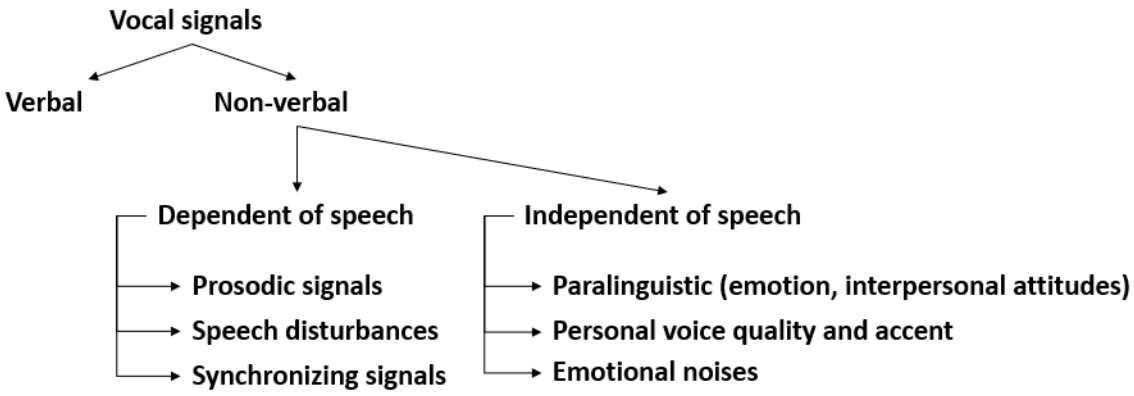


Figure 2.2 Vocal signals

Pitch. Pitch refers to the "highness" or "lowness" of the voice. The pitch of a speaker's voice reflects the f_0 at which his/her vocal folds are vibrating and thereby imposing periodic fluctuations in air pressure. It is measured in hertz, or cycles per second. The fundamental meaning of a *raised pitch* is probably emphasis, interest, and excitement. It

is most of the time accompanied by upward eyebrow movements as well as the movement of mouth, hands and shoulders. Moreover, pitch follows a specific pattern for every language (Mennen et al. [2007]), for different kind of utterances. For English language, patterns include a falling pitch, rising pitch and or employing both in the same utterance. Figure 2.1 depicts a falling pitch at the end of an utterance. For instance, questions starting with the word "*What*" or "*How*" are said with a falling pitch. The questions in which the verb and subject are inverted are said with a rising tone. Pitch patterns can negate the meaning of the spoken utterance, sarcastically, or when the word "*yes*" is said to indicate unwillingness and therefore a "*no*". Changes of pitch can also be used to emphasize certain words. The way the speaker can change his/her voice to convey meaning is referred to as *intonation*, which is manifested acoustically in f_0. The change in f_0, either *upward* or *downward*, is referred to as an *inflection*.

Stress. *Stress* refers to intensity or emphasis placed on a *syllable* or *word* in a spoken utterance, which makes the pronunciation *louder*.A same utterance can convey different meanings by stressing different words and directing attention to them Pierrehumbert [1990], as in "they are *playing* in the garden" or "they are playing *in the garden*" - words in italic are stressed.

Pauses. During speech, the string of words that are emitted by the speaker constitute an utterance (Argyle [2013]) which may be divided into shorter sequences of words separated by pauses. In fact, around half of a speech contains pauses. Pauses may be "*filled*" - with 'ums' or 'ers'; or "*unfilled*" - with silence. For instance, an Inter-Pausal Unit (IPU) is a sequence of continuous stretch of speech in one speaker's channel, delimited by a silence of more than 0.2 seconds. Pauses with a duration less than 0.2 seconds are used for the purpose of emphasis (Argyle [2013]). Longer ones are often used to signal grammatical junctures like the end of utterances. When a topic is difficult, pauses are twice as frequent, as argued by Goldman-Eisler (Goldman-Eisler [1972]). Speakers also employ more pauses and tend to slow down when emphasizing a point. On the contrary, speakers tend to speak faster when saying a subordinate utterance.

Paralinguistic characteristics of vocalization. Paralanguage is the non-lexical component of communication that accompanies speech. Through paralanguage, speakers communicate their emotional state, veracity, and sincerity. It is also known as *vocalics*, and it conveys *emotions* and *attitudes* to other people by the way in which *words* are expressed. There are various paralinguistic features of vocalization including pitch, loudness, tempo, resonance, timbre, syllabic duration, and rhythm. These characteristics overlap with prosodic signals. For instance, the shape of the *pitch contour* - the curve of the perceived pitch of the voice over time - can indicate the end of an utterance, or a paralinguistic signal for emotion.

2.4.2 Hand gestures

During speech, people display a number of bodily movements, particularly with their hands. Bull and Connelly (Bull and Connelly [1985]) found that vocal stress co-occur with movements of the head, hands, or other parts of the body. Numerous gestures are considered "*illustrators*" of the verbal content (Argyle [2013]), since they copy shapes, objects, movements or can have metaphorical meanings. *Synchrony* exists between words

and gestures (Lindenfeld [1971]). In fact, body movements have a hierarchical structure that is synchronized with various sizes of verbal units (Kendon [1972], Scheflen [1964]).

Gestures that exhibit images. Gestures display images that cannot always be expressed in speech (McNeill [1994]), as well as images that may look concealed to the speaker. Gesture and speech must cooperate for the purpose of communication and to express the speaker's meaning. With these kinds of gestures, people unintentionally exhibit their inner thoughts, memories, and their understanding events of the world. The speaker unwittingly render his thoughts visible with these gestures. Gestures belong to the inside world of mental images and thoughts. They are considered thoughts themselves (McNeill [1994]). They are complex, perplexedly interconnected and not in any way like photographs.

Types of hand gestures. Even though this thesis's main focus do not explicitly include modelling the different types of hand gestures, it is important to understand how they are correlated with speech. The major types of gestures are illustrated in figure 2.3.

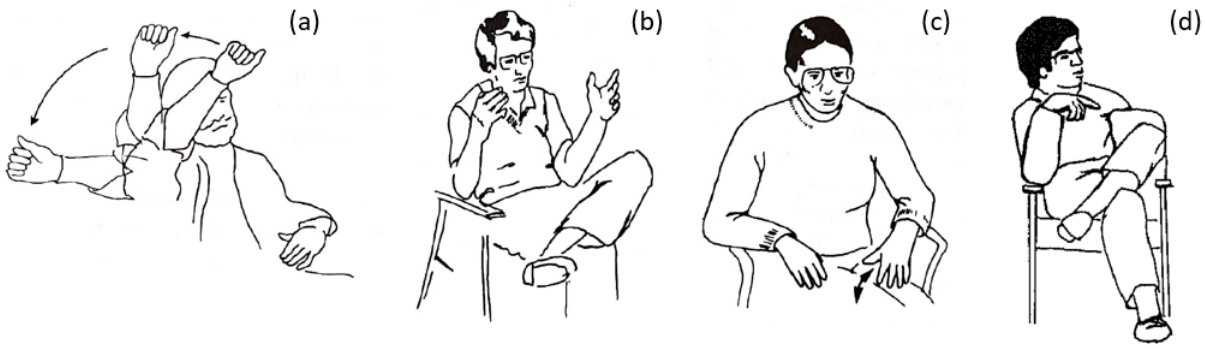


Figure 2.3 Illustration of (a) iconic gestures, (b) metaphoric gestures, (c) beat gestures, and (d) deictic gestures. Figures are taken from McNeill [1994]

- **Iconic gestures.** *Iconic* gestures are the gestures that exhibit semantic content correlated with speech. They reveal the speaker's memory image as well as the point of view that he or she takes towards it. Figure 2.3 (a) illustrates an iconic gesture of a speaker saying "*and he bends it way back*" while his hand appears to grip something and pull it from the upper front space back and down near to the shoulder. The "*stroke*" is the gesture movement, and it occurs with the part of the utterance that holds the same meaning. Figure 2.3 (a) depicts the close connection that exists between speech and gesture. Speech and gesture are complementary and partially overlapping. They jointly give a whole comprehensive perception into the speaker's thinking.

- **Metaphoric gestures.** "*Metaphoric*" gestures are also pictorial, but the pictorial content display an abstract idea, an image of the invisible, rather than a concrete object or event, like the case of *iconic* gestures. The gesture illustrates a concrete metaphor for the concept. Figure 2.3 (b) depicts a metaphoric gesture performed by a speaker saying "*it was a Sylvester and Tweety cartoon*", the idea that he was saying was supported by his hands by raising them up and offering the listener an "object". The speaker is not referring to a specific event, but to the genre of the cartoon, which

is an abstract concept. The metaphor is the concept of a genre of a certain topic presented as a physical object that is bounded and spatially localizable.

- **Beat gestures.** This type of gestures is called "*beat*" because they resemble beating musical time. Beat gestures are performed by moving the hands along with the rhythmical pulsation of speech, and they tend to have the same form regardless of the speech content, unlike metaphoric and iconic gestures (McNeill and Levy [1982]). An example of a beat gesture is quick and short movement of the flick of the hand up and down, or back and forth. This type of gesture has two movement phases: up and down, right and left, etc. ; unlike iconic and metaphoric gestures which typically have three movement phases - preparation, stroke, and retraction. The semiotic value of a beat gesture is that it indexes the word or utterance it follows as being significant, for its discourse-pragmatic content, and not its own semantic content. Figure 2.3 (c) depicts a beat gesture while the speaker is saying the statement "*whenever she*". The beat gesture marks the word "*whenever*", the linguistic segment that references the discourse as whole, and not a specific event.

- **Deictic gestures.** *Deictic* gestures are important for narrative, they are also named the "*pointing*" gestures. They are used to point at objects or events in the concrete or abstract world (Duncan et al. [2007]). Figure 2.3(d) is an illustration of an abstract pointing gesture, while the speaker is saying the statement "*where did you come from before*". The speaker is pointing at a abstract space between himself and the interlocutor, which represents a concept of where he had been before.

2.4.3 Facial gestures

The face is the most important non-verbal channel of communication for expressing emotions and attitudes (Ekman [1992], Argyle [2013]). During social interactions, facial expressions changes rapidly, and are decoded in terms of personality properties. People can make lot of different faces, especially emotional expressions such as happiness, sadness, fear and surprise. To describe a facial expression, Birdwhistell [1974] has proposed thirty-two *kinemes* - basic elements of expression in the face; while Ekman and Friesen (Ekman and Friesen [1982]) refer to 44 Action Units.

Facial Affect Scoring Technique (FAST). More methodical work by Ekman, Friesen, and Tomkins (Ekman et al. [1971]) led to the development of Facial Affect Scoring Technique (FAST). This technique consists of scoring three areas of the face independently, by comparing them against photographs. These areas contain 8 positions of the brows and forehead, 17 for eyes and lids, and 45 positions for the lower face. This scoring technique detects the action of the nervous system, and allows analysis of the effects of opposing muscles. Ekman and Friesen (Ekman and Friesen [1982]) developed ***FACS***, the most elaborate muscle scheme that is based on small facial movements knows by "*Action Units*". These facial movements are based on anatomical principles, they are due to single facial muscles visible to observers, and distinct from one another. FACS is also graded on a scale of intensity which gives a measure of how strong the activation of facial muscles is. Each AU describes one observable movement of a facial feature (e.g. eyebrows) by facial muscles. The FACS manual gives description of the 33 action units, 44 combinations of them, as well as information about other combination. Figure 2.4 illustrates an example of Action Units.

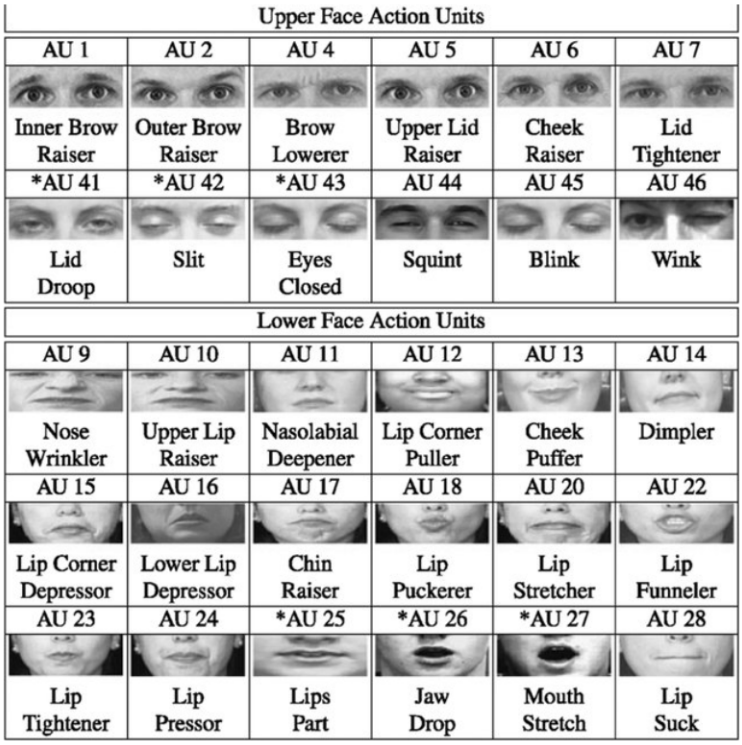


Figure 2.4 Illustration of Action Units. Figures are taken from Fac, HAGER [2002]

To express different emotions, a number of combinations of action units are used, producing different facial expressions and complex movements. For instance, to express surprise, action unit 1 and 2 are added together. Figure 2.5 illustrates some examples.

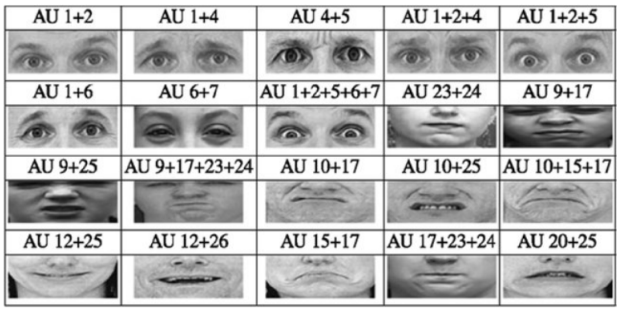


Figure 2.5 Illustration of combinations of action units to express different emotions. Figures are taken from Brahnam et al. [2007] and HAGER [2002]

Eyebrows. There are three action units for the eyebrow: (1) inner brow raised, (2) outer brow raised, (3) brow lowered. Eyebrow motion and speech have been shown to be strongly correlated by Ekman (Ekman [2004]). Eyebrow movement happen during thinking pauses (Cavé et al. [1996], Ekman [2004, 1992]), or to emphasize a word or sequence of words. Eyebrows are either raised or lowered when the speaker is thinking. Eyebrows movements are the most relevant and frequent facial gestures that are used during conversations (Chovil [1991]). Fundamental frequency f_0 variations and eyebrow movements are highly correlated during speech (Cavé et al. [1996]). Therefore, F0 and

eyebrow movements are not directly linked, but they are the results of linguistic and conversational choices. They are also used to reassure the speaker that the attention of his/her listener is still captivated. Eyebrows also mirror the listener's amount of understanding, and can be used as a backchannel (Cavé et al. [1996]).

2.4.4 The verbal message

Language is a system considered in the vocal channel of communication. The meaning of words in a speech can be altered by the different non-verbal signals emitted by the speaker's body. Non-verbal cues are used to "*frame*" the words (Argyle [2013]). A rising pitch indicates a question, other signals, can help a speaker indicate whether or not he / she expects an answer. Different vocal patterns are used with different facial signals (Crystal [1986]) when speakers say the same utterance in different styles (i.e. angry, amused, puzzled)

2.5 Multimodal Human Behavioral Style

Movements and gestures are person-specific and *idiosyncratic* in nature (McNeill et al. [2005]), and each speaker has his or her own gesturing *style*.

Human behavioral style. Humans do not equally gesture as they are speaking. *Behavioral style* involves the ways in which people talk differently. It can be common between different speakers, and unique to a speaker's prototypical gestures produced consciously and unconsciously. The speaker's personality and the social situations he/she is in have a large effect on his/her behavioral style. There is a broad *variability* in the degree to which speakers gesture as they talk (Hostetter and Potthoff [2012]). This *variability* is due to the speaker's personality traits, verbal skills (Hostetter and Alibali [2007]), age (Alibali et al. [2009], Feyereisen and Havard [1999]), or culture (Kita [2009]), etc. It also depends on the subject of the conversation, its role, the interlocutor, and the place in which the conversation takes place. A human's unique personality traits also affects this *variability*. The five factor personality model (McCrae and Costa Jr [1997]) proposes that personality can be best described as a human's unique fusion of five traits: extraversion, neuroticism, agreeableness, conscientiousness, and openness to experience. All humans possess the five traits to some degree, but each human can score relatively high or low on each of the traits. The five personality traits influence individuals including their speech-accompanying gesturing in terms of gesture type, frequency and expressivity (Hostetter and Potthoff [2012]). There is also the *situational variability*, which means that speakers gesture more in some situations than in others. *Behavioral style* could generalize to a group of individuals. An example of such case is that extrovert individuals may use a larger spatial gesturing than introverts (Hostetter and Potthoff [2012]). *Behavioral style* is also continuously attuned as it is co-produced with the audience (Mendoza-Denton [1999]). It can be very self-conscious and at the same time can be extremely routinized to the extent that it resists attempts of being altered (Mendoza-Denton [1999]).

Human behavioral idiosyncrasy. Gestures are not fixed, they are free, and reveal the idiosyncratic imagery of thoughts (McNeill et al. [2005]). They show inter-individual differences and result in different gesturing styles when delivered by different speakers even

if the discourse content is the same. The idiosyncratic nature of gestures defines the gesturing *style* of a speaker. Different speakers use different *style* of body movement, as well as different speaking *style*.

2.5.1 Multimodality of behavioral style.

Human behavior style is a socially meaningful clustering of features found within and across multiple modalities, specifically in linguistic (Campbell-Kibler et al. [2006]), spoken behavior such as the speaking style conveyed by speech prosody (Moon et al. [2022], Obin [2011]), and nonverbal behavior such as hand gestures and body postures (Obermeier et al. [2015], Wagner et al. [2014]). More specifically, a speaker *behavioral style* is defined by verbal, gestural, facial, prosodic and acoustic features. Therefore, we consider behavioral style as being *multimodal* (Knapp et al. [2013]), as it is found in both *verbal* and *non-verbal* modalities of communication. *Style* is omnipresent in speech; it colors the communicative behaviors expressivity.

Verbal Communication Style. Verbal communication reflects communicative styles. Verbal cues reflect different communication styles such as dominance, which is a style that is effective in influencing others, and that is reflected by requests, directives, and assertive statements (Dillard [2010]). In addition, certain verbal cues that are associated with power distinguish the speech styles between men and women (Lakoff and Lakoff [2004]). Examples of less powerful speech associated with females include tag questions, hedges and qualifiers, disclaimers and intensifiers(Lakoff and Lakoff [2004]).

Speech Style. In addition, prosodic features determined by biological, physiological, and sociocultural factors influence two main elements that determine speech style: primary qualities and voice qualifiers (Poyatos [1991]). *Primary qualities* are the prosodic characteristics that are always present in the human voice and include timbre, resonance, loudness, tempo, pitch, intonation range, syllabic duration, and rhythm. *Voice qualifiers* refer to how specific sounds are produced (i.e breathy or husky sounds). For instance, speaking in a moderately loud, rapid, expressive, and fluent voice is associated with dynamism, confidence, competence, and dominance styles (Apple et al. [1979], Buller and Aune [1992]).

Non-Verbal Communication Style. Style differs in the way different people talk in different situations which carry different social meanings (Bell [1984]). It is continuously attuned (Campbell-Kibler et al. [2006]) as it is accomplished and co-produced with the audience (Mendoza-Denton [1999]).

A speaker's nonverbal speech style reflects his or her *personal characteristics* as well as the characteristics of the other person in an interaction. A simple yet well-studied example of such a "target effect" is baby talk, which is a style people employ to talk to young children, that is high-pitch, singsong, slow, rhythmic, repetitive, and simplified (Grieser and Kuhl [1988], Snow and Ferguson [1977]). Even young children adapt their non-verbal speech style when talking to other babies or pets.

Communication Accommodation Theory (CAT)

According to the Communication Accommodation Theory (CAT) (Giles [2016]), people often adjust their communicative style when they are exposed to people. The communication style is accommodated to the communication style of others in a variety of nonverbal behaviors, including facial expressions, smiling, eye behavior, touch, posture, speech rate, pitch, and accent (Giles [1999]); in an effort to either converge or diverge.

Convergence occurs when the speaker adapts his or her style so that it becomes more similar to another person's or group's style. *Divergence* happens when the speaker accommodates his or her communication style to become less similar to the other person or group's style. People typically practice divergence when they dislike the other person or group and want to distance themselves from them or when they want to emphasize their identification within a particular group.

The key points of this Chapter:

Multimodal prosody

- Prosody is *multimodal*. Prosodic structures are produced not only vocally, but also by hand gestures, body motion, facial and head gestures.
- Prosody is considered as a parallel channel of communication, transferring messages that cannot be deduced from the lexical channel.
- Speech and visual prosody are deeply intertwined at the temporal, semantic, and pragmatic levels. For instance, eyebrow motion and f_0 variations are highly correlated.
- Speech and visual prosody can both have elements through which speakers emphasize some components in speech, they can clarify syntactic components, and affect the speaker's emotions, beliefs, and attitudes.

Multimodal Human Communication

- Human communication is explicitly aided by gesture, and could be hindered without it.
- Gestures play a crucial role to aid in the speaker's own process of conveying information.
- Gestures are not simply an addition to speech, but rather an independent expression of thought that reveals underlying mental states, beliefs and intentions of the speaker.

Multimodal Human Behavioral Style

- Behavior style encompasses verbal and non-verbal human behavior styles.
- Behavioral style is *multimodal*. It is found and determined by verbal, gestural, and prosodic features.

Chapter 3

Gesture Generation Approaches

Contents

In this thesis, we are interested in the generation of sequences of *facial* or *body* gestures based on multimodal input sequences of *speech audio* and *text semantics*. Before we dive into the details of the different generative models that were developed,it is necessary to go over the existing approaches for non verbal behavior synthesis. Note that we use indifferently the terms non verbal behavior or gesture to mean all visual modalities, namely facial expression, body movement, hand gesture, head movement and gaze. When referring only to motion of the arm and hand, we use the hand gesture. This Chapter discusses and analyzes the existing gesture generation approaches, their underlying principles, and their limitations. In the following sections, we discuss the different categories of gesture generation models, mainly *rule-based*, *statistical*, and *learning-based* models. We pay a special attention to the *learning based approaches*: we discuss the existing *text-driven* gesture synthesis models, the ones that are *speech-driven*, and finally the ones driven by *multimodal data - speech and text*.

3.1 Nonverbal Behavior Synthesis Approaches for Embodied Conversational Agents

Researchers in the field of *Socially Interactive Agents* (SIA) have been working on gesture generation in the past few years. Both physical and virtual embodied agents were considered in different communities such as *social robotics*, and *intelligent virtual agent* communities.

Discrete Vs. continuous behavior synthesis. Non-verbal behavior synthesis models generate behavior that can be either *discrete* or *continuous*:

1. Some hand gesture synthesis models predict *discrete classes of gesture shapes*. These classes are most often referred by researchers as *gesture lexemes*. In Conversational Analysis, the notion of gesture lexicon is very common, and a *lexeme* holds some generalized information about a gesture form. Different lexemes can imply different shapes and orientations of gestures. For example, if a gesture is labeled as the lexeme "raised index finger", it refers to a hand shape where all fingers are closed except the index finger, and the index finger having an upward orientation (pointing up) (Kipp et al. [2007]).

2. Other gesture synthesis models predict *continuous sequences of gestures*. For facial gesture synthesis, the generative models synthesize either continuous values of AU intensities varying between 0 and 5, or continuous 2D facial landmarks. For hand or body movements synthesis, the output could refer to sequences of 2D or 3D positions - known as *keypoints*. These are usually represented by sequences of poses for a *skeleton* - sometimes referred as *skeleton stick* - in a 2D or 3D space. A skeleton is composed of body joints, and each pose is represented by the *keypoints* of the skeleton joints, which are the *coordinates* of the joints.

Classes of gesture generation models. There are three main classes of gesture generation models:

1. **Rule-based.** Rule-based models are based on a set of rules and conditions. Most of the time, these rules refer to the correspondences existing between patterns produced by human communicative intentions and nonverbal behaviors.

2. **Statistical.** Statistical models generally map between the input features and output gestures based on the statistics in a given dataset of human gestures.

3. **Learning-based.** Learning-based models are the models whose parameters are learned from data. They are often referred to as *data-driven* models, since they are based on machine learning algorithms and are built from large amounts of human communication and behavior data.

These types of models are reviewed in this Chapter, with a special and main focus on learning-based models, since we only consider them in the scope of this thesis.

3.2 Rule-based Approaches

FACE. FACE is a rule-based text-to-expression system that was proposed by Pelachaud et al. [1996] to generate facial expressions, head and eye motion given annotated speech

and using rules. To generate the non-verbal behavior, the authors consider emotions, intonation and information structure. The output is a script in the FACS format and the animation is rendered with Jack software. The input emotions are first mapped into a facial expression. Then, predefined rules map the other input annotations to facial actions units (AUs). The produced AU intensities are dependent on the speech rate. The final output consists of a list of AUs for each phoneme and pause.

Behavior Expression Animation Toolkit (BEAT). BEAT is a rule-based approach that was proposed in an earlier study by Cassell et al. [2001]. BEAT was developed to schedule non-verbal behaviors from plain text. The nonverbal output includes hand gestures, head nods, and gaze. It can produce synchronized speech and gestures given the input text. The system takes an input string text, and predefined gestures are selected based on a linguistic and contextual analysis of the input text and using a set of rules obtained from previous research on human behavior. More specifically, the input text is processed in a pipeline of four modules: (1) language tagging, (2) behavior generation, (3) behavior selection, and (4) behavior scheduling. The input text is first transformed into a tree structure. Text is divided into clauses, and using heuristic rules, clauses are decomposed into themes and rhemes. Their input-to-output pipeline approach is summarized in Figure 3.1.

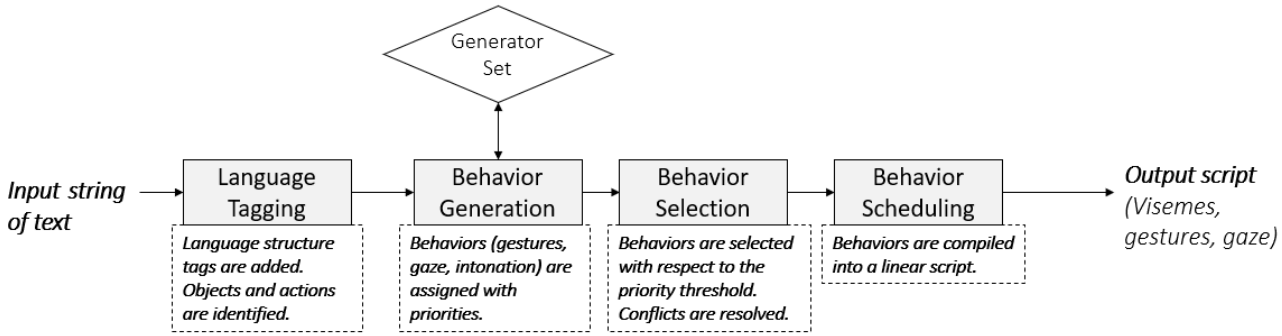


Figure 3.1 BEAT architecture.

Multimodal Assembly eXpert (MAX). Another approach was presented by Kopp et al. [2003] for the conversational agent MAX, which is illustrated in Figure 3.2.

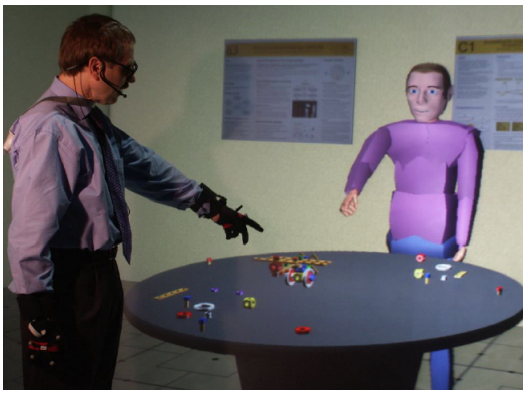


Figure 3.2 Multimodal Assembly eXpert" (MAX) agent interacting with the user. Figure taken from Kopp et al. [2003].

MAX is considered as a concept-to-gesture system. It takes as input annotated speech with communicative intents. The gesture synthesis is based on a lexicon of gestures, where each entry comprises of the gesture's communicative intent and a feature-based representation of the gesture, which represents the gesture spatial constraints. During gesture synthesis, the communicative intent of speech triggers all possible gestures of the gesture lexicon. The system then selects the gesture that matches best the current motion conditions of the agent. Their approach consists of creating gestures during run-time from XML-based specifications, which are descriptions of gestures form, as shown in figure 3.3.

```
<definition>
    ...
    <utterance>
        <specification>
                Insert <time id="t1"/> this screw <time id="t2" chunkborder="true"/> in <time id="t3"/> this bar. <time id="t4"/>
        </specification>
        <!-- refer to screw by gesture -->
        <behaviorspec id="gesture_0">
                <gesture>
                        <function name="refer_to_loc">
                                <argument name="refloc" value="$LocMovedObj"/>
                                <argument name="frame_of_reference" value="world"/>
                        </function>
                        <affiliate onset="t1" end="t2" focus=”diese”/>
                </gesture>
        </behaviorspec>
    ...
    </utterance>
</definition>
```

Figure 3.3 An XML specification from Kopp et al. [2003]

Greta: Gaze and Facial Expression Generation. Moving forward, researchers included affect in embodied conversational agents. Pelachaud et al. [2002] proposed GRETA, a 3D embodied conversational agent whose facial gestures depend on its emotional state. During a dialog with a user, the agent can manifest the affective states that are dynamically activated and de-activated in its mind (De Rosis et al. [2003]). Greta system is a concept-to-gesture system where the input representation encompasses the agent's emotion, beliefs and goals. Gesture synthesis is based on a lexicon of meaning-to-signal mappings. The output consists of text and MPEG-4 commands rendered by a virtual agent with a 3D muscle-based facial and body model 3.4.

Figure 3.4 GRETA embodied conversational agent.

Rule-based systems were incompatible with each other as different embodiment structures were used in each, and gestures were encoded differently. To solve this problem and unify the different frameworks, Kopp et al. [2006] proposed two languages to allow ECA researchers pool their resources to construct more sophisticated ECAs. The first is Function Markup Language (FML) which is used to describe intent. The second language is Behavior Markup Language (BML) which is used to describe desired physical behavior and has become the standard format for rule-based systems since then.

An example of a BML file is illustrated in Figure 3.5. It includes specifications related to the gestures and the accompanying speech.

```
<?xml version="1.0" encoding="utf-8"?>
<bml>
    <speech id="sp1" start="0.0" ready="0.1" stroke="0.1" relax="0.2" end="0.2">
        <text>
            <sync id="T0" time="0.05" />Quite
            <sync id="T1" time="0.38" />
            <sync id="T2" time="0.38" />well
            <sync id="T3" time="0.82" />
            <sync id="T4" time="0.97" />thank
            <sync id="T5" time="1.37" />
            <sync id="T6" time="1.37" />you
            <sync id="T7" time="1.61" />
        </text>

......

    <gesture id="p1_0" lexeme="ask_Ges_B" ready="0.03" relax="1.7">
        <description priority="1" type="gretabml">
            <reference>performative=ask_Ges_B</reference>
            <intensity>1.000</intensity>
            <SPC.value>0.000</SPC.value>
            <TMP.value>0.050</TMP.value>
            <FLD.value>-0.050</FLD.value>
            <PWR.value>0.100</PWR.value>
            <REP.value>-0.100</REP.value>
            <OPN.value>0.000</OPN.value>
            <TEN.value>0.000</TEN.value>
        </description>
    </gesture>
</bml>
```

Figure 3.5 A BML example from Pelachaud [2015]

More recently, rule-based systems were able to produce more complex gestures. Ravenet et al. [2018] proposed a model that can map speech to gestures to synthesize metaphoric gestures. Similarly, Marsella et al. [2013b] proposed an approach to produce speech-driven facial expressions and behaviors for a 3D embodied conversational agents. Their approach is rule-based, the given input speech text is analyzed semantically and contextually, to produce more realistic and appropriate gestures.

Rule-based systems are simple to use and intuitive, but require considerable human effort to determine the rules. They lack gesture diversity and variability, as they can only generate a limited set of gestures. In addition to that, the collection and implementation of rules necessitates a large amount of resources and is time-consuming.

3.3 Statistical Approaches

Statistical systems were proposed later on to overcome the limitations of rule-based systems. As previously mentioned, statistical systems are built based on the statistics in datasets of human non-verbal behavior.

A very early statistical system was proposed by Kipp [2005]. The author's approach produces conversational gestures for an animated agent given an input annotated text. The author's approach used TV show recordings as empirical data to extract gestural key parameters and generate individual gestures. Kipp [2001] used *ANVIL* annotation tool to transcribe gesture and speech of the empirical data, and developed a module that generates individual gesture profiles from the annotations with statistical methods. They then output a linear script in XML format describing the gestures which are created based on the computed gesture profiles and heuristic rules. This script serves to animate a virtual agent.

Neff et al. [2008] propose an approach for producing full-body gesture animation for given input text in the style of a specific performer. The authors have created an *Animation Lexicon* for every speaker. It contains information about the motion of each *gesture lexeme*. Moreover, their approach employs a tool-assisted annotation process which takes as input a video of a specific person whose gesturing style they wish to animate, then builds a statistical model of the person's gesturing style. The statistical model is called "Gesture Profile". Similar to Kipp [2005], their approach produces a gesture script generated from the created *gesture profiles*. The script defines a stream of continuous gestures synchronized with speech. It is given as input to an animation system, which improves the gesture description with additional details. Motion is then simulated and produced based on the description, and using the speaker's animation lexicon.

Another model called GNetIc was proposed later by Bergmann and Kopp [2009a]. It is a framework based on Bayesian decision network for guiding iconic gestures synthesis. More specifically, the network allows for speaker-specific gesture synthesis driven by iconicity and the overall discourse context. Their approach is based on annotated corpora and is supplemented with rule-based decision making. Therefore it is considered hybrid, as it combines rule-based and statistical techniques.

Despite the advantages statistical models provide, the diversity of the produced gestures is still limited. We discuss in the following sections the learning-based approaches, which are also known as data-driven approaches.

3.4 Data-driven Approaches

A great deal of research addresses the issue of gesture generation using data-driven or machine learning approaches. In these approaches, non-verbal behavioral gestures are viewed as the output of some abstract functions that can be produced by the analysis of recorded non-verbal behavioral gestures data. The power of machine-learning approaches is to provide a yardstick against which to compare the generated gestures. Data-driven approaches allow researchers to assess the quality of synthesized motions by evaluating

how much they deviate from the recorded data, which is considered as the *ground truth*. Data-driven approaches typically consist of generative models trained on large amount of recorded data. The learned models are then used to generate novel animations by executing some operations within their learned space.

There have been major efforts in developing data-driven gesture generation systems for embodied conversational agents. Some are driven by audio acoustic features, some others by text information, and few others by features coming from more than one modality, namely speech and text.

Data-driven gesture generation approaches can be grouped in different ways: (1) based on the input modalities such as speech, text, or both; (2) based on the output generated type of gesture such as head motion, eyebrow motion, body gestures, hand gestures, facial gestures, etc. In the following sections we review the different existing approaches, their underlying principles, and their main limitations. We group them according to the input data modalities type.

3.4.1 Speech-Driven Approaches

A large number of speech-driven facial, head, and body gesture generation approaches have been proposed. An early study by Cao et al. [2005] proposed a model for generating expressive facial movement synchronized with the audio of input utterances. The inputs to their system is a spoken utterance and a set of emotional tags. The emotional tags of the input audio can be either determined by the user or extracted from the speech signal using a Support Vector Machine (SVM) classifier. The output of the system is a facial animation synchronized with the input audio and which reflects the specified emotions. Their generative model also maintains accurate synchronization of lip movements. Their approach organizes a large set of recorded gestures and the corresponding speech audio into a data structure called the "Anime Graph".

Zoric et al. [2006] followed on and proposed a *facial* gesture generation model that produces lip movements based on input speech signal. In their work, virtual speakers can read given text and transform it into the corresponding speech and facial gesturing, with automatic Lip Sync process, which are determined from the speech signal. Their method combines lexical analysis of input text, with a statistical model that describes the frequencies and amplitudes of facial movements. Their model is created by conducting an analysis of a training data set that includes multiple speakers videotaped, as well as stenographs of their speech. The lexical analysis of the stenograph texts allowed to correlate the lexical features of input text with the corresponding facial movements.

Hidden Markov Models (HMM). Later on, Hofer and Shimodaira [2007] proposed a speech driven *head* motion sequence prediction based on *Hidden Markov Models*(HMM). Their modelling approach is based on the fact that head motion can be considered as a sequence of short homogeneous units that can be modelled separately. The model is trained on motion units and act as a sequence generator. The training data was collected and annotated by the authors. Their approach is able to distinguish different head motion patterns based on speech features with a 70% accuracy. Ding et al. [2013] proposed an animation model that is also based on HMM to capture the tight relationship between

speech and facial gestures, and then synthesize speech-driven virtual agents' facial gestures. Their model can be parameterized from training samples, is used to capture the mappings between audio and facial gestures.

HMMs have been one of the most popular human behavioral modeling techniques.More complex extensions of HMMs have been used by researchers to model complex activities such as the interaction between two people. These extensions include Parameterized-HMMs (Wilson and Bobick [1998]), Entropic-HMMs (Brand and Kettnaker [2000]), Variable-length HMMs (Galata et al. [2001]), Coupled-HMMs (Brand et al. [1997]), and structured HMMs (Hongeng et al. [2000]). These studies typically use data with short utterances / sequences, and do not know a large variability of expressiveness in both speech and the synthesized gestures. This is due to the constraints that were applied on the number of pre-defined gesture patterns, which limits the variation of expressiveness captured in the data.

Dynamic Bayesian networks (DBNs). Moving beyond the HMM representation and solution paradigm, researchers explored more general temporal dependency models, such as Dynamic Bayesian networks (DBNs), which are also called dynamic graphical models. DBNs have been adopted by several researchers for the modeling and recognition of human activities (Gong and Buxton [1995], Forbes et al. [1995], Liao et al. [2007]). DBNs present several advantages: (1) they can handle incomplete data, missing data, and uncertainty; (2) they are trainable and are capable of avoiding overfitting; (3) they are modular and parallelizable; and (4) they encode causality.

Mariooryad and Busso [2012] developed an approach based on DBNs for facial animation synthesis. Their framework is driven by speech and produces head and eyebrows motion. They propose three DBN models to incorporate different levels of dependencies between head and eyebrow motions. Later on, Sadoughi and Busso [2019] proposed a speech-driven system to predict hand and head motion, using a Dynamic Bayesian Network. Their model is constrained by contextual information and these constraints condition the state configuration between speech and gestures. Despite the many advantages DBNs offer, they pose hard inference problems, mostly with loopy graphs and continuous data. In fact, several efficient optimizations are available for training HMMs, but not for general DBNs.

Deep Neural Networks (DNN). Recently, Deep Neural Networks (DNN) has been applied to gesture generation problem, to overcome some of the limitations found in conventional HMM and DBN approaches. They can capture the large range of variations that are found in expressive data without the need to pre-define gestures patterns. Their hidden layers are capable of learning complex relationships between input and output features and appeared to be more effective than decision trees Ze et al. [2013]. DNNs are also less exposed to over-smoothing and conserve more detail in the output signal. In text-to-speech synthesis, DNNs have widely and successfully outperformed HMM systems Ze et al. [2013].

Seeing the advantages DNNs offer, Ding et al. [2015] were the first to use DNNs for speech-driven head gesture synthesis. They developed and pre-trained a deep belief network (DBN) with stacked restricted Boltzmann machines. On top of the output of the net-

work, they added a target layer for parameter fine-tuning. Haag and Shimodaira [2016] proposed a speech driven head motion synthesis approach based on DNNs. They use them with stacked bottleneck features, along with a Long Short-Term Memory (LSTM) network. In the work of Taylor et al. [2017], they generate lower facial movements based on a deep learning approach that uses a sliding window predictor that learns nonlinear mappings from phonemes to mouth motion. Later on, Suwajanakorn et al. [2017] proposed an approach based on LSTM for synthesizing a video of Obama's speech, they map original audio features to mouth shapes. The model could not perform well in generalizing other identities despite its good accuracy in lip synchronization. On the other hand, Karras et al. [2017] proposed a speech-driven real-time 3D facial animation model based on DNNs. Their network takes as input half a second of speech, and generates the 3D vertex positions of a fixed-topology mesh describing the facial pose at the center of the audio window. The network also takes a description of the emotional state corresponding to the speech as input. Emotional states are then learned from the training data without having been pre-labeled. More specifically, the network consists of a *convolutional* and *fully connected* layers. Their approach suffers from lack of variability in the synthesized gestures. This is due to the small size of the training dataset (3 - 5mins) that was used, in addition to the lack of fine detail in the performance capture data. Moreover, the text is not taken into account, limiting the range and variability of the produced gestures. LSTM networks driven by speech were also used to predict sequences of gestures (Hasegawa et al. [2018]) and body motions (Shlizerman et al. [2018], Ahuja et al. [2019]).

Later on, Kucherenko et al. [2019] proposed a speech driven gesture generation by learning a lower dimensional representation of human movements using a denoising autoencoder neural network, and an encoder network that maps speech to movement representation with a low dimensionality. Their model takes speech as input and generates gestures in the form of a sequence of 3D coordinates. The main shortcoming of their approach is the lack of appropriate dynamic range of the produced motion. Another limitation is that text is also not taken into account. On another hand, Song et al. [2018] suggested an audio-driven approach based on *conditional recurrent generation network*, which merges image and audio features into a recurring unit and produce facial animation by time-dependent coupling. The main limitation of this work is the weakness of their training performance, as well as the low resolution of the synthesized video. Later on, Ginosar et al. [2019] proposed an approach driven by speech to produce body gestures. They learn a mapping from speech to gesture using *L1* regression to temporal stacks of *2D* pose keypoints. They additionally use an adversarial discriminator to make sure that the generated motion is plausible w.r.t the typical motion of the speaker. The main limitation of this work is that it was trained on single speakers, and therefore cannot generalize to new speakers.

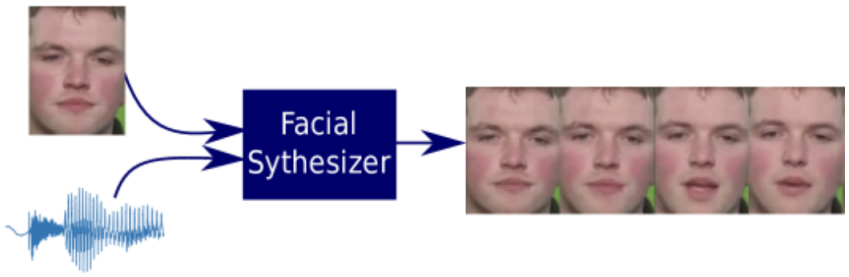


Figure 3.6 Facial synthesizer of Vougioukas et al. [2019] (Figure taken from Vougioukas et al. [2019]).

Vougioukas et al. [2019] propose an approach for synthesizing videos of talking heads based on a person's image, and audio speech (Figure 3.6). They generate lip motion that is in sync with speech, as well as facial expressions like blinks and eyebrow motion. Their approach is based on *Generative Adversarial Networks (GAN)* that uses three discriminators whose goal is to produce reasonable expressions, and audio-visual synchronization. Later on, Oh et al. [2019] proposed a speech-driven model trained on a large number of videos to reconstruct facial images from short audio recordings of the corresponding speaker. They design and train a DNN to perform this task using a large amount of Internet/YouTube videos of people speaking. Their model is self-supervised: during training, it learns the complex relationships between speech and facial gestures which allows the generation of images that capture various physical attributes of the speakers, mainly age, gender and ethnicity.

Other speech-driven approaches have been proposed by Jamaludin et al. [2019], Duarte et al. [2019], Garrido et al. [2015], Cudeiro et al. [2019] and Lu and Shimodaira [2020] for facial gestures synthesis and each approach has its limitations. Jamaludin et al. [2019] presented a model that integrates an auto-encoder to learn the correspondence between audio features and video data. Generated animation of their talking faces lack continuity. Duarte et al. [2019] proposed a method to synthesize facial videos, but the results are ambiguous. Garrido et al. [2015] also proposed a an approach that synthesizes the speaker's face by moving the mouth shape of the speaker in the dubbing video to the target video. Cudeiro et al. [2019] proposed VOCA (Voice Operated Character Animation), an approach for synthesizing speech-driven 3D facial animation. They train a DNN which takes any speech signal of any person, in any language as input and generates animations displaying a wide range of facial expressions. The model was trained on English language but can generalize to non-English sentences. The main limitation of this work is that the synthesized motions are mostly present in the lower face, and lacks realism especially for upper-face motions.

More recently, Lu and Shimodaira [2020] presented an approach that predicts head motion based on speech waveforms. They suggest a canonical-correlation-constrained autoencoder (CCCAE), in which hidden layers are trained to maximize the canonical correlation with head motion while simultaneously minimizing error. Ferstl et al. [2020] proposed an approach based on generative adversarial training in combination with a recurrent neural network to map speech to 3D gesture motion, and they used a gesture phase classifier as an additional adversarial loss. However, the produced motion lacks

realism. Jonell et al. [2020] propose a probabilistic approach based on normalizing flows for synthesizing facial gestures in dyadic settings, and based on multimodal inputs. The main weakness of this approach is that authors evaluated the produced gestures without revealing the audio to the evaluators, and this limitation most likely may have affected the evaluators' responses.

The aforementioned works have focused on producing nonverbal behaviors (facial expression, head movement, gestures in particular) driven namely by speech. In addition to the main limitations discussed in these audio-driven approaches, they do not generate complex gestures. More specifically, gestures lack diversity and variability and this is due to the absence of text modality and semantics, limiting the range of the produced gestures. Researchers recently turned to investigate the relationship existing between text transcriptions and speech, since there exist relationships between speech, gestures and text, and it could enable the synthesis of more diverse and complex gestures. We discuss the main works related to text-driven approaches in the next section.

3.4.2 Text-Driven Approaches

Text-driven approaches were also recently proposed for the task of hand and body gesture generation. Ishi et al. [2018] proposed an approach to synthesize gestures from text input through a series of probabilistic function where words were mapped to word concepts using WordNet (Fellbaum [2010]). The word concepts were similarly mapped to a gesture function, which maps them to clusters of 3D hand gestures. On the other hand, Yoon et al. [2019] proposed a sequence to sequence RNN based network to map utterance text to gestures. More specifically, they used *Gated Recurrent Units (GRU)*, a type of RNNs with gating mechanisms. The synthesized gestures were simulated on a NAO humanoid robot.

Even though text-driven gesture synthesis approaches may learn important relationships existing between text and gestures, they fail to capture the strong relationship between gestures and speech (Feyereisen et al. [1991]). This is why approaches that are semantically-aware and speech-driven combine both text and speech. We discuss such existing approaches in the next section.

3.4.3 Text and Speech Driven Approaches

Chiu et al. [2015] proposed an approach that combines audio signals and text transcripts for synthesizing gestures. Their model is a combination of a feed-forward neural network and Conditional Random Fields (CRFs). The main limitation of this work is that they map text and audio to discrete predefined gestures, therefore the variability of gesturing remains limited to these pre-defined gestures.

Moving forward, Kucherenko et al. [2020] proposed a speech and text driven gesture generation that maps speech acoustic and semantic gestures into continuous 3D body gestures. Their approach generates different gesture types including the acoustically-linked gestures and the ones that are semantically-aware. They separately encode audio features and text for each frame, then concatenate them before feeding them to multiple fully-connected layers. The output pose is given as input to the model in an autoregressive way. The main weakness of their approach is that it is deterministic and not stochastic, therefore

producing identical gesturing for a given input and therefore does not allow variability in gestures. Yoon et al. [2020] presented an automatic gesture generation model that uses the multimodal context of speech text, audio, and speaker identity "ID" to generate gestures. Ahuja et al. [2020a] studied the links between spoken language and co-speech gestures. They propose "Adversarial Importance Sampled Learning" (AISLe) which combines adversarial learning with importance sampling. They introduce the usage of transformers for gesture synthesis conditioned on speech, based on neural cross-attention architecture, which helps with the alignment between language and gestures. The attention mechanism is applied on audio and text to combine them. They used Generative Adversarial Networks to synthesize 2D pose gestures. These works (Kucherenko et al. [2020], Yoon et al. [2020] and Ahuja et al. [2020a]) were the first to combine both audio and text, and synthesize continuous gestures in contrast to Chiu et al. [2015] who worked with discrete classes of gestures.

3.5 Conclusion

Data-driven approaches have yielded some of the most expressive and realistic gesture synthesis models to date. However, most of these works have focused on the synthesizing gestures using one modality of human communication; in most of the cases it is speech audio, and in few works it is text transcriptions. In reality, *facial*, *hand* and *body* motion are not only dependent on speech, or only on text, but they are a function of both. For instance, a same utterance spoken in a joyful way would result in different gesturing if spoken in a sad way. On the other hand, a speaker would gesture differently if saying two different utterances in a happy way. Text includes semantic information that impacts the type of gestures a speaker displays. This association between speech and text has long been studied by researchers. Kucherenko et al. [2020], Yoon et al. [2020] and Ahuja et al. [2020a] are amongst the first to fuse speech and text for the task of continuous gesture synthesis. However, they have focused on synthesized *body* and *hand* gesturing, and they train their model on only *one speaker data*. In the next part (Chapter 5), we present our work for synthesizing speech-driven and semantically-aware *facial* gestures.

Part II

Multimodal Gesture Synthesis

Chapter 4

Corpora

Contents

In this Chapter, we present two corpora that we have used in the context of this thesis. We first present *TEDx Corpus*, a corpus that we collected to train, test, and validate our models that generate speech-driven and semantically-aware facial gestures (research questions 1, 2). More specifically, *TEDx Corpus* aims to study relationships governing three modalities: *speech audio*, *text semantics*, and *facial cues* - more precisely *eyebrow gestures* and *head motion*. The second corpus we introduce in this Chapter is *PATS corpus*, a corpus that was previously built by Ahuja et al. [2020b], and that is used in this thesis to train, test, and validate our models that are related to *body visual prosody expressivity* (research questions 1, 2) and *multimodal style modelling and transfer* (research questions 3, 4, 5). *PATS* corpus was originally proposed to study the correlations between multimodal features related to *speech audio*, *text semantics*, and *body pose gestures*. We extend *PATS* corpus to include additional multimodal features related to *2D facial landmarks*, and *dialog tags* - tags of the dialog acts of the spoken utterance.

4.1 TEDx Corpus

We introduce *TEDx Corpus*, the first corpus that was collected in the context of this thesis, for the purpose of studying the correlations between *speech audio*, *text semantics*, *eyebrow gestures* and *head motion*. This corpus aims to help us develop generative models that can produce expressive facial gestures for ECAs.

4.1.1 TEDx Talks

Figure 4.1 TEDx Talks

TED (Technology, Entertainment, Design)(ted) is an organization devoted to spreading ideas during TED events, generally in the form of short, powerful talks. They are often called "TED talks" or "TED conferences".

TEDx talks. During TED conferences, speakers share their major research or ideas from multiple disciplines with their audience (Figure 4.1). These talks contain myriad of presentation topics, each presented by a unique speaker.

Quality of TEDx talks. TEDx talks are quality talks, as they are well-structured, and well-recorded. TEDx speakers are coached and well-prepared in advance. They are taught on how to intrigue, inspire and put forward their "ideas worth sharing". They get expert coaching on how to deliver the best possible presentation, while employing expressive gesturing, vocalising, pace, and tone.

Speakers expressivity. Each speaker has his/her communicative style, and all of them have the same goal which is to captivate the audience. Speakers' speech and gestures are highly expressive, intense, and energized. Speakers employ expressive non-verbal cues for the purpose of delivering their presentation in the best possible way. Visual prosody cues such as facial expressions, hand/arm gestures, and body postures are strongly present throughout their speech.

4.1.2 TEDx Corpus

Given that TED speakers are competent communicators and use expressive verbal and non-verbal signals for conveying their ideas, it is very advantageous to use TEDx videos for modelling the relationships between facial visual prosody, speech prosody, language,

and speech context.

To answer our research questions 1, and 2 related to facial visual prosody expressivity, we gathered the *TEDx Corpus* that consists of multimodal speech and facial features extracted from TEDx talks.

TEDx talks were obtained from YouTube. We collected the same 1760 talks that were used in Yoon et al. [2019], along with their transcripts. After extracting the multimodal features, we processed the resulting data and cleaned them before using them in our studies. The collected 1760 videos have average length 13 minutes (minimum length is 1 min, and maximum length is 47 mins), with a frame rate equal to 24 FPS.

4.1.3 Features

The *TEDx Corpus* was collected to study the relationships between the main multimodal features that are involved in a human communication, for the purpose of developing models capable of synthesizing human-like and expressive facial gestures. More specifically, this corpus includes multimodal features related to: (1) *speech audio* - f_0, jitter, shimmer, Harmonic-to-Noise Ratio, and Hammarberg index, (2) *text semantics* - *BERT embeddings* ; (3) *eyebrow gestures* - $AU1$, $AU2$, $AU4$, $AU5$, $AU6$ and $AU7$; and (4) *head motion* - R_X, R_Y and R_Z. These features are listed in Table 4.1 and discussed in the following sections.

Features	***Collection Methodology***	***Available Representations***
Audio	SWIPE estimator	Fundamental Frequency - f_0
	OpenSmile	Jitter - $Jitt$
		Shimmer - $Shimm$
		Harmonic-to-Noise Ratio - HNR
		Hammarberg index - $Hamm$
Action Units (AUs)	OpenFace	$AU1$ - *Inner Brow Raiser*
		$AU2$ - *Outer Brow Raiser*
		$AU4$ - *Brow Lowerer*
		$AU5$ - *Upper Lid Raiser*
		$AU6$ - *Cheek Raiser*
		$AU7$ - *Lid Tightener*
Head Motion	OpenFace	Roll Euler angle - R_X
		Pitch Euler angle - R_Y
		Yaw Euler angle - R_Z
Text	Pre-trained "bert_base_uncased" model	$BERT\ embeddings$

Table 4.1 *TEDx Corpus* Features

Audio Features

The audio features we are considering in this corpus are prosodic and voice quality features, which are characteristics of voice expressivity (Monzo et al. [2014]).

More specifically, we consider f_0 variations, Jitter ($Jitt$), Shimmer ($Shimm$), Harmonic-to-Noise Ratio (HNR), and the Hammarberg index ($Hamm$). F0 variations capture pitch changes, which are essential for conveying intonation and melodic contour. Jitter and Shimmer provide insights into vocal fold stability, contributing to speech rhythm and fluency. HNR assesses the presence of unwanted noise, affecting prosodic cues. The Hammarberg Index helps identify speech pathology, which can impact prosody.

f_0 variations were extracted using SWIPE estimator (Camacho and Harris [2008]). The remaining voice quality features were extracted with OpenSmile (Eyben et al. [2013]). In this corpus, f_0 values were restricted to the range of 50 to 550Hz, which is enough to enclose the vocal ranges of both male and female speakers. In fact, the vocal speech of a typical adult male has a f_0 ranging from 85 to 180 Hz. That of a typical adult female ranges from 165 to 255 Hz (Baken and Orlikoff [2000], Titze [1994]).

Audio features alignment. IrcamAlign (Lanchantin et al. [2008]) was used to perform alignment for speech signals into phones and diphones, providing a confidence level for each phone. It was also used to extract the phonological structure such as syllables, words and breath sequences from the resulting aligned sequences of phones.

Action Units Features

As discussed in Chapter 2, facial movements are represented by "*Action Units*"(AUs), as defined in the Facial Action Coding Systems (FACS) manual that was developed by Ekman et al. [1971]. The AUs we study in the scope and context of this thesis are the ones related to *eyebrows* and *eyelids* movements. We do not consider the other facial movements involved in articulatory movements, since we do not model them in the scope of this thesis. *Eyebrows* and *eyelids* motion are represented by the six action units $AU1$, $AU2$, $AU4$, $AU5$, $AU6$, and $AU7$. The latter six upper-face AUs are listed and described in Table 4.1. We extracted eyebrows and eyelids action units using the tool OpenFace (Figure 4.2, Baltrušaitis et al. [2016]).

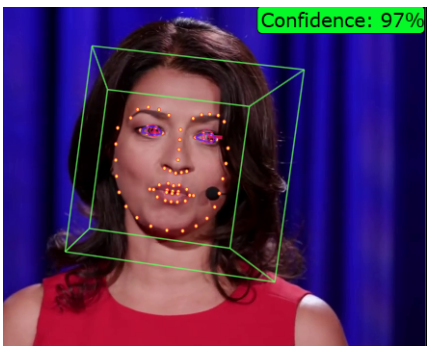


Figure 4.2 Open Face AU detection

AUs extraction. As previously discussed in Chapter 2, Action Units are represented by values of intensity, which is a measure of how strong the activation of facial muscles is. In OpenFace, AU intensities are continuous values ranging from 0 - *lowest* intensity - to 5 - *highest* intensity. Each generated AU intensity is given a "Success" score, that is equal to 1 in case OpenFace was able to detect the speaker's face, or 0 otherwise. Similarly, each intensity is given a "Confidence" value, which is a value between 0 and 1, representing the confidence level of OpenFace.

Head motion features

Head motion is represented by 3D head angles. Head rotations have three degrees of freedom, represented by the Euler angles: *roll*, *pitch* and *yaw*. The latter angles are represented by R_X, R_Y and R_Z, which are the rotations of the head with respect to X, Y, and Z axes.

Head motion extraction. R_X, R_Y and R_Z were also extracted using the tool OpenFace. Each head angle value has a success score and confidence level.

Text features

Transcripts of all collected TEDx talks were collected when downloading TEDx talks from Youtube. They include the timestamps of the *start* and *end* of each word in all utterances. In this corpus, speech text is represented by a sequence of words, and each word is encoded as a *BERT embedding* (Devlin et al. [2018], Wolf et al. [2019]). *BERT embeddings* were generated by the pre-trained "*bert_base_uncased*" model. We chose BERT as it is the first deeply bidirectional unsupervised language representation, that was trained using a large corpus of sentences, and produces powerful representation of words: its embeddings are jointly conditioned on both left and right contexts simultaneously, unlike other tools (Pennington et al. [2014], Peters et al. [2018]).

4.1.4 Data Cleaning

The main purpose of the *TEDx Corpus* is to provide clean, structured, and aligned multimodal features of the TEDx speakers. As we started extracting the multimodal features from the collected TEDx talks, we identified several situations where the speakers' extracted features were either missing, or very noisy. We list the major undesired situations:

- During the recording of TEDx talks, multiple cameras from various angles are used, which could sometimes display a variety of shots that include the audience faces, still pictures such as slides, or the back of the speaker. These shots are unnecessary for us and ought to be deleted.
- The speakers face in the videos are very small, and the extracted action units are noisy (see Figure 4.3).
- Faces of people displayed in the slides of the presenters may be extracted by OpenFace, as shown in Figure 4.4.
- Faces of people in the audience are also detected by OpenFace with a high *confidence level*, as illustrated in Figure 4.5.

Figure 4.3 The speaker's face displayed on screens extracted with OpenFace.

Figure 4.4 Faces extracted in slides of the speaker's presentation by OpenFace.

Figure 4.5 OpenFace detecting faces of people in the audience.

We worked out a series of solutions to overcome these issues, which are explained in the following sections.

4.1.5 Videos segmentation and shots filtering

We are only interested by the shots where only the speaker's face is detected with a large confidence level (confidence > 90), and where the face is visible enough and close to the camera.

For this purpose, we segmented the collected TEDx videos into shots using the tool PySceneDetect (Castellano), a python library that analyzes videos, looking for scene changes or cuts. Videos were split into individual shots that correspond to different scenes. To avoid having noisy features, and to minimize errors in our future studies, TEDx shots were filtered and selected such that:

- The speaker's face is visible to the camera
- The speaker's face is not far from the camera (not small)
- There is one face is in the shot - speaker's face, not the audience faces.
- The speaker is facing the camera that is in front of her/him
- There are no still pictures or slides in the shot

We applied those rules to the 1760 TEDx videos. We obtained 266,000 segments.

4.1.6 Data Processing

Our next step consists in extracting the multimodal features that we're interested in from these shots. The intensity values extracted by OpenFace are noisy and may be missing for some frames. We applied additional data smoothing and fitting techniques to further eliminate noise within these shots, and filling in the gaps by determining the missing values.

Median Filtering. *Median filtering* is a smoothing technique that is frequently used to remove noise from an image or a signal. It is especially effective at removing noise in smooth regions, while preserving the edges. For each extracted AU intensity of each shot of interest, we applied a *median filter* to remove noises and deal with the cases where OpenFace's confidence level is low. Figure 4.6 depicts the effect of applying window sizes equal to 3, 5 and 7 to $AU1$ for the purpose of filtering out noises while preserving and highlighting edges. After testing with different window sizes, we applied the median filter with a *window size* equals to 7, since it eliminates noise and maintains the edges, as shown in Figure 4.6.

Linear interpolation. Median filtering removed unwanted random noise generated by OpenFace. However, even after filtering unwanted shots and applying median filtering, we identified some cases where OpenFace did not detect well the speaker's face, and for these cases, *Success* score was equal to 0 and no intensity values were extracted. For this reason, we further applied linear interpolation, to fill in the gaps where OpenFace failed to detect the speaker's face, as shown in Figure 4.7.

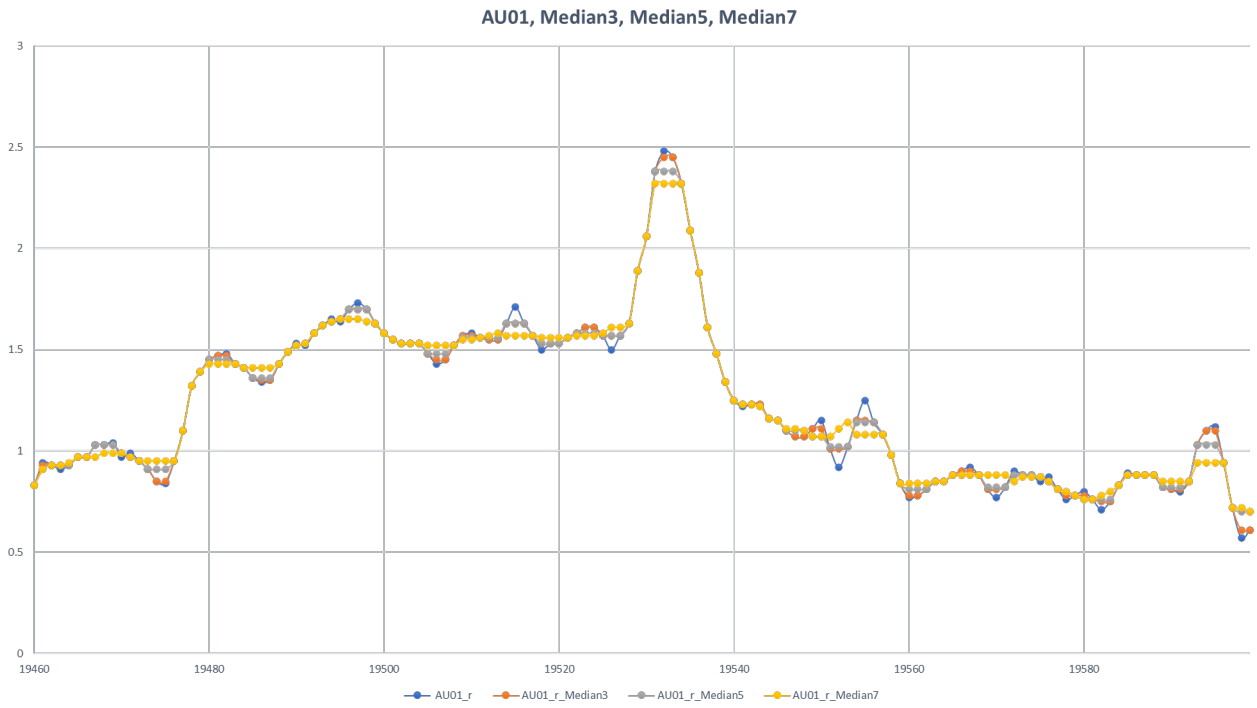


Figure 4.6 This figure is a plot of different median filtering window sizes applied to $AU1$ signal. The original signal is plotted in blue. The orange curve represents $AU1$ after applying median filtering with window size equal to 3. Median filtering with window size equal to 5 is plotted in grey. Median filtering with window size equal to 7 applied to $AU1$ is plotted in yellow.

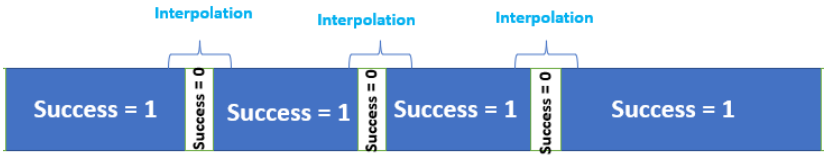


Figure 4.7 Linear Interpolation is applied on the frames where OpenFace's *success score* is equal to 0.

Similar to Action Units features, and for the purpose of removing noise and dealing with the cases where OpenFace's *confidence* is low, a median filter of window size equal to 7 was also applied on head angles values. In addition, we applied linear interpolation to deal with the frames where OpenFace failed to detect the speaker's face (*success score* equal to 0).

Audio Processing

"Voiced" and "Unvoiced" speech segments. Speech is composed of *phonemes*, which are generated by the vocal cords and the vocal tract. During the production of speech, the air coming out of lungs through the trachea is disrupted periodically by the vibrating vocal folds. Voiced signals are generated as a result of the vocal cords vibration, which is produced when the speaker pronounces a phoneme. Grossly, when we look at the speech signal waveform, if it looks nearly periodic in nature, then it can be marked as *voiced speech*. *Unvoiced signals*, by contrast, do not entail the use of the vocal cords, they are generated by the lips or the glottis constrictions.

Silence regions. The speech production is a succession of *voiced* and *unvoiced* speech, separated by *silence* regions. *Silence* exists when there is no excitation supplied to the vocal tract and thus no speech output is produced. As a matter of fact, silence is an integral part of speech signal. Without its presence between *voiced* and *unvoiced* speech segments, the speech will not be comprehensible. For instance, Figure 4.8 illustrated the f_0 contours generated when a speaker is saying the word "tired".

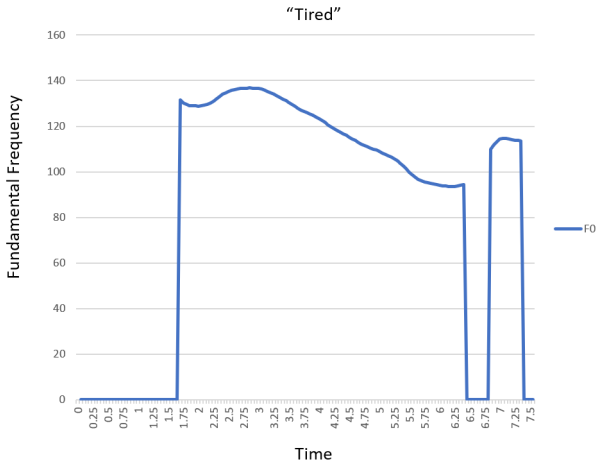


Figure 4.8 f_0 variations resulting from a speaker saying the word "Tired".

To overcome this lack of f_0 values we applied the following process.

Fundamental frequency processing. For each sequence of f_0 values corresponding to a word, we applied linear interpolation and extrapolation in order to get a complete sequence of non-zero f_0 values, as depicted in Figures 4.10, 4.11, and 4.9. In this corpus, f_0 values were restricted to the range of 50 to 550Hz, which is enough to enclose the vocal ranges of both male and female speakers. In fact, the vocal speech of a typical adult male has a f_0 ranging from 85 to 180 Hz. That of a typical adult female ranges from 165 to 255 Hz (Baken and Orlikoff [2000], Titze [1994]).

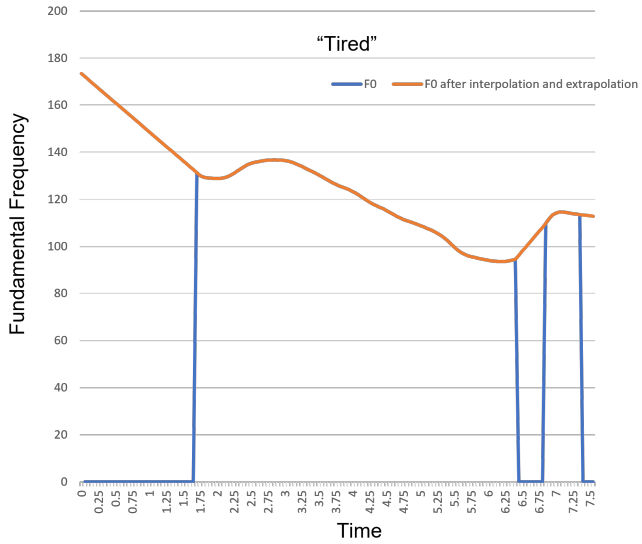


Figure 4.9 f_0 variations resulting from a speaker saying the word "Tired" Vs. f_0 contours after applying linear interpolation and extrapolation between *voiced speech* and *silence regions*.

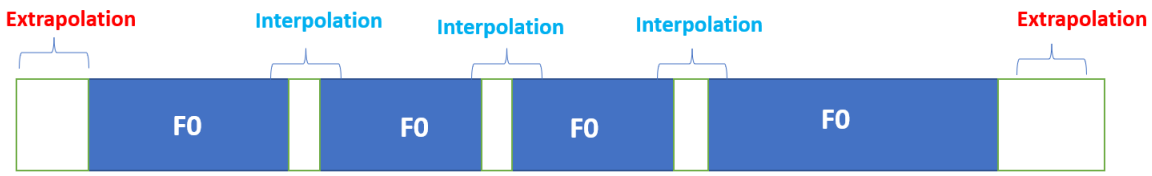


Figure 4.10 Linear interpolation and extrapolation are applied on f_0 unvoiced segments. Voiced segments are illustrated in blue, and unvoiced segments are illustrated in white

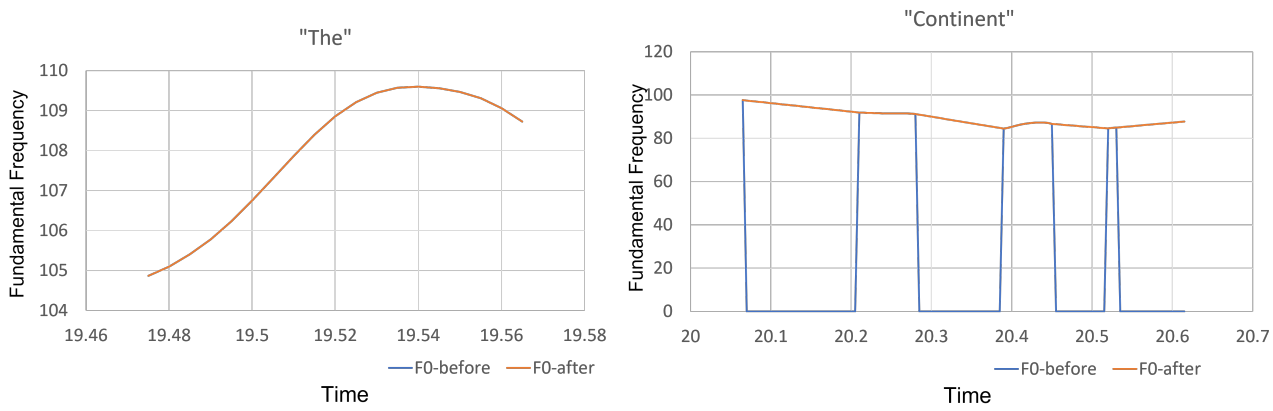


Figure 4.11 An example of word-level F0 contours corresponding to the utterance "The continent", before and after applying linear interpolation and extrapolation.

4.2 PATS Corpus

4.2.1 PATS (Pose, Audio, Transcript, Style)

PATS (Ahuja et al. [2020a, 2022]) is a corpus that was proposed by Ahuja et al. [2020b] to study the correlation of gestures with audio and text modalities. The corpus contains different and large amount of aligned 2D upper-body pose, speech and text transcripts. *PATS* was used in the context of this thesis to challenge the research questions related to *visual prosody expressivity* and *multimodal style modelling and transfer* (research questions Q3, Q4, and Q5).

4.2.2 PATS Data

The *PATS Corpus* consists of transcribed Pose data with aligned audio and texts. It includes data of 25 speakers with different communicative styles. Ginosar et al. [2019] collected 10 of these speakers. PATS contains 251 hours of data, with $84,000$ intervals and a mean duration equal to 10.7 seconds per interval. The standard deviation is 13.5 seconds per interval. An interval corresponds to an utterance consisting of 64 timesteps.

4.2.3 PATS Features

PATS Corpus consists of a large amount of multimodal features related to: (1) Speech, (2) Pose and (3) Transcripts.

Speech. Speech audio was collected from YouTube, and is provided by *Log-mel Spectrograms* representations.

Pose. Pose are represented using upper-body 2D skeletal keypoints, which were extracted using the tool OpenPose (Cao et al. [2017]). The following Figure 4.12 depicts the main 2D poses, relative to the upper-body joints of PATS speakers.

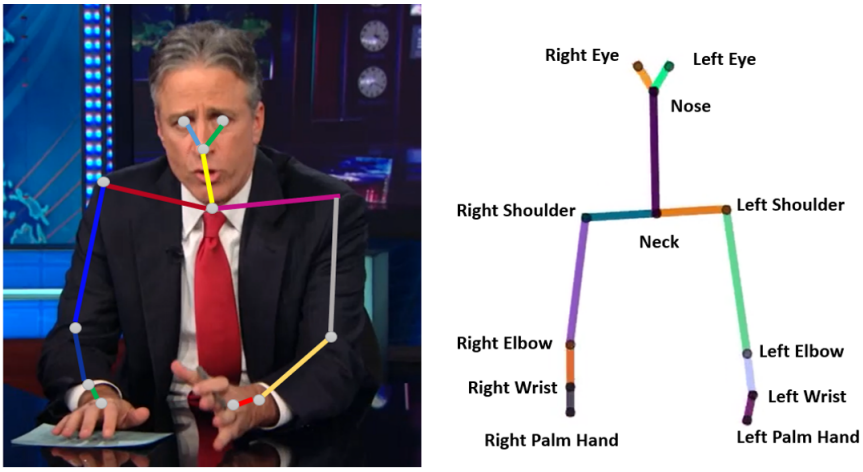


Figure 4.12 Figure (b) illustrates the two-dimensional (2D) skeleton of joints relative to the speaker in Figure (a).

Transcripts. Transcripts are represented by *Word Tokens*, *Bert Embeddings*, and *Word2Vec Embeddings*. The transcription of *Word Tokens* was done using *Google ASR* (goo). *Bert Embeddings* were generated using "bert_base_uncased" pretrained model provided by HuggingFace (Devlin et al. [2019]), and *Word2Vec Embeddings* (wor).

4.2.4 PATS Speakers

PATS consists of 25 speakers whose types of speech can be categorized as follows: 15 talk show hosts, 5 lecturers, 3 YouTubers, and 2 televangelists. Each speaker has his/her own communicative style, and lexical and gesture diversity.

Lexical Diversity Vs. Spatial Extent

A speaker *lexical diversity* reflects his/her lexical "richness", and refers to the diversity of the words used by him/her. It is computed by calculating the ratio of different unique word with respect to the total number of words (tokens). As shown in Figure 4.13 (published by Ahuja et al. [2020b]), the PATS speakers have diverse lexical content and diverse gestures. Each speaker has a specific position in the 2D space of lexical diversity Vs. average gesture spatial extent. As shown in Figure 4.13, speakers in the same domain are part of a same cluster, they share similar lexical diversity and spatial extent. We can notice that TV hosts are more expressive with their gestures and have a richer vocabulary than the speakers in other categories.

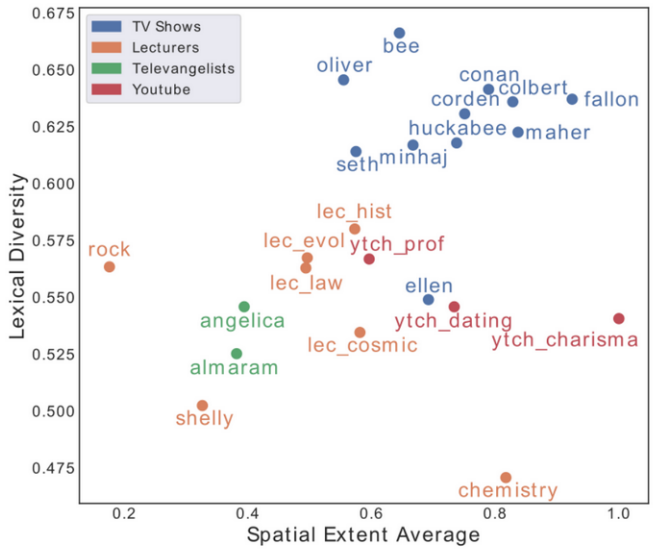


Figure 4.13 Lexical Diversity Vs. Spatial Extent

Lexical Similarity

To identify the common lexical vocabulary between the different speakers, we computed the *lexical similarity* of all speakers, by clustering the provided *Bert embeddings* using *K-means* clustering algorithm, with k=3. Results are illustrated in table 4.2.

Lexical Similarity - Kmeans Clustering of Bert Embeddings, k=3		
Cluster 1	***Cluster 2***	***Cluster 3***
angelica	lec_evol	colbert
almaram	chemistry	bee
ytch_prof	lec_hist	corden
ytch_dating	lec_law	fallon
ytch_charisma	lec_cosmic	jon
		minhaj
		oliver
		seth
		shelly
		conan
		ellen
		huckabee
		noah
		rock
		shelly

Table 4.2 Lexical Similarity between PATS speakers divided into 3 clusters.

We can notice that most of the speakers belonging to a same category belong to a same cluster. Cluster 3 includes all *TV hosts*. Cluster 1 encompasses all *YouTubers* and *televangelists*. We note that Rock and Shelly lecturers are included in Cluster 3, which means that their lexical vocabulary is close to the one of TV hosts. Cluster 2 gathers the other *lecturers*.

Acceleration, Jerk, Velocity

To further understand each speaker's communicative style, we looked at their behavior expressivity. We computed the mean wrists acceleration, jerk, and velocity of the different PATS speakers. Results are shown in Figure 4.14 (a), (b) and (c). We can notice that wrists acceleration, velocity and jerk are correlated, which was expected since jerk is the derivative of acceleration, and second derivative of velocity, and since acceleration is the first derivative of velocity.

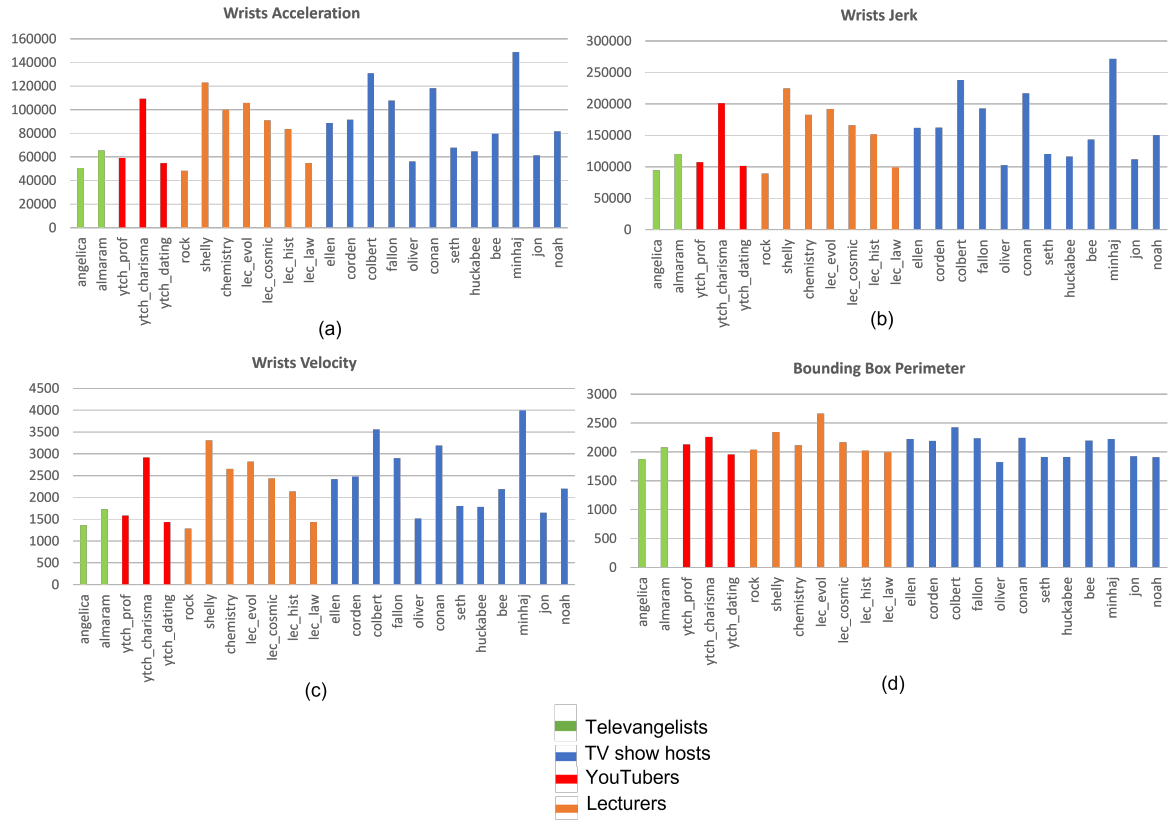


Figure 4.14 The wrists mean acceleration, jerk, velocity, and bounding box perimeter of PATS speakers

Bounding Box Perimeter

The bounding box perimeter was additionally computed for all PATS speakers. Results are shown in Figure 4.14 (d). We additionally computed the Pearson Correlation Coefficient (PCC) score between bounding box perimeter of all speakers and their wrists acceleration. Results show that average bounding box perimeter of each speaker is highly correlated with their wrists acceleration, as *PCC score* is equal to 0.7322.

4.3 PATS Extension

The *PATS Corpus* originally included 2D upper-body joints keypoints, that are aligned with the given speech Mel spectrogram and the different transcript representations.

We extended the *PATS Corpus* to include additional multimodal features related to *2D facial landmarks* which are facial keypoints, as well as *dialog tags*, which are tags that reflect additional text context information.

4.3.1 2D Facial Landmarks

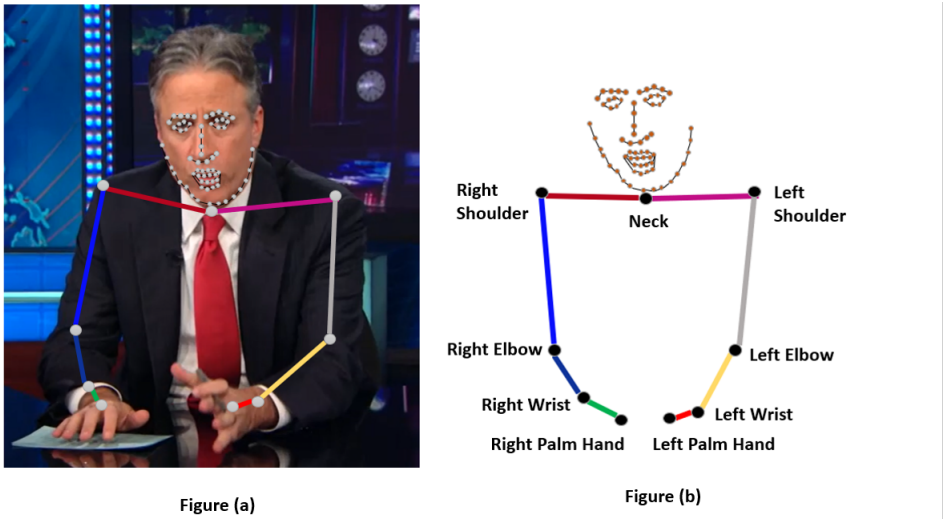


Figure 4.15 Figure (b) illustrates the two-dimensional (2D) facial landmarks, and two-dimensional (2D) upper-body skeleton of joints of the speaker in Figure (a).

We additionally extracted the 2D facial landmarks of the different speakers in PATS using the tool OpenPose (Cao et al. [2017]), to make sure that extracted data and the 2D upper-body keypoints extracted by Ahuja et al. [2020b] are aligned and have the same extraction quality. The extracted features consist of 70 facial landmarks that correspond to the aligned multimodal features that were already extracted by Ahuja et al. [2020b]. Figure 4.15 (b) illustrated the 2D facial landmarks of the speaker in Figure 4.15 (a).

Figure 4.16 provides a closer look at the 70 facial landmarks extracted by OpenPose.

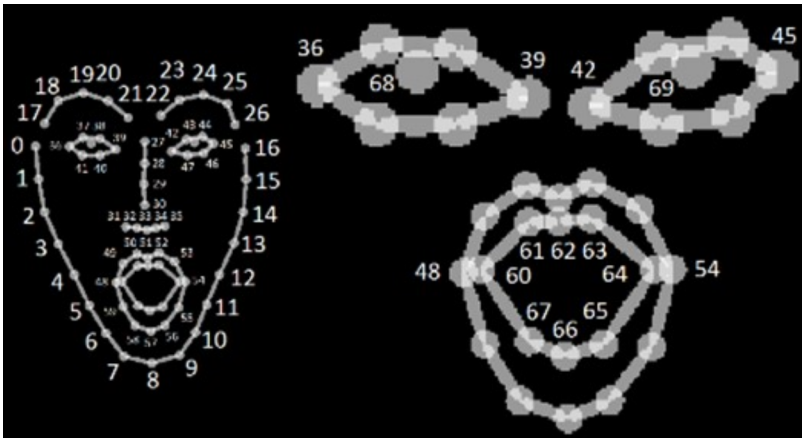


Figure 4.16 70 Facial Landmarks - OpenPose

4.3.2 Dialog Tags

Dialog tag classification is the task of classifying an utterance with respect to the role it serves in the speech. We used the tool "DialogTag" (bha) to extract dialog tags from PATS utterances. The tool "DialogTag" considers 38 different tags that are listed in the following Table 4.3.

TAG	EXAMPLE
Statement-non-opinion	*Me, I'm in the legal department.*
Acknowledge (Backchannel)	*Uh-huh.*
Statement-opinion	*I think it's great*
Agree/Accept	*That's exactly it.*
Appreciation	*I can imagine.*
Yes-No-Question	*Do you have to have any special training?*
Yes answers	*Yes.*
Conventional-closing	*Well, it's been nice talking to you.*
Uninterpretable	*But, uh, yeah*
Wh-Question	*Well, how old are you?*
No answers	*No.*
Response Acknowledgement	*Oh, okay.*
Hedge	*I don't know if I'm making any sense or not.*
Declarative Yes-No-Question	*So you can afford to get a house?*
Other	*Well give me a break, you know.*
Backchannel in question form	*Is that right?*
Quotation	*You can't be pregnant and have cats*
Summarize/reformulate	*Oh, you mean you switched schools for the kids.*
Affirmative non-yes answers	*It is.*
Action-directive	*Why don't you go first*
Collaborative Completion	*Who aren't contributing.*
Repeat-phrase	*Oh, fajitas*
Open-Question	*How about you?*
Rhetorical-Questions	*Who would steal a newspaper?*
Hold before answer/agreement	*I'm drawing a blank.*

Table 4.3 continued from previous page

TAG	EXAMPLE
Negative non-no answers	*Uh, not a whole lot.*
Signal-non-understanding	*Excuse me?*
Conventional-opening	*How are you?*
Or-Clause	*or is it more of a company?*
Dispreferred answers	*Well, not so much that.*
3rd-party-talk	*My goodness, Diane, get down from there.*
Offers, Options Commits	*I'll have to check that out*
Self-talk	*What's the word I'm looking for*
Downplayer	*That's all right.*
Maybe/Accept-part	*Something like that*
Tag-Question	*Right?*
Declarative Wh-Question	*You are what kind of buff?*
Apology	*I'm sorry.*
Thanking	*Hey thanks a lot*

Table 4.3 Table listing the 38 dialog tags that can be extracted using the tool "DialogTag" (bha).

4.4 Conclusion

In this Chapter, we presented *TEDx Corpus* and an extension of *PATS Corpus*. *TEDx Corpus* presents a large amount of data that includes speech audio features, text semantics, upper-facial features, and head features. The extension of *PATS Corpus* includes multimodal features related to 2D facial landmarks, and dialog tags.

The key points of this Chapter:

TEDx Corpus

- *TEDx Corpus* was built to challenge the research questions 1 and 2, which are related to producing speech-driven and semantically-aware facial gestures.
- *TEDx Corpus* aims to study the relationships governing the following modalities of communication: *speech audio*, *text semantics*, *eyebrow* and *head* motion.
- *TEDx Corpus* consists of a large amount of aligned *word-level multimodal features*, which are: upper-facial Action Units, Head Rotations, Text semantics, Voice Prosody, and Voice Quality features.
- *Text* is represented by a sequence of words. Each word is encoded as a *BERT embedding* (Devlin et al. [2019], Wolf et al. [2019]).
- *Voice Prosody* is represented by f_0 word-level contours.
- *Action Units*(AUs) correspond to eyebrow and eyelid motion, which are: $AU1$, $AU2$, $AU4$, $AU5$, $AU6$, and $AU7$.
- *Head Rotations* are represented by 3D head angles, which are R_x, R_y, and R_z: the rotations of the head with respect to X, Y, and Z axes.

PATS Corpus

- *PATS corpus* was originally proposed by Ahuja et al. [2020b] to study the correlations between multimodal features related to speech audio, text semantics, and upper-body pose gestures.
- We extended *PATS corpus* to include additional multimodal features related to 70 *2D facial landmarks*, and 38 different *dialog tags*.

Chapter 5

Semantically Aware and Speech-Driven Facial Gestures Synthesis

Contents

5.1 Introduction

The first form of communication in the lifespan of humans is non-verbal communication. Before humans evolved their ability to speak and use language, they were able to communicate using non-verbal channels of communication (Knapp et al. [2013]). All non-verbal cues are involved in Human-Human Interaction (HHI): body, face, voice, appearance, touch, spatial distancing, and other physical cues. Non-verbal behaviors convey tremendous information to the interlocutors.

One important channel of communication in HHI is the human face. During speech, the face can carry a variety of meaning related to the *verbal* message, the speaker's *emotions*, *attitudes* and *opinions*. Humans use their gaze to convey their desire to switch speaking turns, and their face and body movements to express their thoughts (Burgoon et al. [2016]). During speech, humans frequently use various facial gestures, known as "*visual prosody*" (Graf et al. [2002]). These gestures include *facial* or *head* movement, and are produced jointly with verbal communication and speech prosody. More specifically, speech-driven facial gestures are associated with *speech prosody* and *paralinguistic information*, which both involve various speech characteristics to convey emotions and intentions. Facial gestures are consciously or unconsciously used to adjust speech, accentuate words or word segments, or mark speech pauses. These gestures involve different head movements, eyebrow gestures, frowning, nose wrinkling or lips moistening (Zoric et al. [2007]). One of the key challenges in creating Embodied Conversational Agents is to produce *expressive* visual prosody. Previous data-driven generative models have focused on synthesizing gestures using one modality of human communication; in most of the cases it is *speech*, and in few works it is *text*.

As a first step in this thesis, we focus on addressing the research questions Q1 and Q2 discussed earlier (in Chapter 1), and developing an approach to synthesize coherent human-like facial expressivity in ECAs. In particular, we focus on developing an approach that predicts *expressive facial movements* such as *eyebrows*, *eyelids*, and *head* movements. We do not model mouth movements, as they are not considered in the scope of this thesis. The synthesized gestures are based on different modalities, specifically *audio* data, *text* data, and *facial* data. To the best of our knowledge, synthesizing facial movements based on all the previously discussed modalities has not been investigated at the time of this research.

Model 1: LSTM-based model (*Baseline*). As a starting point, we propose an end-to-end *sequence to sequence LSTM neural network architecture* (Figure 5.1) that predicts *upper-face gestures* and *head motion*, based on both *speech prosody* and *text semantics* (Fares [2020]).

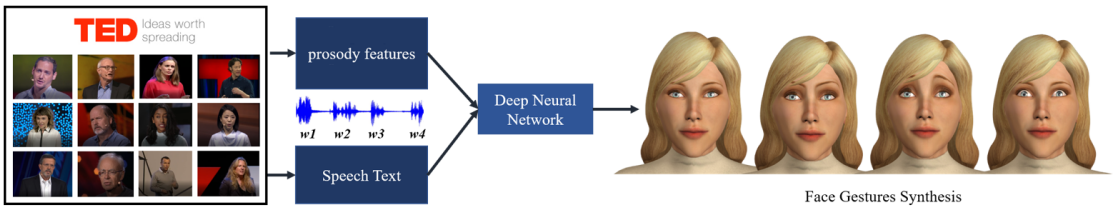


Figure 5.1 An end-to-end LSTM-based neural network is used to generate upper-face gestures and is trained using facial gestures, audio features, and speech text extracted from *TEDx Corpus*.

Recurrent Neural Networks (RNN), specifically *Long Short-Term Memory (LSTM) architectures* are used to exploit the temporal dependencies in the audio data and model prosody variations. We consider multimodal cues when modelling the ECA's facial gestures, since data coming from multiple sources improve RNNs, produce complementary information, and convey patterns that are not discernible when working with individual modalities. We trained, validated, and tested our model using *TEDx Corpus* (discussed in Chapter 4), which we have developed for studying the complex relationships between *speech audio*, *text semantics*, and *facial cues* - more precisely *eyebrow gestures* and *head motion*. This LSTM-based network serves as our *baseline*.

Model 2: Transformer-based model. To overcome some limitations of *Model 1* (lack of learning dependencies and of distributed computations) we propose a second novel approach that makes use of *Transformers* and *Convolutions* to synthesize upper-facial gestures, and head motion based on *speech audio*, and *text semantics*. More specifically, our second approach is also an end-to-end *sequence to sequence model* that takes as input sequences of speech audio and the corresponding text semantics, to generate *eyebrows*, *eyelids* and *head* motion. We compare this approach to the *baseline* model previously described. Similar to the baseline, we train this model on *TEDx Corpus*.

This Chapter is organised as follows. Before we dive into the details of our approach, we first review the main existing approaches for *facial* and *head* motion synthesis and their limitations. We then define and explain the problem we're trying to solve. Next, we review the multimodal input and output features used for training our networks. Then we explain ***Model 1***, its architecture, training regime, and the main objective evaluation conducted on this model. Next, we explain ***Model 2*** architecture, and discuss the objective and subjective evaluations conducted on it. We compare the objective results with those of ***Model 1***.

5.2 Related Works and Limitations

As previously discussed in Chapter 3, a large number of *data-driven* gesture generative models have been proposed with the aim of predicting the behavior of ECAs, and they are principally based on sequential generative parametric models such as *HMMs*, and gradually moving towards deep neural networks enabling spectacular advances over the last few years. Multiple *head motion generation* systems has been proposed in previous works (Hofer and Shimodaira [2007], Haag and Shimodaira [2016], Lu and Shimodaira [2020],

Vougioukas et al. [2019], Mariooryad and Busso [2012], Sadoughi and Busso [2019], Wang et al. [2021]). Hidden Markov Models (HMM) (Hofer and Shimodaira [2007]), Recurrent Neural Networks (RNN) (Wang et al. [2021], Haag and Shimodaira [2016]), and Dynamic Bayesian Networks (DBN) (Mariooryad and Busso [2012], Sadoughi and Busso [2019]) have been used to generate head motion from speech; Generative Adversarial Networks (GAN) have been proposed to produce facial gestures from speech (Karras et al. [2017], Vougioukas et al. [2019]). LSTM networks driven by speech were recently used to predict sequences of gestures (Hasegawa et al. [2018]) and body motions (Shlizerman et al. [2018], Ahuja et al. [2019]). GANs were proposed to generate realistic head motion (Sadoughi and Busso [2018]) and body motions (Ferstl et al. [2019]). *HMMs* were used to predict head motion driven by prosody (Sargin et al. [2008]), and body motion (Levine et al. [2009], Marsella et al. [2013a]).

The aforementioned approaches have multiple limitations:

- Their approaches are not multimodal, they do not exploit multimodal data for synthesizing multimodal behavior. More specifically, they consider as input one modality, namely speech without making their approach semantically-aware. At the output of their models, they synthesize one modality (either facial gestures, head motion, or body movements alone). However, it is necessary to create and define architectures capable of processing sequences of different modalities. It is very important to model the inter-correlations between input and output modalities, possibly occurring on different time scales (utterances, words, etc.). The architectures should be *multimodal* and *multi-scale*.
- When synthesizing facial gestures, they do not model their correlation with head movements which is crucial to produce natural animation. For instance, Hofer and Shimodaira [2007], Haag and Shimodaira [2016], Lu and Shimodaira [2020] and Sadoughi and Busso [2019] do not generate eyebrow motion along with head motion, which are both correlated to f_0 (Yehia et al. [2002]).

In this Chapter, we propose novel paradigm for synthesizing semantically-aware and speech driven facial gestures, and to address the limitations of previous works. As a starting point, we propose an *LSTM-based network* (*Model 1*) for synthesizing the facial gestures, given speech-prosody and text semantics as inputs. To improve the results of this network, we propose another *Transformer-based network* (*Model 2*), whose results surpassed those of the LSTM-based.

5.3 Multimodal Input/Output Features

We used *TEDx Corpus* that we have gathered and which was previously discussed in Chapter 4 to train and test both of our models. It consists of a large amount of multimodal speech and facial features extracted from TEDx talks. We consider the following multimodal features in our approach:

- **Action Units (*AU*)**: *Eyebrows* and *eyelids* motion are represented by the six action units $AU1$, $AU2$, $AU4$, $AU5$, $AU6$, and $AU7$.
- **Head motion (R)**: Head motion is represented by 3D euler head angles: *roll*, *pitch* and *yaw*. The latter angles are represented by R_X, R_Y and R_Z, which are the rotations of the head with respect to X, Y, and Z axes.

- **Audio**: *Speech* is represented by *prosody*, and more specifically f_0 since previous studies (Yehia et al. [2002], Bolinger [1989]) have demonstrated the strong correlation between f_0 and facial gestures (*eyebrows* and *head* motion). We consider f_0 values with a confidence level > 0.3. The ones with confidence level < 0.3 were replaced by the value 0. For each sequence of f_0 corresponding to a word, we applied linear interpolation and extrapolation in order to get a complete sequence of non-zero f_0 values.
- **Text**: *Text semantics* are represented by *BERT embeddings*, since they capture important semantic information about words in context, as they were extracted using BERT, a transformer architecture that have marked the field of *Natural Language Processing*.

Given that AU, R and f_0 values are continuous, they were quantized to generate a finite range of discrete integers. In fact, in deep learning, quantized representations are used to highly reduce the model size and energy consumption, by storing weights using a compact format such as integers instead of floating numbers (Guo [2018]).

Features	
Upper-Face Action Units	$AU1$
	$AU2$
	$AU4$
	$AU5$
	$AU6$
	$AU7$
Head Rotation	R_X
	R_Y
	R_Z
Fundamental Frequency	f_0
Bert	*Bert Embeddings*

Table 5.1 Multimodal Input/Output Features

5.4 Problem Definition

The problem we are trying to solve in this Chapter consists of predicting sequences of Action Units (AUs) relative to *eyebrows* and *eyelids* motion, and sequences of head Euler angles representing *head* motion, based on two the input modalities (1) *speech prosody* and (2) *text semantics*. The input features are sequences of f_0 and $BERTembeddings$ corresponding to a spoken utterance, more specifically to a spoken Inter-Pausal Unit (IPU). An IPU includes a sequence of words, which in turn comprises audio and visual frame sequences. One of the objectives of this work is to propose an encoding schema covering several temporal scales, with the *word-level* as main pivot scale. More specifically,

the input f_0 sequences are encoded both on a *frame-level* and on *word-level*; the input $BERTembeddings$ are at the *word-level*. f_0 values are restricted to the range of 50 to 550Hz (as previously discussed in Chapter 4).

This problem is similar to the *neural machine translation* problem which consists of mapping a sequence of words to another sequence of words in another language (Neubig [2017], Sutskever et al. [2014]). The model used in such problems is called *Sequence to Sequence (Seq2Seq)*. Both of our networks - ***Model 1*** and ***Model 2*** - (which are discussed in the following sections) are *Seq2Seq* models that take as input:

1. Sequence of f_0-values represented by X_{speech}=(f_{01}, ..., f_{0N_W}), where N_W is the number of f_0-values corresponding to the spoken word ***W***,
2. *BERT embedding* vector of the spoken word ***W***, represented by X_{text}.

The outputs of the network are:

1. Sequence of AU-values represented by $\widehat{Z}^{(j)}_{AU}(X) = (AU^{(j)}_1, \ldots, AU^{(j)}_{N'_W})$ where j denotes the j^{th} ***AU*** and N'_W the number ***AU***-values corresponding to the word ***W***
2. Sequence of R-values represented by $\widehat{Z}^{(k)}_{R}(X) = (R^{(k)}_1, \ldots, R^{(k)}_{N'_W})$) where k denotes the k^{th} ***R*** and N'_W the number ***R***-values corresponding to the word ***W***.

For the sake of clarity, we will denote $\widehat{Z}^{(j)}_{AU} = \widehat{Z}^{(j)}_{AU}(X)$ and $\widehat{Z}^{(k)}_{R} = \widehat{Z}^{(k)}_{R}(X)$ in the remaining of this dissertation. Unlike previous works, we encode modalities at the word-scale, and operate the *Seq2Seq* transcoding from these word representations.

5.5 Model 1: LSTM-based Network for Facial Gestures Synthesis

We first propose a learning-based facial gestures generation model. The model architecture is an end-to-end sequence-to-sequence neural network model that consists of an encoder $E_{\textbf{speech, text}}$ to encode the input features, and a decoder D_{face} to generate the sequences of the AU/Rs that are related to eyebrows, eyelids, and head movements.

The objective is to build the architecture described in the previous section. To build it and to facilitate learning, we pre-trained two auto-encoders AE_{speech} and AE_{face}, to learn compressed representations of f_0 and AU/R modalities, and be able to reconstruct the original data from that representation. For the f_0 modality, only the encoder part is then kept, and for the AU/R modality only the decoder part is used. We present in the following sections the autoencoders AE_{speech} and AE_{face}. We then discuss *Model 1*'s overall network architecture, its training regime, and an objective evaluation that we have conducted to assess it, as well as the results.

5.5.1 Multimodal Pre-Net Encoders

In this ***Model 1***, we used autoencoders for *dimensionality reduction*. We developed and trained two different autoencoders AE_{face} and AE_{speech} that compress AU/R and f_0 values into a lower-dimensional representation, and then convert them back to a reconstruction of the original input. For each autoencoder, the encoder - E_{speech} or E_{face} - and

decoder - D_{speech} or D_{face} - were trained jointly and used independently as different components in *Model 1*'s overall network architecture, which is described in the next section. The autoencoders AE_{face} and AE_{speech} are based on Long-Short Term Memory (LSTM), a neural architecture to model sequences, with better transmission of temporal information than the transmission in a classical recurrent neural network (Hochreiter and Schmidhuber [1997]).

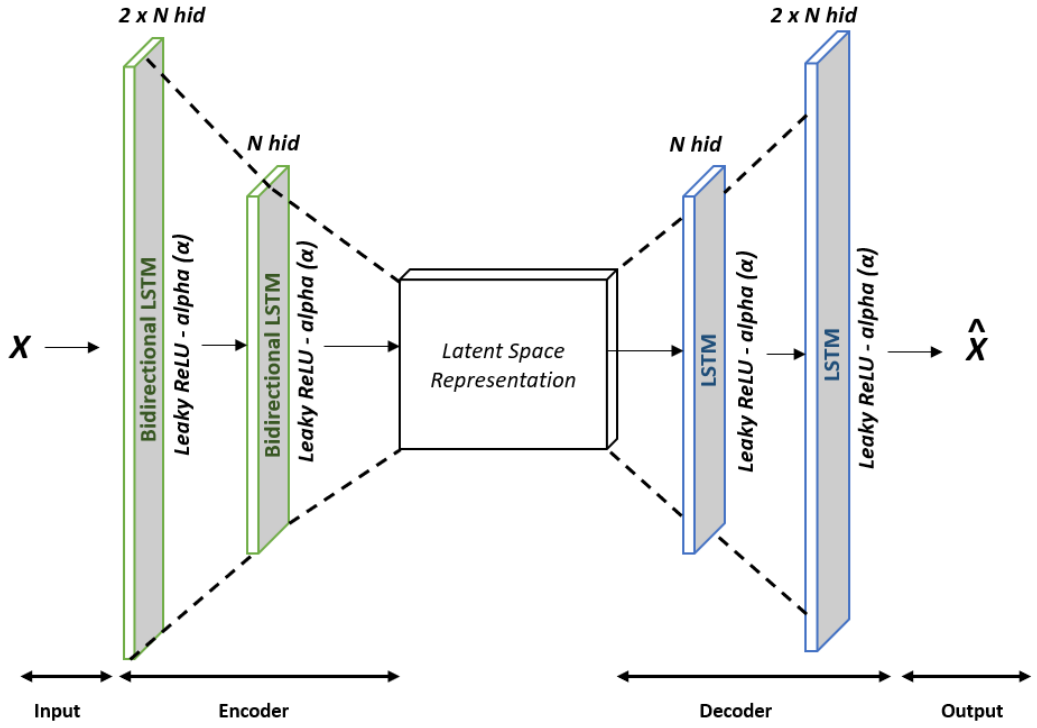


Figure 5.2 AE_{face} and AE_{speech} architecture. It takes as input X_{speech} (X_{face} resp.), encodes it into a latent representation h_{speech} (h_{face} resp.) then generate $\widehat{X}_{speech}$ ($\widehat{X}_{face}$ resp.) which is the reconstruction of the input. It is composed of an encoder E_{speech} (E_{face} resp.) and a decoder D_{speech} (D_{face} resp.).

AE_{face} and AE_{speech} have the same architecture, which is illustrated in Figure 5.2. Each autoencoder is composed of an encoder $\boldsymbol{E}$ (E_{speech} and E_{face}) and a decoder $\boldsymbol{D}$ (D_{speech} and D_{face}).

Encoder (E_{speech} and E_{face}). The encoder E_{speech} takes X_{speech} (f_0) as input, and the output of E_{speech} is the vector h_{speech}, which is the latent space representation of the input X_{speech}. Similar to E_{speech}, E_{face} takes X_{face} (AU/R) as input and generates the latent space representation of X_{face} which is the vector h_{face}. The latent space representations h_{speech} and h_{face} can be written as follows:

$$\begin{aligned} h_{speech} &= E_{speech}(X_{speech}) \\ h_{face} &= E_{face}(X_{face}) \end{aligned} \tag{5.1}$$

Decoder (D_{speech} and D_{face}). The vector h_{speech} (resp. h_{face}) is then given as input to the decoder D_{speech} (resp. D_{face}) which then generates the output $\widehat{X}_{speech}$ (resp. $\widehat{X}_{face}$),

which is the reconstruction of the input X_{speech} (resp. X_{face}). $\widehat{X}_{speech}$ and $\widehat{X}_{face}$ can be written as follows:

$$\begin{aligned} \widehat{X}_{speech} &= D_{speech}(h_{speech}) \\ \widehat{X}_{face} &= D_{face}(h_{face}) \end{aligned} \tag{5.2}$$

The loss function used for both networks is MSE (Mean Squared Error) and it is calculated between the probability distribution of $\widehat{X}_{speech}$ and X_{speech} (resp. $\widehat{X}_{face}$ and X_{face}). It can be written as follows:

$$\begin{aligned} \mathcal{L}_{speech}(E_{speech}, D_{speech}) &= \mathbb{E}_{X_{speech}}(X_{speech} - \widehat{X}_{speech})^2 \\ \mathcal{L}_{face}(E_{face}, D_{face}) &= \mathbb{E}_{X_{face}}(X_{face} - \widehat{X}_{face})^2 \end{aligned} \tag{5.3}$$

5.5.2 Pre-Net Encoders Implementation Details

E_{speech} (resp. E_{face}) is composed of two layers of Bi-directional LSTMs with $2 \times \mathbf{N}_{hid}$ units for the first layer and $\mathbf{N}_{hid}$ units for the second, where $\mathbf{N}_{hid}$ is equal to 100.

D_{speech} (resp. D_{face}) is composed of two layers of LSTM with $\mathbf{N}_{hid}$ and $2 \times \mathbf{N}_{hid}$ units respectively. Each LSTM and Bidirectional LSTM layer in this network is followed by a *Leaky ReLU* activation function with $\alpha = 0.05$.

We used $Adam$ optimizer for training both networks. We tuned both networks by performing a *Grid Search*, which is an exhaustive search to find the optimal following hyperparameters: $\mathbf{N}_{hid}$, batch size BS, and number of epochs N_{ep}. The optimized hyperparameters we found for each network are summarized in Table 5.2. We reached an accuracy equals to 96% for AE_{speech}, and 94% for AE_{face}.

Network	Hyperparameter	Value
AE_{face}	$\mathbf{N}_{hid}$	100
	BS	128
	N_{ep}	300
AE_{speech}	$\mathbf{N}_{hid}$	100
	BS	10
	N_{ep}	500

Table 5.2 Optimal hyperparameters found after performing a *Grid Search* for AE_{face} and AE_{speech}

5.5.3 Model 1: Sequence to Sequence Neural Architecture

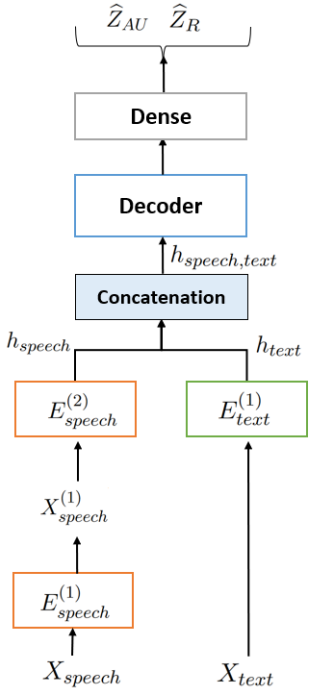


Figure 5.3 *Model 1* Network Architecture.

The overall network architecture of ***Model 1*** is depicted in Figure 5.3. The encoder E_{speech} of AE_{speech} is used to encode X_{speech}, and the decoder D_{face} of AE_{face} is used for decoding the AUs.

The network has two encoding levels for X_{speech}, and one decoding level for Z_{face}. X_{speech} is first encoded by $E^{(1)}_{speech}$ which corresponds to the pre-trained E_{speech} of AE_{speech}. The output vector $X^{(1)}_{speech}$ is then given as input to $E^{(2)}_{speech}$ which is composed of 2 layers of bidirectional LSTMs, and produces the output vector h_{speech}. Bert embedding X_{text} is given as input to $E^{(1)}_{text}$ which is also composed of 2 layers of bidirectional LSTMs, and produces the vector h_{text}. The vectors h_{speech} and h_{text} are then concatenated together and the resulting vector is $h_{speech,text}$ which can be written as follows:

$$h_{speech,text} = [h_{speech}, h_{text}] = [E^{(2)}_{speech}(E^{(1)}_{speech}(X_{speech})), E^{(1)}_{text}(X_{text})] \tag{5.4}$$

$h_{speech,text}$ is then transmitted to 9 D_{face} to produce the corresponding $\widehat{Z}_{AU}$ and $\widehat{Z}_R$. For simplicity, we only illustrate one D_{face} in Figure 5.3. The decoders are followed by *Dense* layers with *Softmax Activation*, and produce the corresponding $\widehat{Z}_{AU/R}$. $\widehat{Z}_{AU}$ can be written as:

$$\widehat{Z}_{AU} = D^{j}_{face}(h_{speech,text}) \tag{5.5}$$

where j denotes the j^{th} AU; and $\widehat{Z}_R$ can be written as:

$$\widehat{Z}_R = D^{k}_{face}(h_{speech,text}) \tag{5.6}$$

where k denotes the k^{th} R.

The categorical cross-entropy is calculated between the probability distribution of $\widehat{Z}_{AU}$ and Z_{AU}, and between the probability distribution of $\widehat{Z}_R$ and Z_R. It can be stated formally as *H(T, S)* where *H(.)* is the cross-entropy function, T the target distribution (Z_{AU} or Z_R) and S is the approximation of the target distribution ($\widehat{Z}_{AU}$ or $\widehat{Z}_R$). The total loss can therefore be written as follows:

$$\begin{aligned}\mathcal{L}_{total}(E^{(1)}_{speech}, E^{(2)}_{speech}, E^{(1)}_{text}, D^{n}_{face}) &= \textstyle\sum_j \mathbb{E}_{\widehat{Z}^{(j)}_{AU}} ||H(Z^{(j)}_{AU}, \widehat{Z}^{(j)}_{AU})||_2 \\ &+ \textstyle\sum_k \mathbb{E}_{\widehat{Z}^{(k)}_{R}} ||H(Z^{(k)}_{R}, \widehat{Z}^{(k)}_{R})||_2\end{aligned} \tag{5.7}$$

where n denotes the n^{th} D_{face}, j denotes the j^{th} *AU*, and k denotes the k^{th} *R* .

5.5.4 Implementation Details

The hyperparameters of **Model 1** are as follows. For both $E^{(2)}_{speech}$ and $E^{(1)}_{text}$, each first layer of bidirectional LSTM has $\mathbf{N}_{hid}$=200 units, and the second one has $\mathbf{N}_{hid}$=100 units.

The hyperparameters used in both $E^{(2)}_{speech}$ and $E^{(1)}_{text}$ of **Model 1** are summarized in Table 5.3.

Component	Layer	Parameter	Value
$E^{(2)}_{speech}$	Bidirectional LSTM - layer 1	$\mathbf{N}_{hid}$	200
	Bidirectional LSTM - layer 2	$\mathbf{N}_{hid}$	100
$E^{(1)}_{text}$	Bidirectional LSTM - layer 1	$\mathbf{N}_{hid}$	200
	Bidirectional LSTM - layer 2	$\mathbf{N}_{hid}$	100

Table 5.3 **Model 1** hyperparameters

The activation function used in these layers is *LeakyReLU* with $\alpha = 0.01$. We trained it on N_{ep}=300 epochs, using a batch size $BS = 128$. We used *Root Mean Squared Propagation* ($RMSProp$) optimizer. The training hyperparameters are summarized in Table 5.4.

Hyperparameter		Value
Batch Size	BS	128
Number of epochs	N_{ep}	300
LeakyReLU	α	0.01

Table 5.4 Baseline training hyperparameters

5.5.5 Material and Experimental Setups

We trained, validated, and tested our model on a subset of the *TEDx Corpus* (Chapter 4), containing preprocessed ***AUs/R***, f_0s, and $BERTembeddings$ of filtered shots where speakers' face and head are visible and close to the camera. Our subset consists of the features of 200 videos. Videos vary between 2 and 25 minutes, with a frame rate of 24 FPS, the total numbers of IPUs is 919, and of words is 62307. We shuffled all the IPUs, then split them into: training set (80%), validation set (10%) and test set (10%).

There are two test conditions: SD ($SpeakerDependent$) and SI ($SpeakerIndependent$). The SD condition aims to assess to what extent the model can generalize on new sentences pronounced by a speaker seen during training - training set includes multiple speakers. The SI condition aims to assess the extent to which gestures predictions can be extrapolated to unseen speakers (not part of the training set). ***Model 1*** is only evaluated on the SD set, the SI condition is used later on for assessing ***Model 2***.

5.5.6 Objective Evaluation

To assess the quality of the generated gestures, we used the following measures:

1. Root Mean Squared Error (***RMSE***),
2. Pearson Correlation Coefficient (***PCC***),
3. Activity Hit Ratio (***AHR***)
4. Non-Activity Hit Ratio (***NAHR***).

AHR and ***NAHR*** were proposed by Ong et al. [2017], to evaluate the performance of Voice Activity Detector (VAD) systems. We also considered them since evaluating AU activity looks similar to VAD evaluation. We considered an AU as ***"Activated"*** when its value is greater than ***0.5***, otherwise it is ***"Not-Activated"***. ***AHR*** is the percentage of predicted AU activation with respect to ground truth. If it is greater than 100%, it means that the model is predicting more activation than the amount of activation that is in the ground truth. ***NAHR*** is the same but for non-activity.

We assessed the full model using the ***SD*** test set.

5.5.7 Objective Evaluation Results and Discussion

Table 5.5 reports the objective evaluation results of ***Model 1***, using the ***SD*** test set. Results reveal that RMSE errors ($0.20 \leq error \leq 0.97$)) are low for some ***AU/R*** such as $AU1$, $AU7$ and R_Z. Error is higher for other ***AU/R*** like $AU5$, $AU5$ and R_Y. The ***PCC*** scores are close to zero for all the features, which means that there is almost no correlation between the generated ***AU/R*** values and the ground truth. AHR and NAHR were calculated to measure the activation of AUs only, since they are not applicable for R. The ***AHR*** is less than 60% for all the features. This means that the percentage of the predicted AUs' activation is low compared to the activation in the ground truth. The amount of non-activation is high for all the features ($107\% \leq$ ***NAHR*** $\leq 135.9\%$). This means that the model is predicting more non-activation than the amount of non-activation that is in the ground truth.

Model 1 has some limitations which are due to the large number of operations required to relate signals from two random input or output positions when processing sequences. It is difficult to learn dependencies between distant positions (Hochreiter et al. [2001], Vaswani et al. [2017]). These limitations are due to the large number of operations required in LSTM layers to connect signals from input or output positions when processing sequences. It is very complicated to learn dependencies between distant positions when using LSTMs for processing temporal sequence (Hochreiter et al. [2001], Vaswani et al. [2017])s.

In a Transformer Network, the number of operations is reduced and is maintained constant, although this reduction is done at the cost of reduced effective resolution due to averaging attention-weighted positions. This effect is overcome by the usage of *attention mechanisms*, and especially the *Multi-Head Attention* which will be described in later sections. The Transformer network relies entirely on the attention mechanisms to compute representations of its input and output without using any sort of recurrence like in RNNs (Vaswani et al. [2017]).

	Model 1 (SD)			
	RMSE	**PCC**	**AHR**	**NAHR**
AU1	0.20	-0.012	42.56	115.95
AU2	0.48	-0.002	34.15	107.08
AU4	0.53	-0.012	60.31	121.85
AU5	0.50	-0.011	21.12	135.95
AU6	0.48	-0.002	33.11	135.05
AU7	0.33	-0.053	20.21	131.52
$\mathbf{R}_X$	0.53	0.018	*NA*	*NA*
$\mathbf{R}_Y$	0.97	-0.024	*NA*	*NA*
$\mathbf{R}_Z$	0.22	0.003	*NA*	*NA*

Table 5.5 Objective Evaluation of LSTM-based Model (*Model 1*)

5.6 Model 2: Transformer-based Model for Facial Gesture Synthesis

As discussed in the previous section (5.5.7), ***Model 1*** has some limitations including low activation of the predicted AUs, low correlation between the predictions and the ground truth, and high RMSE errors for some AU/R predictions. Before 2017, the LSTM was the most optimal architecture for neural machine translation tasks. Later on, within the past few years, the Transformer architecture was explored as an alternative, and has been assessed to outperform the LSTM within these neural machine translation tasks. In the Transformer network (Vaswani et al. [2017]), the number of operations is reduced and is constant. This is achieved by the usage of *attention mechanisms*, and especially the *Multi-Head Attention*. They are used to compute representations of its input and output and accomplish dependencies without using any sort of recurrence like in RNNs (Vaswani et al. [2017]). Hence, the Transformer allows for notably more parallelization and can attain a new state of the art in translation quality. Moreover, Transformer networks and attention mechanisms have been recently proved to be very efficient for *sequence-to-sequence modelling*, with particular advances for modelling *multimodal processes*. For instance, they were previously used for translating speech to text (ASR) (Hrinchuk et al. [2020], Mohamed et al. [2019]), and multimodal learning of images based on text (Yao and Wan [2020]).

To overcome the weaknesses of ***Model 1***, we propose a novel approach for *upper-facial* and *head* gestures generation based on a multimodal Transformer network. A transformer network is presented to handle the sequence-to-sequence modelling of multimodal upper-facial and head movements at the word-level. As inputs, the transformer exploits both

acoustic - X_{speech} - and semantic - X_{text} - information, by adding BERT semantic embeddings to word-level encoded f_0 contours. As outputs, modelling the correlation between head motion and upper-facial gestures allows the generation of a more coherent and natural behavior of the agent [1].

We propose a new architecture that includes:

1. A transformer network operating on multi-modal input text and speech information in order to generate upper-facial and head movements,
2. A cross-attention module that can efficiently exploit semantic and speech information.

We discuss the architecture, implementation details as well as the conducted objective and subjective evaluations in the following sections.

5.6.1 Neural Transformer Architecture with Cross-Attention

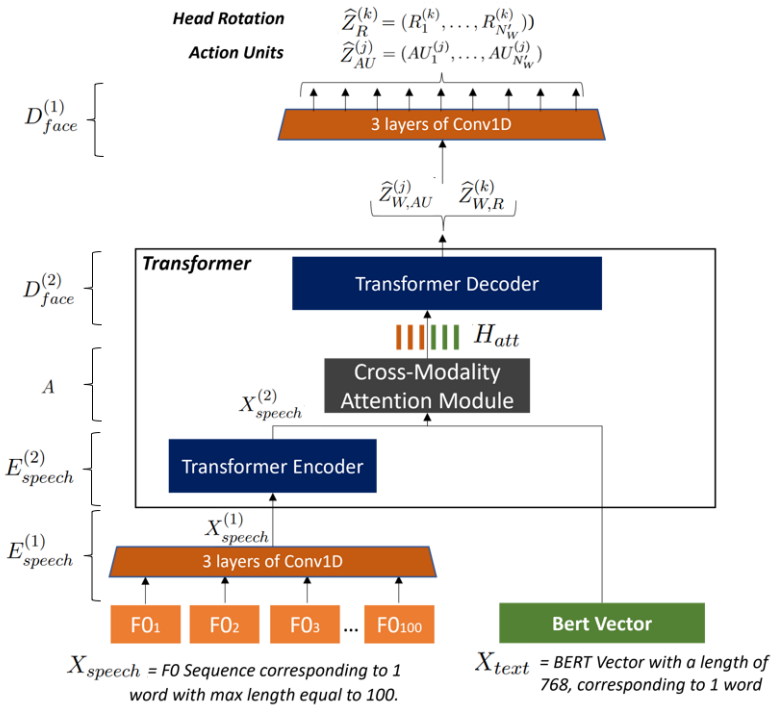


Figure 5.4 *Model 2* Architecture - The Transformer network operates on multi-modal input text and speech information to generate upper-facial and head movements. The network takes word-level X_{speech} and X_{text} as input and generates the corresponding word-level $\mathbf{Z}_{W,AU}^{(j)}$ and $\mathbf{Z}_{W,R}^{(k)}$. A *cross-attention* mechanism is applied on both encoded input modalities to exploit semantic and speech information and generate an embedding that represents efficiently both modalities.

[1] Video samples of our model's gestures predictions and other related material can be found in: https://github.com/mireillefares/VAAnimation/blob/main/README.md

The proposed architecture aims at mapping the multimodal speech and text feature sequence into continuous facial and head gestures. As previously stated, this problem is treated as a multimodal *sequence-to-sequence* (S2S) problem, for which a transformer network operating at the word level is presented. The network is illustrated in Figure 5.4. The inputs and outputs of the transformer network consist of one feature vector for each word ***W*** of the input text sequence, which corresponds to an IPU.

In order to handle continuous flow of input and output information with different timing, the Transformer is wrapped with a f_0 encoder module - $E^{(1)}_{speech}$ - at its input and a AU/R decoder - $D^{(2)}_{face}$ - at its output. The objective of $E^{(1)}_{speech}$ is to encode the continuous f_0-values at the word level and the objective of $D^{(2)}_{face}$ is to reconstruct the continuous values for Z_{face}, more specifically $\widehat{Z}_{AU}$ and $\widehat{Z}_{R}$, from word-level encodings $\widehat{Z}_{W,AU}$ and $\widehat{Z}_{W,R}$.

At the input of the network, each spoken word ***W*** is represented by:

1. A word-level f_0 embedding vector $X^{(1)}_{speech}$. $X^{(1)}_{speech}$ corresponds to the encoding of X_{speech} which is the sequence of f_0-values $X_{speech} = (f_{01}, \ldots, f_{0N_W})$, where N_W is the number of f_0 values corresponding to the spoken word ***W***.
2. A word embedding vector X_{text}. X_{text} is the word *BERT embedding* vector corresponding to the spoken word ***W***, including silences. In our approach, silences less than 0.2 secs may belong to IPUs. Silences do not have a contextual *BERT embedding*. Hence, we replaced them by a comma - ",".

At the output of the network, each spoken word ***W*** is represented by:

1. A word-level ***AU*** vector $\mathbf{Z}^{(j)}_{W,AU}$ which is the encoding of the sequence of values $\mathbf{Z}^{(j)}_{AU} = (AU^{(j)}_1, \ldots, AU^{(j)}_{N'_W})$ where j denotes the j^{th} ***AU*** and N'_W the number ***AU/R***-values corresponding to the word ***W***.
2. A word-level ***R*** vector $\mathbf{Z}^{(k)}_{W,R}$ which is the encoding of the sequence of values $\mathbf{Z}^{(k)}_{R} = (R^{(k)}_1, \ldots, R^{(k)}_{N'_W})$ where k the k^{th} ***R***, and N'_W the number ***AU/R***-values corresponding to the word ***W***.

Fundamental Frequency Encoder ($E^{(1)}_{speech}$). $E^{(1)}_{speech}$ takes as input X_{speech} which the f_0 sequence corresponding to a ***W***, and projects it into a word-level representation of f_0 contours covering local context of f_0 variations. The generated output vector of the latter layers $X^{(1)}_{speech}$ is then fed as an input to a *Transformer Encoder* $E^{(2)}_{speech}$. The output vector can be written as follows:

$$X^{(1)}_{speech} = E^{(1)}_{speech}(X_{speech}) \tag{5.8}$$

Transformer Encoder ($E^{(2)}_{speech}$). $E^{(2)}_{speech}$ takes as input the encoded $X^{(1)}_{speech}$. $E^{(2)}_{speech}$ consists of multiple encoding blocks, which employ the *self-attention* mechanism to enrich each token (embedding vector) with contextual information from the input encoded sequence $X^{(1)}_{speech}$. The *self-attention* mechanism uses multiple heads (parallel attention computations) so that the model can tap into multiple embedding subspaces. The output of the last encoding block is $X^{(2)}_{speech}$, which can be written as follows:

$$X^{(2)}_{speech} = E^{(2)}_{speech}(X^{(1)}_{speech}) \tag{5.9}$$

Cross-Modality Attention Module (A). The output $X^{(2)}_{speech}$ of the Transformer encoder $E^{(2)}_{speech}$, as well as X_{text} are fed as inputs to the *Cross-Modality Attention Module (A)* (see Figure 5.4). This Module has the same structure as the Transformer decoder in Vaswani et al. [2017]. It generates H_{att}, a representation that can take into account both modalities, text semantics and speech prosody. The representation learning is done in a master/slave manner, where one modality - the master - is used to highlight the extracted features in the other modality - the slave. This module takes X_{text} - text modality - as master, and $X^{(2)}_{speech}$ - speech modality - as slave. Thus, it performs cross-attention such that the attention mask is derived from text modality, and is harnessed to leverage the latent features from the speech modality. More specifically, cross-attention combines asymmetrically the two input modalities, in contrast to the *self-attention* used in $E^{(2)}_{speech}$, which applies attention on only the speech spectral features. The A module processes the input vectors as follows:

1. A takes as input $X^{(2)}_{speech}$ and X_{text}.
2. It then computes the *key* K and *value* V from $X^{(2)}_{speech}$.
3. The *Queries* Q are then calculated from X_{text}.
4. An attention matrix is then computed from K and Q; and Q is applied to the attention matrix.
5. The output vector H_{att} have the same dimension as X_{text}.

To compute the feature representations K, Q, and V, the corresponding input vector X ($X^{(2)}_{speech}$ or X_{text}) of n tokens of dimensions d, $X \in R^{n \times d}$, is projected using the following 3 matrices:

$$\begin{aligned} W_Q &\in R^{n \times d_q} \\ W_K &\in R^{n \times d_k} \\ W_V &\in R^{n \times d_V} \end{aligned} \tag{5.10}$$

Then, K, Q, and V are calculated as follows:

$$\begin{aligned} Q &= X \times W_Q \\ K &= X \times W_K \\ V &= X \times W_V \end{aligned} \tag{5.11}$$

H_{att} can therefore be written as follows:

$$\begin{aligned} H_{att} &= A(X^{(2)}_{speech}, X_{text}) \\ &= Softmax((W_Q \times X_{text}) \times (W_K \times X^{(2)}_{speech})^T) \times (W_V \times X^{(2)}_{speech}) \\ &= Softmax(Q \times K^T) \times V \end{aligned} \tag{5.12}$$

where $A(.)$ denotes cross-attention.

Transformer Decoder ($D^{(2)}_{face}$). $D^{(2)}_{face}$ takes as input H_{att} which is the multimodal representation of the input modalities, and outputs the corresponding word-level $\widehat{Z}^j_{W,AU}$ and $\widehat{Z}^k_R$. $D^{(2)}_{face}$ consists of multiple decoding blocks which can be thought of encoding blocks

generating enriched embeddings useful for translation outputs. $D^{(2)}_{face}$ architecture is similar to $E^{(2)}_{speech}$, except it calculates the H_{att} - target attention. $D^{(2)}_{face}$ has a supplemental third sub-layer, unlike $E^{(2)}_{speech}$. This latter layer performs *multi-head attention* over the output of the $E^{(2)}_{speech}$ stack. We use 9 *Transformer decoder*s, one for each ***AU/R***. For simplicity, Figure 5.4 only illustrates one decoder. The output vectors $\widehat{Z}^{j}_{W,AU}$, car be written as follows:

$$\widehat{Z}^{j}_{W,AU} = D^{(2)^j}_{face}(H_{att}) \tag{5.13}$$

where j denotes the j^{th} *AU*. $\widehat{Z}^{k}_{R}$ can be written as follows:

$$\widehat{Z}^{k}_{W,R} = D^{(2)^k}_{face}(H_{att}) \tag{5.14}$$

where k denotes the k^{th} *R*.

AU/R decoder ($D^{(1)}_{face}$): As depicted in Figure 5.4, the $D^{(2)}_{face}$ outputs are concatenated together, then fed to $D^{(1)}_{face}$ to learn the correlation between the output features, and therefore the correlation between *facial* and *head* movements. Finally, a *Dense* layer with a *Softmax* activation function is applied on each of the outputs, to convert the outputs to predicted next-token probabilities. The final output sequences are $\widehat{Z}^{(j)}_{AU}$ and $\widehat{Z}^{(k)}_{R}$. $\widehat{Z}^{(j)}_{AU}$ can be written as follows:

$$\widehat{Z}^{(j)}_{AU} = D^{(1)}_{face}(\widehat{Z}^{j}_{W,AU}) \tag{5.15}$$

where j denotes the j^{th} *AU*. $\widehat{Z}^{(k)}_{R}$ can be written as follows:

$$\widehat{Z}^{(k)}_{R} = D^{(1)}_{face}(\widehat{Z}^{k}_{W,R}) \tag{5.16}$$

where k denotes the k^{th} *R*.

Loss $\mathcal{L}_{total}$. The loss function used is the *categorical cross-entropy*, which is calculated between the probability distribution of $\widehat{Z}^{j}_{AU}$ and Z^{j}_{AU}, and between the probability distribution of $\widehat{Z}^{k}_{R}$ and Z^{k}_{R}. The total loss can be written as follows:

$$\begin{aligned}\mathcal{L}_{total}(E^{(1)}_{speech}, E^{(2)}_{speech}, A, D^{(2)}_{face}, D^{(1)}_{face}) &= \textstyle\sum_j \mathbb{E}_{\widehat{Z}^{(j)}_{AU}} ||H(Z^{(j)}_{AU}, \widehat{Z}^{(j)}_{AU})||_2 \\ &+ \textstyle\sum_k \mathbb{E}_{\widehat{Z}^{(k)}_{R}} ||H(Z^{(k)}_{R}, \widehat{Z}^{(k)}_{R})||_2\end{aligned} \tag{5.17}$$

where j denotes the j^{th} *AU*, and k denotes the k^{th} *R*.

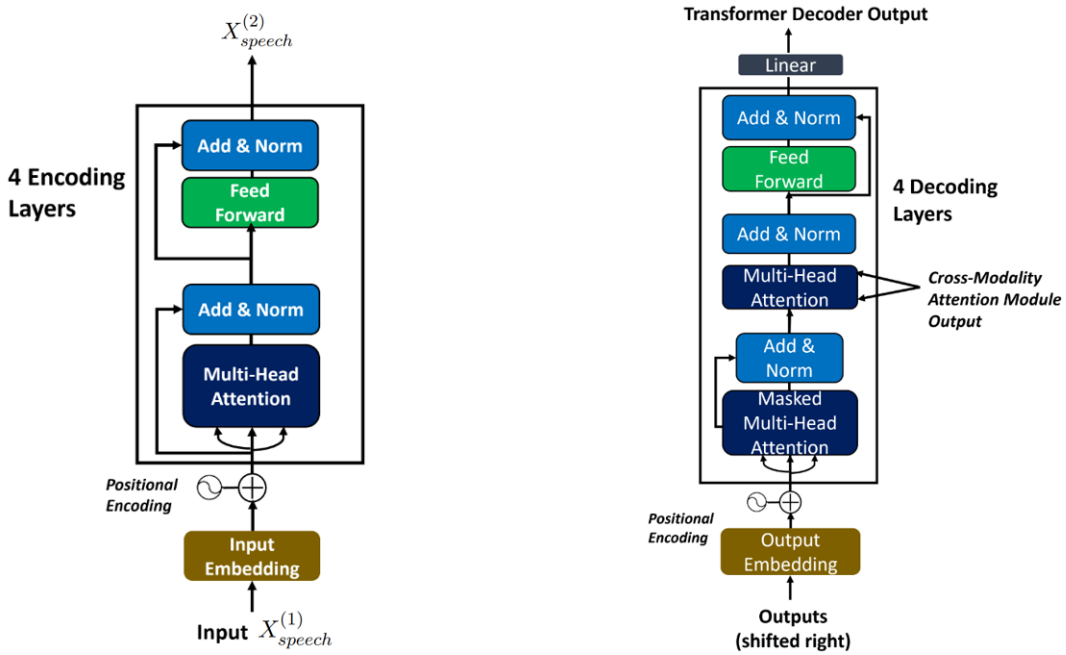
5.6.2 Implementation Details

Below we discuss the implementation details of ***Model 2***.

Fundamental Frequency Encoder ($E^{(1)}_{speech}$). As depicted in Figure 5.4, for each ***W***, three one-dimensional convolutional layers are applied to project the input f_0 sequence X_{speech} into a word-level representation of f_0. These convolutional layers include a number of filters N_{filt} equal to *64*, with a kernel size K_{size} equal to *3*.

Transformer Encoder **($E^{(2)}_{speech}$).** The *Transformer Encoder* architecture is depicted in Figure 5.5 (a) ; it is similar to the one proposed in Vaswani et al. [2017]. In our work, it is composed of a stack of $N_{enc} = 4$ identical encoding layers. Each layer has two sub-layers: the first one is a multi-head self attention mechanism with $N_h = 4$ attention heads, and the second one is a position-wise fully connected feed-forward network. As the original transformer encoder, we employ a residual connection around each of the 2 sub-layers, followed by layer normalization.

Transformer Decoder ($D^{(2)}_{face}$). The Transformer decoder is composed of $N_{dec} = 4$ identical decoding layers, with $N_h = 4$. Similar to the one proposed in Vaswani et al. [2017], it is composed of residual connections applied around each of the sub-layers, followed by layer normalization. As depicted in Figure 5.5 (b), the self-attention sub-layer in the decoder stack is modified to prevent positions from attending to subsequent positions. The output predictions which are offset by one position, and this masking ensure that the predictions for position index ***j*** (resp. ***k***) depend only on the known outputs at positions less than ***j*** (resp. ***k***). $D^{(2)}_{face}$'s sub-layers, as well as the sub-layers in $E^{(2)}_{speech}$, apply residual connections around each of the sub-layers and then performs layer normalization. The sub-layers in the $D^{(2)}_{face}$ stack are *masked*, so that we can prevent positions from attending to subsequent positions.



(a) Transformer Encoder (b) Transformer Decoder

Figure 5.5 *Transformer Encoder and Decoder*

AU/R decoder ($D^{(1)}_{face}$). As depicted in Figure 5.5, the *Transformer Decoder* outputs are concatenated together, then fed to 3 one-dimensional convolutional layers that include N_{filt} equals to 64 filters, with a kernel size K_{size} equals to 3, to learn the correlation between the output features, and therefore the correlation between *facial* and *head* movements. Finally, a *Dense* layer with a *Softmax* activation function is applied on each of the outputs, to convert the outputs to predicted next-token probabilities. The final output se-

quences are $\widehat{Z}_{AU}^{(j)}$ and $\widehat{Z}_{W,R}^{(k)}$.

Transformer Sub-Layers. $E_{speech}^{(2)}$ and $D_{face}^{(2)}$ have attention sub-layers, and contain fully connected feed-forward networks which are applied to each position separately and identically. Similarly to other sequence to sequence models, we use learned embeddings to convert the input tokens and output tokens to vectors of dimension $d_{model} = 64$. All sub-layers and embedding layers therefore use this dimension. The inner feed-forward layers are of dimension $N_{hid} = 400$. Positional encodings are applied to the inputs of the transformer encoder and decoders. They have the same dimension as the embeddings, so that they can be added together. We use sine and cosine functions, similar to Vaswani et al. [2017].

X_{speech}, Z_{AU} and Z_R have a variable length. We set the maximum X_{speech} input sequence length to 100, and the maximum $Z_{AU/R}$ output length to 124. Shorter sequences were padded to the maximum length, and longer ones were truncated. *Model 2* hyperparameters were chosen empirically and are summarized in Table 5.6.

Component	Hyperparameter	Value
$E_{speech}^{(1)}$	N_{filt}	64
	K_{size}	3
$D_{face}^{(2)}$	N_{dec}	4
	N_h	4
	N_{hid}	400
$E_{speech}^{(2)}$	N_{enc}	4
	N_h	4
	N_{hid}	400
$D_{face}^{(1)}$	N_{filt}	64
	K_{size}	3
Overall Network	d_{model}	64

Table 5.6 Model Hyperparameters

Each training batch contained BS=128 pairs of X_{text}, X_{speech}, and their corresponding Z_{AU}^{j} and Z_R^k. We used *Adam optimizer* with $\beta_1 = 0.9, \beta_2 = 0.98$ and $\epsilon = 10^9$. We used a *Learning Rate Scheduler* as in Vaswani et al. [2017], with $W_{steps} = 4000$ warmup steps. We applied a dropout $Drop$ equals to 0.1 to the output of each sub-layer of the transformer, and to the sums of the positional encodings in the Transformer encoder and decoder stacks. All features values were normalized between 0 and 1. The total number of the model's parameters is 2051133.

The training hyperparameters are summarized in Table 5.7.

Hyperparameter		Value
Batch Size	BS	128
$Adam\ Optimizer$	β_1	0.9
	β_2	0.98
	ϵ	10^9
Learning Rate Scheduler	W_{steps}	4000
	$Drop$	0.1

Table 5.7 Training Hyperparameters

5.6.3 Objective Evaluation

To assess the quality of the generated gestures, we used the same objective measures used for evaluating our baseline: ***RMSE***, ***PCC***, ***AHR*** and ***NAHR***. We also evaluate ***Model 2*** using the ***SD*** test set. We additionally evaluated its capacity to generalize on the ***SI*** set in order to assess the extent to which gestures predictions can be extrapolated to unseen speakers.

To evaluate the different parts of our architecture, we conduct an ablation study as follows:

1. Speech ablation,
2. Text ablation,
3. A ablation,
4. AU/R decoder ablation

The ablation study is evaluated using $RMSE$ metric. We also compare the results of the objective studies of **Model 1** and **Model 2**.

Objective Evaluations Results

	Model 2 (Transformer based)				Model 1 (Baseline)			
	RMSE	**PCC**	**AHR**	**NAHR**	**RMSE**	**PCC**	**AHR**	**NAHR**
AU1	0.193	0.81	92.1	109.0	0.20	-0.012	42.56	115.95
AU2	0.142	0.92	100.0	100.0	0.48	-0.002	34.15	107.08
AU4	0.171	0.78	91.2	120.0	0.53	-0.012	60.31	121.85
AU5	0.199	0.67	101.1	100.0	0.50	-0.011	21.12	135.95
AU6	0.219	0.89	102.3	100.2	0.48	-0.002	33.11	135.05
AU7	0.14	0.88	98.1	99.5	0.33	-0.053	20.21	131.52
$\mathbf{R}_X$	0.29	0.62	*NA*	*NA*	0.53	0.018	*NA*	*NA*
$\mathbf{R}_Y$	0.24	0.81	*NA*	*NA*	0.97	-0.024	*NA*	*NA*
$\mathbf{R}_Z$	0.32	0.74	*NA*	*NA*	0.22	0.003	*NA*	*NA*

Table 5.8 Objective Evaluation: comparison of proposed transformer model vs. baseline lstm-based model

Table 5.8 reports the model's as well as the **Baseline**'s objective evaluation results using the same ***SD*** test set. Results reveal that RMSE errors are much smaller for **M** ($0.14 \leq error \leq 0.32$) than **Baseline** ($0.2046 \leq error \leq 0.9786$). On the other hand, PCC coefficients show that **M**'s predictions ($0.62 \leq PCC \leq 0.92$) are more correlated than **Baseline**'s predictions ($-0.0525 \leq PCC \leq 0.0179$) to **GT**. AHR and NAHR were calculated to measure the activation of AUs only, since they are not applicable for R. Results show that **M** predicts better the activation rate AHR ($AHR \geq 91.2$) than **Baseline** ($20.211 \leq AHR \leq 60.314$). The non-activation rate is higher for **Baseline** ($107.084 \leq NAHR \leq 135.95$) than for **M**($99.5 \leq NAHR \leq 120$). This constitutes objective validation that **M** gives better results than **Baseline** in terms of *error*, *correlation*, and AU's *activation rate*.

AU/R Decoder ablation resulted in even higher RMSE errors especially for head rotations. AUs and R RMSE scores increased after A ablation (i.e. AU1 RMSE increased to 0.202). This constitutes an objective validation that the use of multi-modal inputs (speech and text modalities) in **M** improves predictions. Thus we can also conclude that A module is an efficient and a key component of our model, as it improves the generation accuracy of face gestures and head rotations. As mentioned previously, we also tested our model on the ***SI*** set. RMSE errors are between 0.301 and 0.89 for AUs, and between 0.25 and 0.93 for R. As we could expect, we got higher errors than the errors we had for the ***SD*** condition (Table 5.8) since the speakers in ***SI*** set were not seen by our model during the training phase.

5.6.4 Subjective Evaluation

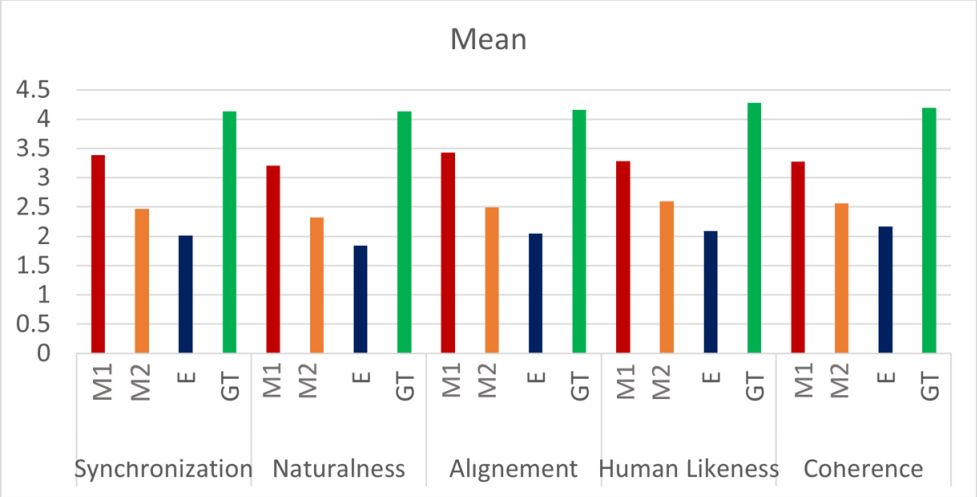
To investigate human perception of the facial gestures produced by our model, we conducted two different experimental studies using the virtual agent Greta (Pelachaud [2017]). We followed the recommendations proposed in Wolfert et al. [2022], by adapting them to facial gesture generation and assessed the *naturalness*, *coherence*, and *human-likeness* of the virtual agent's gestures. Since we are not evaluating deictic and iconic gestures, we did not use the metrics *appropriateness*, and *intelligibility* as proposed in Wolfert et al. [2022]. Those two metrics focus on the shape of the gestures that we do not model explicitly. We added the metrics *synchronization*, and *alignment* to evaluate the gestures' temporal property with speech. Participants in both studies were fluent in English, with a University degree, and recruited on Prolific, a crowd sourcing website. We added attention checks at the beginning of our perceptual evaluations, to filter out inattentive participants.

The first study was done by 35 participants, and consisted of presenting 16 videos: each video showed the virtual agent saying a sequence of words that corresponds to a sequence of IPUs. We considered 4 conditions: 4 videos (condition **M2**) used our full *Model 2* of ***SD*** gestures predictions; 4 videos (condition **GT**) were simulated using the gestures extracted from TED videos, which serve as ground truth; 4 videos of the virtual agent were simulated using *Model 1* which is the LSTM-based Baseline model of ***SD*** predictions (condition **M1**). The remaining 4 videos were produced using predicted gesture animation of IPUs with the sound of other IPUs (condition **E**). The latter condition serves as a *control condition*.

The second study was conducted by 55 participants. The goal of the second study was to evaluate our model when simulated with **SI** data, and therefore its capability to generalise

to new speakers. It included 8 videos: 4 were simulated with our model's **SI** predictions, and 4 using **SI** gestures extracted from SI set, which serve as ground truth. For each video in both studies, participants were asked to rate the 5 factors, namely *naturalness*, *coherence*, *human-likeness*, *synchronization*, and *alignment* of the virtual agent's gestures on a 1 to 7 likert scale (Wolfert et al. [2022]). The questions were listed in a random order. The agent's mouth movements were blurred to prevent participants from getting distracted by these gestures which were not inferred by our model, and therefore focus on the model's generated gestures.

Subjective Evaluation Results and Discussion



.

Figure 5.6 Subjective evaluation results obtained on the $SpeakerDependent$ (SD) set to assess the *naturalness*, *human-likeness*, and *coherence* of the predicted gestures, as well as the *synchronization* and *alignment* of the gestures with the speech. The assessment was conducted for the 4 conditions: *Model 2* denoted by ***M2***, *Model 1* baseline denoted by ***M1***, the ground truth ***GT***, and the error condition ***E***

For our first perceptive study conducted on ***SD*** data, Figure 5.6 shows the mean scores obtained on the 5 factors for the 4 conditions: *Model 2* (***M2***), *Model 1* which is the baseline (***M1***), the ground truth (**GT**), and the error (**E**). Participants perceived our model ***M2*** as being close to the ground truth **GT** for all conditions *Coherence*, *Human-Likeness*, *Naturalness*, *Alignment*, and *Synchronization*. ***M1*** is perceived less closer to the ***GT*** than ***M2*** for all the factors. The error condition ***E***, which serves as a *control condition*, is perceived as the farthest to ***GT*** amongst all conditions which implies that both of our models ***M2*** and ***B*** perform better than the error condition in terms of the 5 factors. Moreover, for all the factors, ***M2*** is perceived even better than ***M1*** as it is the closest to ***GT*** and the farthest to ***E***. More specifically, for the 5 factors, **M2** is perceived as much closer to **GT** than **Baseline** and **E**, especially in terms of *Alignment* and *Synchronization* between speech and gestures. The mean difference between **M2** and **GT** is 0.72 for *Alignment* and 0.74 for *Synchronization* (Figure 5.6).

We performed a post-hoc *Fisher's LSD* test to do pair-wise comparisons of the means between the factors of all conditions. Significant results ($\boldsymbol{p}<0.007$) were found when comparing ***M2*** and ***M1*** which means that ***M2*** is significantly higher than ***M1*** for all the factors. In addition to that, significant results ($\boldsymbol{p}<0.001$) were also found between ***M1*** and ***E***, which means that our baseline ***M1*** performs significantly better than the error condition ***E***. For the pairs (***GT***, ***M2***), as well as (***GT***, ***M1***). Fisher's LSD test also resulted in $\boldsymbol{p}<0.001$ which means that the ***GT*** is perceived significantly higher than ***M2*** and ***M1*** in terms of all the factors. This result was expected, as the ***GT*** condition is the simulation of the raw speakers' gesturing data. This constitutes experimental validation that when used with ***SD*** data, condition **M2** is perceived significantly closer to the ***GT*** than **Baseline** and **E** for all the factors, and that ***M1*** performs significantly better than the error condition ***E***.

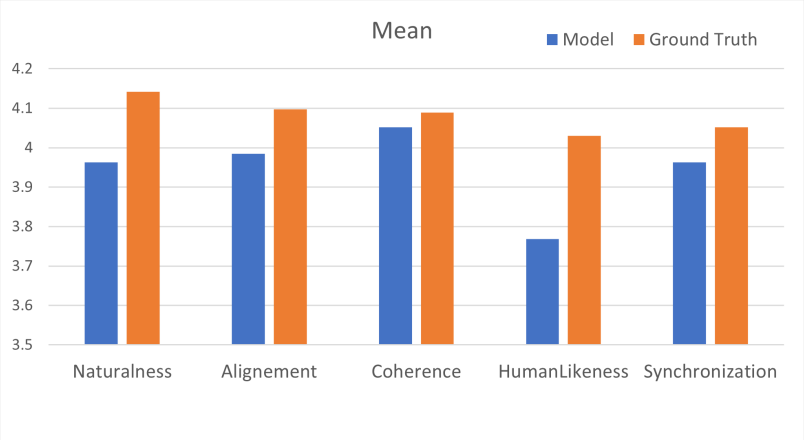


Figure 5.7 Subjective evaluation results obtained on the *Speaker Independent (SI)* set to assess the *naturalness*, *coherence*, and *human-likeness* of **M2**'s predicted gestures, as well as the *alignment* and *synchronization* of the gestures with speech and its content. These factors were evaluated for both conditions **M2** and **GT**.

For our second perceptive study conducted on ***SI*** data set, Figure 5.7 shows the means scores obtained for the factors *Naturalness*, *Coherence*, *Human-Likeness*, *Alignment* and *Synchronization* for conditions **M2** (*Model 2*) and **GT**. This second evaluation aims to measure **M2**'s generalization performance on new speakers that were not seen by **M2** during training, on a 5-point likert scale. For the 5 factors, the **GT** condition received a score above 4. **M2**'s *coherence* is the closest to the **GT**, as the mean difference between **M2** and **GT** is equal to 0.04 for the *Coherence* factor. The mean difference between **GT** and **M2** for the *Synchronization* and *Alignment* factors is equal to 0.07 and 0.11, respectively. Hence, when testing with unseen speakers, the coherence of **M2**'s produced gestures, as well as their alignment and synchronization with speech and its content are very close to **GT**'s coherence, alignment and synchronization. Our model **M2** captures well the temporal relation between speech and gestures.

The mean difference between **M2** and **GT** for the factors *Naturalness* and *Human-Likeness* is greater than the mean difference of the other factors. More specifically, the mean difference for the *Naturalness* factor is 0.16; the one for *Human-likeness* is equal to 0.24. Thus, **M2**'s synthesized gestures of unseen speakers are not as natural and human-like as the **GT**'s gestures. These two factors are linked to the quality of the resulting animations,

while the other three factors are linked to the temporal relationship with speech. While our model captures well the temporal relationship, the quality of the resulting animations needs to be improved. This could be done by adding smooth filters at the output of our model. Moreover, as **M2** was only trained on a subset of the *TEDx Corpus*, the generalization could be leveraged by training it on more speakers' data to capture more variability in multimodal behaviors.

Even though **M2** was trained on only 200 speaker's data, the coherence of the synthesized gestures as well as their alignment, and synchronization with speech are very close to the ground truth, as shown by this second perceptive study, while the quality of the resulting animations needs further improvements.

5.7 Conclusion

In this Chapter, we focus on addressing the research questions Q1 (*multimodality synthesis*), and Q2 (*generalization*) discussed earlier (Chapter 1), and developing an approach to synthesize coherent human-like facial expressivity in ECAs. In particular, we focus on developing an approach that predicts *expressive facial movements* such as *eyebrows*, *eyelids*, and *head* movements. The synthesized gestures are based on different modalities, specifically *speech* data, *text* data, and *facial* data. To the best of our knowledge, synthesizing speech-driven and semantically-aware facial gestures was never investigated at the time of this research. As a starting point, we proposed **Model 1**, an end-to-end *sequence to sequence LSTM neural network architecture* that predicts *upper-face gestures* and *head motion*, based on both *speech prosody* and *text semantics*. The objective evaluation results conducted on **Model 1** revealed the presence of high RMSE errors for some features, and the absence of correlation between predictions are ground truth. The AUs activations are low, and the non-activations are high. To overcome the weaknesses of **Model 1**, we proposed a second novel approach that makes use of *Transformers* and *Convolutions* to synthesize the upper-face and head gestures, based on *speech audio*, and *text semantics*. We compare this approach to the *baseline* model, and results showed that **Model 2** surpasses the baseline in terms of RMSE errors, PCC, AUs activation and AUs non-activation. The subjective evaluation study conducted on the ***Speaker Dependent*** data set showed that when testing with seen speakers, **Model 2**'s synthesized gestures are perceived significantly closer to the ground truth than the baseline and the error condition in terms of *naturalness*, *coherence*, *human-likeness* of the gestures, as well as the *synchronization* and *alignment* with speech and its content. Hence, we experimentally validated that ***Model 2*** performs well with seen speakers in terms of the previously mentioned factors. Moreover, the results of the second perceptive study which was conducted on the ***Speaker Independent*** data set constitute an experimental validation that **Model 2** can generalize its performance to new unseen speakers, especially in terms of *coherence* of the gestures, and their *alignment* and *synchronization* with speech. These results allow us to answer the research questions Q1 and Q2. More specifically, we demonstrated that **Model 2**, the transformer-based computational model can synthesize speech-driven and semantically-aware facial gestures for speakers that were seen during training. We demonstrated that it can also generalize the learned gestural space to new speakers that were unseen during the training phase of our generative model. Our approach exploits human multimodal behavior - speech prosody, visual prosody and language - to synthesize expressive body visual prosody in ECAs.

The key points of this Chapter:

Addressing Research Questions Q1, and Q2

- Research question Q1 - How can we exploit human multimodal behavior - speech prosody, visual prosody, and language - to generate expressive and human-like *facial* and *body* visual prosody in Embodied Conversational Agents? How to design computational models that can capture the relationship between these different modalities?
- Research question Q2 - How can we generalize the learned gestural latent space to new speakers data, that are unseen during the training phase of our generative model?

Model 1

- As a starting point, we propose **Model 1**, an end-to-end *sequence to sequence LSTM neural network architecture* that predicts *upper-face gestures* and *head motion*, based on both *speech prosody* and *text semantics*.
- The objective evaluation results showed the presence of high RMSE errors for some features, and the absence of correlation between predictions are ground truth. The AUs activations are low, and the non-activations are high.

Model 2

- To overcome the weaknesses of **Model 1**, we proposed a second novel approach that makes use of *Transformers* and *Convolutions* to synthesize the upper-face and head gestures, based on *speech audio*, *text semantics*.
- Objective evaluation showed that **Model 2** surpasses the baseline in terms of RMSE errors, PCC, AUs activation and AUs non-activation.
- We experimentally validate that **Model 2** performs well with seen speakers. More specifically, the first perceptive study showed that when testing with seen speakers, **Model 2**'s synthesized gestures are perceived significantly closer to the ground truth than the baseline and the error condition in terms of *naturalness*, *coherence*, *human-likeness* of the gestures, as well as their *synchronization* and *alignment* with speech.
- **Model 2** can generalize its performance to new unseen speakers, especially in terms of ***coherence*** of the gestures, and the *alignment* and *synchronization* of the speech and its content.

Part III

Gesturing with Style: Modelling Behavioral Style for Gesture Synthesis

Chapter 6

Style Transfer for Upper-Body Gesture Animation using Adversarial Disentanglement of Multimodal Style Encoding

Contents

6.1 Introduction

Modeling virtual agents with *behavior style* is one factor for personalizing human-agent interaction. In this Chapter, we propose an efficient yet effective machine learning approach to synthesize gestures driven by *prosodic features* and *text* in the style of different speakers including those unseen during training. Our model performs *zero-shot multimodal style transfer* driven by multimodal data from the *PATS Corpus* (Chapter 4) containing videos of various speakers.

We view *behavioral style* as being pervasive while speaking; it colors the communicative behaviors expressivity while speech content is carried by multimodal signals and text. This disentanglement scheme of *content* and *style* allows us to directly infer the style embedding even of speakers whose data are not part of the training phase, without requiring any further training or fine-tuning.

The first goal of our model is to generate the gestures of a *source speaker* based on the *content* of two input modalities – Mel spectrogram and text semantics. The second goal is to condition the *source speaker*'s predicted behavior expressivity on the multimodal behavior *style* embedding of a *target speaker*. The third goal is to allow zero-shot style transfer of speakers unseen during training without re-training the model.

Our system consists of two main components:

1. A *speaker style encoder network* that learns to generate a fixed-dimensional speaker embedding *style* from a target speaker multimodal data (mel-spectrogram, pose, and text)

2. A *sequence-to-sequence synthesis network* that synthesizes gestures based on the *content* of the input modalities - text and mel-spectrogram - of a source speaker, and conditioned on the speaker style embedding.

We evaluate that our model is able to synthesize gestures of a *source speaker* given the two input modalities, and transfer the knowledge of *target speaker* style variability learned by the speaker style encoder to the gesture generation task in a zero-shot setup, indicating that the model has learned a high quality speaker representation.

We conduct objective and subjective evaluations to validate our approach and compare it with the baseline *Mix-StAGE*.

This Chapter is organized as follows. We first discuss how we view *behavioral style* and the key challenges for modeling it. We then review the existing *behavioral style* modeling approaches, and discuss their limitations. Then, we explain our approach, followed by the details of the architecture we propose and a description of the training regime we follow for training the model. We then conduct objective and subjective evaluations and discuss the results.

6.2 Context

Behavioral style is multifaceted. We discuss in this Section how we view *behavioral style* and the key technical challenges for modeling it.

Human behavior. Human behavior involves *verbal* and *non-verbal* behavior. Non-verbal behavior includes *speech style*, which refers to the prosodic features that are determined by biological, physiological, and sociocultural factors. Some prosodic characteristics are always present in the human voice, and include *timbre*, *resonance*, *loudness*, *tempo*, *pitch* and *intonation*. *Speech style* is also determined by how specific sounds (such as whisper or breathy) are produced, and these sounds are most of the time associated with various styles such as dominance, competence, and confidence. Moreover, *non-verbal* behavior includes movements and gesturing, which are *person-specific* and *idiosyncratic* in nature. In other words, each speaker has his or her own *behavioral style*. *Behavioral style* is the way speakers express themselves, using their facial expressions, body language, gestures, tone of voice, and non-verbal cues.

Human behavioral style. As previously discussed in Chapter 2 Section 2.5, human behavior style is *multimodal*, it is a socially meaningful clustering of features found within and across multiple modalities, specifically in linguistic (Campbell-Kibler et al. [2006]), spoken behavior such as the speaking style conveyed by speech prosody (Moon et al. [2022], Obin [2011]), and nonverbal behavior such as hand gestures and body posture (Obermeier et al. [2015], Wagner et al. [2014]). Behavioral style is specifically related to the multimodal behaviors and their expressivity specific to each speaker (Pelachaud [2009], Bergmann and Kopp [2009b]). Speakers gesture differently and there is a large *variability* in gesturing, due to speaker's *personality traits*, *verbal* skills, *age*, and *culture*, etc.

Technical challenges - modeling behavior style. modeling behavioral style for ECAs constitutes a stimulating technical challenge. The behavior generation model should not simply learn an overall style from multiple speakers, but should remember each speaker's specific behavioral style generated in a specific lexical content context and behavior expressivity. The model should be able to capture the behavioral style that are common throughout speakers, as well as the ones that are unique to a speaker's prototypical gestures produced consciously and unconsciously.

Technical challenges - modeling non-verbal behavior. As discussed in the previous Chapters, verbal and non-verbal behavior plays a crucial role in communication in human-human interaction (Norris [2004]). Generative models that aim to predict communicative gestures of ECAs must produce meaningful and naturalistic gestures that are aligned with speech (Cassell [2000]). Non-verbal behavior must be generated and synchronized in conjunction with verbal and prosodic behavior to define their shape and time of occurrence (Salem et al. [2011]). This constitutes another technical challenge, to enable a smooth and engaging interaction between humans and ECAs by making sure that ECAs produce semantically-aware, natural, expressive and coherent gestures aligned with speech and its content.

ECAs' multimodal behavioral style. Our work considers the presence of style in a speaker's multimodal behavior, encompassing verbal and non-verbal human behavior styles.

As previously discussed, verbal, gestural, and prosodic features determine the speaker's *behavioral style*. The context of the speech is reflected by verbal cues, and it is linked to the speaker's behavior style, as it determines the situation he or she is in. In our work, to synthesize an ECA's gestures in the style of *target speakers*, we consider *behavioral style* as being conveyed through multimodal behavior.

6.3 State of the Art - Existing Behavioral Style Modeling Approaches

Beyond realistic generation of human non-verbal behavior, several early works have explored the relationship between personality and body motion in virtual characters, enabling the synthesis of stylized virtual character animations. Durupinar et al. [2016] established a formal association between personality and body motion, utilizing Laban Movement Analysis and user studies to animate expressive characters with personality. Smith and Neff [2017] focused on the impact of gesture edits on perceived character personality, identifying dimensions of plasticity and stability. Brand and Hertzmann [2000] introduced stylistic motion synthesis, learning motion patterns from diverse sequences to generate virtual motion-capture in various styles. Hartmann et al. [2006] presented a computational model of gesture quality, emphasizing modifications to convey desired expressive content while preserving original semantics.

Other related works have focused on modeling and controlling style in gesture to create more expressive behaviors that can be tailored to specific audiences (Neff et al. [2008], Karras et al. [2017], Cudeiro et al. [2019], Ahuja et al. [2020b], Ginosar et al. [2019], Alexanderson et al. [2020], Ahuja et al. [2022]). Neff et al. [2008] propose a system that produces full body gesture animation driven by text, in the style of a specific performer. They focus on hand movements, and certain movements were avoided since they require deeper model of semantics. Alexanderson et al. [2020] propose a generative model for synthesizing speech-driven gesticulation, they exert directorial control over the output style such as gesture level and speed. Karras et al. [2017] propose a model for driving 3D facial animation from audio. Their main objective is to model the style of a single actor by using a deep neural network that outputs 3D vertex positions of meshes that correspond to a specific audio. Cudeiro et al. [2019] also propose a model that synthesizes 3D facial animation driven by speech signal. The learned model, VOCA (Voice Operated Character Animation) takes any speech signal as input–even speech in languages other than English–and realistically animates a wide range of adult faces. Conditioning on subject labels during training allows the model to learn a variety of realistic speaking styles. VOCA also provides animator controls to alter speaking style, identity-dependent facial shape, and pose (i.e. head, jaw, and eyeball rotations) during animation. Ginosar et al. [2019] propose an approach for generating gestures given audio speech, their approach uses models trained on *single* speakers.

The aforementioned works have focused on generating nonverbal behaviors (facial expression, head movement, gestures in particular) that are either aligned with speech or text. They have not considered *multimodal data* when modeling style, as well as when synthesizing gestures. Moreover, their generative models are trained on *single speaker* data.

To our knowledge, the only attempts to model and transfer the style from multi-speakers database have been proposed by Ahuja et al. [2020b] and Ahuja et al. [2022]. Ahuja et al. [2020b] presented Mix-StAGE, a speech driven approach that trains a model from multiple speakers while learning a unique style embedding for each speaker. They created PATS, a dataset designed to study various styles of gestures for a large number of speakers in diverse settings. In their proposed neural architecture, a *content* and a *style* encoder are used to extract content and style information from speech and pose. To disentangle style from content information, they assume that style is only encoded through the pose modality, and the content is shared across speech and pose modalities. A style embedding matrix whose each vector represents the style associated to a specific speaker from the training set. During training, they further propose a multimodal GAN strategy to generate poses either from the speech or pose modality. During inference, the pose is inferred by only using the speech modality and the desired style token.

However, their generative model is conditioned on gesture style and driven by audio. It does not include verbal information. It cannot perform zero-shot style transfer on speakers that were not seen by their model during training. In addition, the style is associated with each unique speaker, which makes the distinction unclear between each speaker's specific style - idiosyncrasy -, the style that is shared among a set of speakers of similar settings (i.e. TV show hosts, journalists, etc...), and the style that is unique to each speaker's prototype gestures that are produced consciously and unconsciously.

Moreover, the style transfer is limited to the styles of *PATS speakers*, which prevents the transfer of style from an unseen speaker. Furthermore, the proposed architecture is based on the disentangling of content and PATS style information, which is based on the assumption that style is only encoded by gestures. However, both text and speech also convey style information, and the encoding of style must take into account all the modalities of human behavior. To tackle those issues, Ahuja et al. [2022] presented a *few-shot* style transfer strategy based on neural domain adaptation accounting for cross-modal grounding shift between source speaker and target style. This adaptation still requires 2 minutes of the style to be transferred.

To the best of our knowledge, our approach is the first to synthesize gestures from a source speaker, which are semantically-aware, speech driven and conditioned on a multimodal representation of the style of target speakers, in a zero-shot configuration i.e., without requiring any further training or fine-tuning.

6.4 Our Approach

We propose a novel approach to model *behavioral style* in ECAs and tackle the different challenges. Our approach aims at:

1. Synthesizing natural and expressive upper body gestures of a source speaker, by encoding the *content* of two input modalities – text semantics and Mel spectrogram,

2. Conditioning the *source speaker*'s predicted gesture on the multimodal *style* representation of a *target speaker*, and therefore rendering the model able to perform style transfer across speakers,

3. Allowing zero-shot style transfer of newly coming speakers that were not seen by the model during training.

We trained our model on the database *PATS* (Chapter 4), which was proposed in Ahuja et al. [2020b] and designed to study gesture generation and style transfer. It includes 3 main modalities that we are considering in our approach: text semantics represented by *BERT embeddings*, *Mel spectrogram* and *2D upper body poses*.

We propose the first approach for zero-shot multimodal style transfer approach for 2D pose synthesis. At inference, an embedding style vector can be directly inferred from multimodal data (text, speech and poses) of any speaker, by simple projection into the embedding style space (similar to the one used in Jia et al. [2018]). The style transfer performed by our model allows the transfer of style from any "unseen" speakers, without further training or fine-tuning of our trained model. The model learns a style space based on the speakers observed during training, but it can generalize to new speakers who were not present in the training data. This latent space captures the distribution of style vectors, enabling effective representation of the behavioral styles of the speakers encountered during training. During inference, the style encoder takes as input the target speaker's multimodal data and generates a behavioral style vector specific to that speaker. Remarkably, even for speakers not encountered during training, the model is still capable of generating high-quality behavioral style vectors. This means that the model is not constrained to the styles of speakers in a specific database. Moreover, it facilitates "style preservation" by generating gestures for multiple speakers while retaining their unique characteristics.

To design our approach, we make the following assumptions for the separation of style and content information:

1. *Style* is possibly encoded across all modalities (text, speech, pose) and varies little over time and in some cases does not;

2. *Content* is encoded only by text and speech modalities and varies over time.

To implement theses assumptions, we propose an architecture for encoding and disentangling *content* and *style* information from multiple modalities. On one side, a content encoder is used to encode a content matrix from text and speech signal; on the other hand, a style encoder is used to encode a style vector from all text, speech, and body pose modalities. A fader loss is introduced to effectively disentangle content and style encodings (Lample et al. [2017]). The encoding of the style takes into account 3 features: body poses, text semantics, and speech - Mel spectrograms. These features are important to generate behaviors (Kucherenko et al. [2019], Ginosar et al. [2019]) and are linked to style.

We evaluate the generated behaviors by conducting objective and subjective evaluations.

6.5 Zero-Shot Multimodal Style Transfer Model 1.0 (ZS-MSTM 1.0) for Gesture Animation driven by Text and Speech

We propose **ZS-MSTM 1.0** (**Z**ero-**S**hot **M**ultimodal **S**tyle **T**ransfer **M**odel), a Transformer-based architecture for stylized upper-body gesture synthesis, driven by the content of a

source speaker's speech - text semantics represented by BERT embeddings and audio Mel spectrogram -, and conditioned on a target speaker's multimodal style embedding. The stylized generated gestures correspond to the style of target speakers that have been seen and unseen during training.

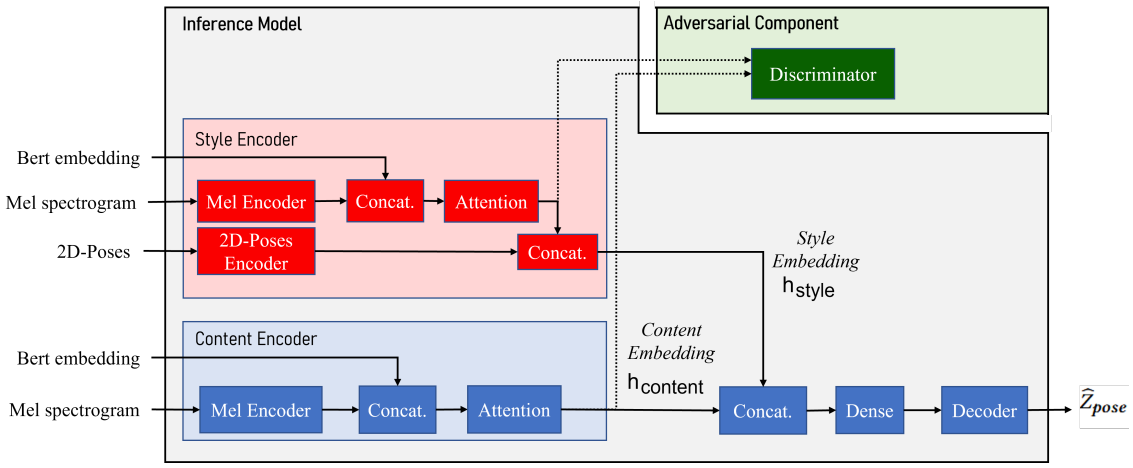


Figure 6.1 **ZS-MSTM 1.0** (**Z**ero-**S**hot **M**ultimodal **S**tyle **T**ransfer **M**odel) architecture. The content encoder (further referred to as $E_{content}$) is used to encode content embedding $h_{content}$ from BERT text embeddings X_{text} and speech Mel-spectrograms X_{speech} using a speech encoder $E^{content}_{speech}$. The style encoder (further referred to as E_{style}) is used to encode style embedding h_{style} from multimodal text X_{text}, speech X_{speech}, and pose X_{pose} using speech encoder E^{style}_{speech} and pose encoder E^{style}_{pose}. The generator G is a transformer network that generates the sequence of poses $\widehat{Z}_{pose}$ from the sequence of content embedding $h_{content}$ and the style embedding vector h_{style}. The adversarial module relying on the discriminator Dis is used to disentangle content and style embeddings $h_{content}$ and h_{style}.

As depicted in Figure 6.1, the system is composed of three main components:

1. A **speaker style encoder** network that learns to generate a fixed-dimensional speaker embedding style from a *target speaker* multimodal data: 2D poses, BERT embeddings, and Mel spectrogram, all extracted from videos in a database.

2. A **sequence to sequence gesture synthesis** network that synthesizes upper-body behavior (including hand gestures and body poses) based on the content of two input modalities - text embeddings and Mel spectrogram - of a *source speaker*, and conditioned on the *target speaker* style embedding. A *content encoder* is presented to encode the content of the Mel spectrogram along with BERT embeddings.

3. An **adversarial component** in the form of a fader network (Lample et al. [2017]) is used for disentangling style and content from the multimodal data.

At inference time, the adversarial component is discarded, and the model can generate different versions of poses when fed with different style embeddings. Gesture styles for the same input speech can be directly controlled by switching the value of the style embedding vector h_{style} or by calculating this embedding from a target speaker's multimodal data fed as input to the *Style Encoder*.

ZS-MSTM 1.0 illustrated in Fig. 6.1 aims at mapping multimodal speech and text feature

sequences into continuous upper-body gestures, conditioned on a speaker style embedding. The network operates on a segment-level of 64 timesteps (duration of 4.26 seconds): the inputs and output of the network consist of one feature vector for each segment ***S*** of the input text sequence. The length of the segment-level input features (text and audio) corresponds to $t = 64$ timesteps (as provided by *PATS Corpus*). The model generates a sequence of gestures corresponding to the same segment-level features given as inputs. Gestures are sequences of 2D poses represented by *x* and *y* positions of the joints of the skeleton. The network has an embedding dimension d_{model} equal to 768.

6.5.1 Content Encoder

The content encoder $E_{content}$ illustrated in Figure 6.1 takes as inputs BERT embedding X_{text} and audio Mel spectrograms X_{speech} corresponding to each **S**. X_{text} is represented by a vector of length 768 - BERT embedding size used in *PATS Corpus*. X_{speech} is encoded using *Mel Spectrogram Transformer (AST)* pre-trained *base384* model (Gong et al. [2021]).

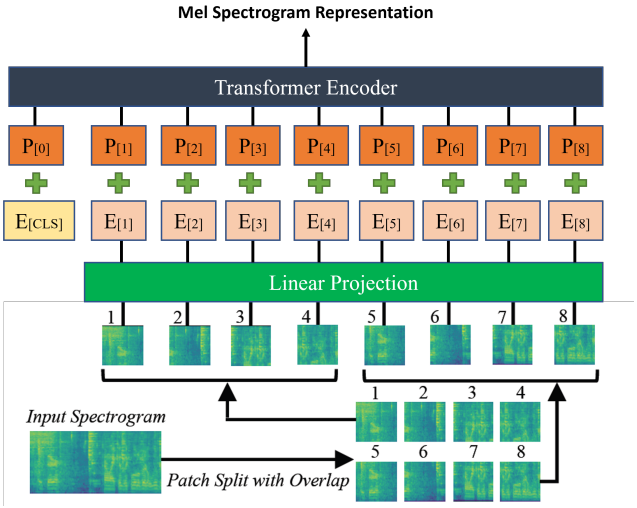


Figure 6.2 AST Architecture

AST operates as follows: the input Mel spectrogram which has 128 frequency bins, is split into a sequence of 16x16 patches with overlap, and then is linearly projected into a sequence of 1D patch vectors, which is added with a positional embedding. We append a [*CLS*] token to the resulting sequence, which is then input to a *Transformer Encoder*. *AST* was originally proposed for audio classification. Since we do not intend to use it for a classification task, we remove the linear layer with sigmoid activation function at the output of the *Transformer Encoder*. We use the *Transformer Encoder*'s output of the [*CLS*] token as the Mel spectrogram representation **S**. The *Transformer Encoder* has an embedding dimension equals to d_{model}, N_{enc} equals to 12 encoding layers, and N_h equals to 12 attention heads.

The segment-level encoded Mel spectrogram is then concatenated with the segment-level BERT embedding. A self-attention mechanism is then applied on the resulting vector. The multi-head attention layer has N_h equals to 4 attention heads, and an embedding size d_{att} equals to $d_{att} = d_{model} + 768$. The output of the attention layer is the vector $h_{content}$, a content representation of the source speaker's segment-level Mel spectrogram and text

embedding, and it can be written as follows:

$$h_{content} = sa\left(\left[E^{content}_{speech}(X_{speech}), X_{text}\right]\right) \tag{6.1}$$

where: sa(.) denotes self-attention.

6.5.2 Style Encoder

As discussed previously, *behavior style* is a clustering of features found within and across modalities, encompassing verbal and non-verbal behavior. It is not limited to gestural information. We consider that *behavior style* is encoded in a speaker's multimodal - text, speech and pose - behavior. As illustrated in Figure 6.1, the style encoder E_{style} takes as input, at the segment-level, Mel spectrogram X_{speech}, BERT embedding X_{text}, and a sequence of (X, Y) joints positions that correspond to a target speaker's 2D poses X_{pose}. *AST* is used to encode the audio input spectrogram. N_{lay} equals to 3 layers of LSTMs with a hidden-size equal to d_{model} are used to encode the vector representing the 2D poses. The last hidden layer is then concatenated with the audio representation. Next, a multi-head attention mechanism is applied on the resulting vector. This attention layer has N_h equals to 4 attention heads and an embedding size equals to d_{att}. Finally, the output vector is concatenated with the 2D poses vector representation. The resulting vector h_{style} is the output speaker style embedding that serves to condition the network with the speaker style. The final style embedding h_{style} can therefore be written as follows:

$$h_{style} = \left[sa\left(\left[X_{text}, E^{style}_{speech}(X_{speech})\right]\right), E^{style}_{pose}(X_{pose})\right] \tag{6.2}$$

where: sa(.) denotes self-attention.

6.5.3 Sequence to sequence gesture synthesis

The stylized 2D poses are generated given the sequence of content representation $h_{content}$ of the source speaker's Mel spectrogram and text embeddings obtained at S-level, and conditioned by the style vector embedding h_{style} generated from a target speaker's multimodal data. For decoding the stylized 2D-poses, the sequence of $h_{content}$ and the vector h_{style} are concatenated (by repeating the h_{style} vector for each segment of the sequence), and passed through a $Dense$ layer of size d_{model}. We then give the resulting vector as input to a *Transformer Decoder*. The *Transformer Decoder* is composed of $N_{dec} = 1$ decoding layer, with $N_h = 2$ attention heads, and an embedding size equal to d_{model}. Similar to the one proposed in Vaswani et al. [2017], it is composed of residual connections applied around each of the sub-layers, followed by layer normalization. Moreover, the self-attention sub-layer in the decoder stack is altered to prevent positions from attending to subsequent positions. The output predictions are offset by one position. This masking makes sure that the predictions for position index ***j*** depends only on the known outputs at positions that are less than ***j***. For the last step, we perform a permutation of the first and the second dimensions of the vector generated by the transformer decoder. The resulting vector is a sequence of 2D-poses which corresponds to:

$$\widehat{Z}_{pose} = G(h_{content}, h_{style}) \tag{6.3}$$

where: G is the transformer generator conditioned on latent content embedding $h_{content}$ and style embedding h_{style} . The generator loss of the transformer gesture synthesis can

be written as:

$$\mathcal{L}_{rec}^{gen}(E_{content}, E_{style}, G) = \mathbb{E}_{\widehat{Z}_{pose}}||\widehat{Z}_{pose} - G(h_{content}, h_{style})||_2 \tag{6.4}$$

6.5.4 Adversarial Component

Our approach of disentangling style from content relies on the fader network disentangling approach (Lample et al. [2017]), where a fader loss is introduced to effectively separate content and style encodings. The fundamental feature of our disentangling scheme is to constrain the latent space of $h_{content}$ to be independent of the style embeddings h_{style}. Concretely, it means that the distribution over $h_{content}$ of the latent representations should not contain the style information. A fader network is composed of: an encoder which encodes the input information X into the latent code $h_{content}$, a decoder which decodes the original data from the latent, and an additional variable h_{style} used to condition the decoder with the desired information (a face attribute in the original paper). The objective of the fader network is to learn a latent encoding $h_{content}$ of the input data that is independent on the conditioning variable h_{style} while both variables are complementary to reconstruct the original input data from the latent variable $h_{content}$ and the conditioning variable h_{style}. To do so, a discriminator *Dis* is optimized to predict the variable h_{style} from the latent code $h_{content}$; on the other side the auto-encoder is optimized using an additional adversarial loss so that the classifier *Dis* is unable to predict the variable h_{style}. Contrary to the original fader network in which the conditional variable is discrete within a finite binary set (0 or 1 for the presence or absence attribute), in this paper the conditional variable h_{style} is continuous. We then formulate this discriminator as a regression on the conditional variable h_{style}: the discriminator learns to predict the style embedding h_{style} from the content embedding $h_{content}$, as:

$$\widehat{h}_{style} = Dis(h_{content}) \tag{6.5}$$

While optimizing the discriminator, the discriminator loss $\mathcal{L}^{dis}$ must be as low as possible, such as:

$$\mathcal{L}^{dis}(D) = \mathbb{E}_{h_{style}}||h_{style} - Dis(h_{content})||_2 \tag{6.6}$$

In turn, optimizing the generator loss including the fader loss $\mathcal{L}_{adv}^{gen}$, the discriminator must not be able to predict correctly the style embedding h_{style} from the content embedding $h_{content}$ conducting to a high discriminator error and thus a low fader loss. The adversarial loss can be written as,

$$\mathcal{L}_{adv}^{gen}(E_{content}, E_{style}) = \mathbb{E}_{h_{style}}||1 - (h_{style} - Dis(h_{content}))||_2 \tag{6.7}$$

To be consistent, the style prediction error is preliminary normalized within 0 and 1 range.

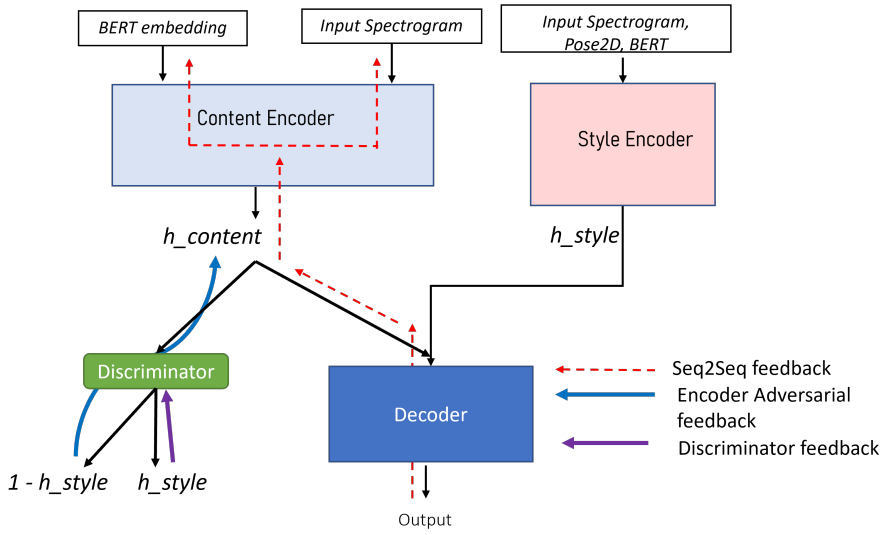


Figure 6.3 Fader network for multimodal content and style disentangling.

Finally, the total generator loss can therefore be written as follows:

$$\mathcal{L}^{gen}_{total}(E_{content}, E_{style}, G) = \mathcal{L}^{gen}_{rec}(E_{content}, E_{style}, G) + \lambda \mathcal{L}^{gen}_{adv}(E_{content}, E_{style}, G) \quad (6.8)$$

where λ is the adversarial weight that starts off at 0 and is linearly incremented by 0.01 after each training step.

The discriminator Dis and the generator G are then optimized alternatively as described in Lample et al. [2017].

All **ZS-MSTM 1.0** hyperparameters were chosen empirically and are summarized in Table 6.1.

Component	Hyperparameter		Value
AST (base384 model)	Embedding size	d_{model}	768
	Encoding layers	N_{lay}	12
	Attention heads	N_h	12
Content Encoder	Attention heads	N_h	4
	Embedding size	d_{att}	1536
Style Encoder	2D Pose LSTMs	N_{lay}	3
		N_{hid}	768
	Attention heads	N_h	4
	Embedding size	d_{att}	1536
Sequence to Sequence Component	Transformer Decoder	N_{dec}	1
	Attention heads	N_h	2
	Embedding size	d_{model}	768

Table 6.1 ZS-MSTM 1.0 hyperparameters

6.6 Training

This section describes the training regime we follow for training **ZS-MSTM 1.0**. We trained our network using the *PATS Corpus* (Ahuja et al. [2020b]) which was previously discussed in Chapter 4. PATS was created to study various styles of gestures. The dataset contains upper-body 2D pose sequences aligned with corresponding Mel spectrogram, and BERT embeddings. Each PATS speaker is characterized by their lexical diversity and the spatial extend of their arms. While in PATS arms and fingers have been extracted, we do not consider finger data in our work. That is we do not model and predict 2D finger joints. This choice arises as the analysis of finger data is very noisy and not very accurate. We model 11 joints that represent upper body and arm joints.

We consider two test conditions: *Seen Speaker* and *Unseen Speaker*. The *Seen Speaker* condition aims to assess the style transfer correctness that our model can achieve when presented with speakers that were seen during training as target style. On the other hand, the *Unseen Speaker* condition aims to assess the performance of our model when presented with unseen target speakers, to perform zero-shot style transfer. Seen and unseen speakers are specifically selected from PATS to cover a diversity of stylistic behavior with respect to lexical diversity and spatial extent as reported by Ahuja et al. [2020b][1].

For each PATS speaker, there is a train, validation and test set already defined in the database. For testing the *Seen Speaker* condition, our test set includes the test sets of 16 PATS speakers. Six other speakers are selected for the *Unseen Speaker* condition, and their test sets are also used for our experiments. These six speakers differ in their behavior style and lexical diversity. *Seen* and *Unseen* speakers are listed in Table 6.2.

Condition	Speakers
Seen	"Shelly", "Jon", "Fallon", "Bee", "Ellen", "Oliver", "Lec_cosmic", "Lec_hist", "Ytch_prof", "Ytch_dating", "Seth", "Conan", "Angelica", "Rock", "Noah", and "Lec_law"
Unseen	"Lec_evol", "Almaram", "Huckabee", "Ytch_charisma", "Minhaj", and "Chemistry"

Table 6.2 Seen and Unseen PATS Speakers

Each training batch contains $BS = 24$ pairs of word embeddings, Mel spectrogram, and their corresponding sequence of (X, Y) joints of the skeleton (of the upper-body pose). We use Adam optimizer with $\beta_1 = 0.95, \beta_2 = 0.999$. For balanced learning, we use a scheduler with an initial learning rate Lr equals to 1e-5, with W_{steps} equals to $20,000$. We train the network for $N_{ep} = 200$. All features values are normalized so that the dataset mean and standard deviation are 0 and 0.5, respectively. Table 6.3 summarizes all hyperparameters used for training.

[1] https://chahuja.com/pats/

Hyperparameter		Value
Batch Size	BS	24
Number of epochs	N_{ep}	200
Adam Optimizer	β_1	0.95
	β_2	0.999
Scheduler	W_{steps}	20,000
	Lr	1e-5

Table 6.3 Training Hyperparameters

6.7 Objective Evaluation

To validate our approach and assess the stylized generated gestures, we conduct an objective evaluation for the two conditions *Seen Speakers* and *Unseen Speakers*.

6.7.1 Objective Metrics

In our work, we have defined *behavioral style* by the *behavior expressivity* of a speaker. To evaluate objectively our works, we define metrics to compare the *behavior expressivity* generated by our model, with the target speaker's *behavior expressivity*, and source speaker's *behavior expressivity*.

Following works on *behavior expressivity* by Wallbott [1998] and Pelachaud [2009], we define 4 objective *behavior dynamics* metrics to evaluate the style transfer of different target speakers: *acceleration*, *jerk* and *velocity* that are averaged over the values of all upper-body joints, as well as the speaker's average *bounding box perimeter* (BB perimeter) of his/her body movements extension. In addition, we compute the *acceleration*, *jerk* and *velocity* of only the *left* and *right wrists*, to obtain information on the *arms movements expressivity* (Wallbott [1998], Kucherenko et al. [2019]).

For both conditions SD and SI, we define two sets of distances:

1. ***Dist.(****Source*, *Target*)**: representing the average distance between the source style and the target style,
2. ***Dist.(****ZS-MSTM 1.0*, *Target*)**: representing the average distance between our model's gestures style and the target style.

More specifically, after computing the *behavior expressivity* and *BB perimeter* of our model's generated gestures, the ones of source speakers, and the ones of the target speakers, we calculate the average distance as follows:

$$\textbf{Dist}_{avg}\textit{(x, Target)} = \frac{\textbf{Dist.}\textit{(x, Target)}}{\textbf{Dist.}\textit{(Source, Target)} + \textbf{Dist.}\textit{(ZS-MSTM 1.0, Target)}} \times 100 \quad (6.9)$$

Where x denotes *Source* for computing $\textbf{Dist}_{avg}(Source, Target)$ and *ZS-MSTM 1.0* for computing $\textbf{Dist}_{avg}(\textit{ZS-MSTM}, Target)$. The reason we use relative distance metrics is to allow for comparisons withing the corresponding conditions and provide a meaningful assessment of the relationships and variations between them.

6.7.2 Objective Evaluation Results

Objective evaluation experiments are conducted for evaluating the performance of our model in the *Seen Speaker* and *Unseen Speaker* conditions. For *Seen Speaker* condition, experiments are conducted on the test set that includes the 16 speakers that are seen by our model during training. For *Unseen Speaker* condition, experiments are also conducted on another test set that includes the 6 speakers that were not seen during training.

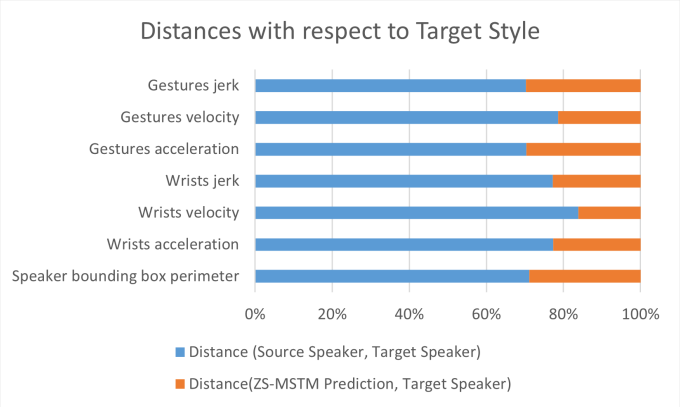


Figure 6.4 Distances between the target speaker style and each of the source style and our model's generated gestures style for seen target speakers

Figure 6.4 reports the experimental results on the *Seen Speaker* test set. It illustrates the results of ***Dist.(****Source*, *Target***)** in terms of *behaviors dynamics* and speaker *bounding box perimeter* between the target speaker style and the source speaker style.

For *Seen Speaker* condition (Figure 6.4), ***Dist.(****Source*, *Target***)** is higher than 70% of the total distance for all behavior dynamics metrics; thus ***Dist.(****ZS-MSTM 1.0*, *Target***)** is less than 30% of the total distance for all behavior dynamics metrics. Wrists velocity, jerk and acceleration results reveal that the virtual agent's arms movements show the same expressivity dynamics as the target style (***Dist.(****ZS-MSTM 1.0*, *Target***)** $<$ 22%).The style transfer from target speaker "Shelly" to source speaker "Angelica" - knowing that Angelica is a *Seen Speaker* - shows that the distance of predicted gestures' behavior dynamics metrics are close (distance $<$ 20%) to "Shelly" (*target style*), while the ones between "Angelica" and "Shelly" are far (distance $>$ 80%). The perimeter of the prediction's bounding box (BB) is closer (distance $<$ 30 %) to the target speaker's BB perimeter than the source . The closeness between predictions dynamics behavior metrics values are shown for all speakers in the *Seen Speaker* condition, specifically for the following style transfers - *target* to *source* - : "Fallon" to "Shelly", "Bee" to "Shelly", "Conan" to "Angelica", "Oliver" to "lec_cosmic", which are considered having different lexical diversity, as well as spatial average extent, as reported by the authors of PATS (Ahuja et al. [2020b]).

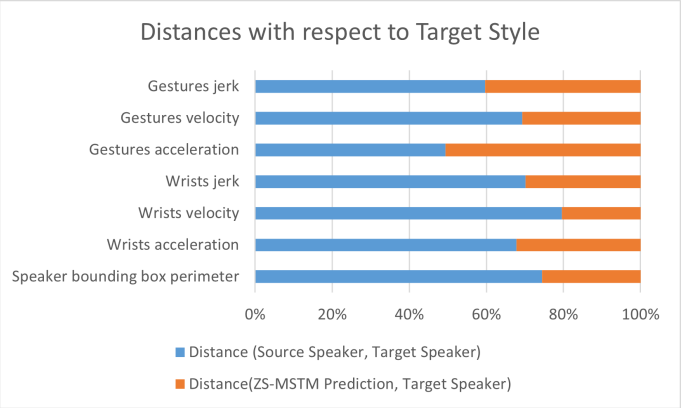


Figure 6.5 Distances between the target speaker style and each of the source style and our model's generated gestures style for unseen target speakers

Experimental results for the *Unseen Speaker* test set are depicted in Fig. 6.5. Results reveal that our model is capable of reproducing the style of the 6 unseen speakers. As depicted in Fig. 6.5, for all behavior dynamics metrics, as well as the bounding box perimeter, ***Dist.(****Source*, *Target***)** is higher than 50% of the total distances for all metrics. Results show that for wrists velocity, jerk and acceleration, ***Dist.(****ZS-MSTM 1.0*, *Target***)** is less than 33%. Thus, arm movement's expressivity produced by *ZS-MSTM 1.0* is close to the one of the target speaker style. Moreover, the perimeter of the prediction's bounding box is close (distance < 30 %) to the target speaker's, while the distance between the BB perimeter of the source and the target is far (distance > 70 %). While our model has not seen "Lec_evol"'s multimodal data during training, it is yet capable of transferring his behavior expressivity style to the source speaker "Oliver". It is also capable of performing zero-shot style transfer from the target speaker "Minhaj" to the source speaker "Conan". In fact, results show that wrists acceleration and jerk values of our model's generated gestures are very close to those of the target speaker "Minhaj". We observe the same results for the 6 speakers for the *Unseen Speaker* condition.

We additionally conduct a Fisher's LSD Test to do pair-wise comparisons on all metrics, for the two set of distances - ***Dist.(****Source*, *Target***)**, and ***Dist.(****ZS-MSTM 1.0*, *Target***)** - in both conditions. We find significant results ($p < 0.003$) for all distances in both conditions.

6.7.3 Additional t-SNE Analysis

In this work, the style encoder is agnostic: it is the attention weights that make it possible to exploit the different modalities given as input to the style encoder.

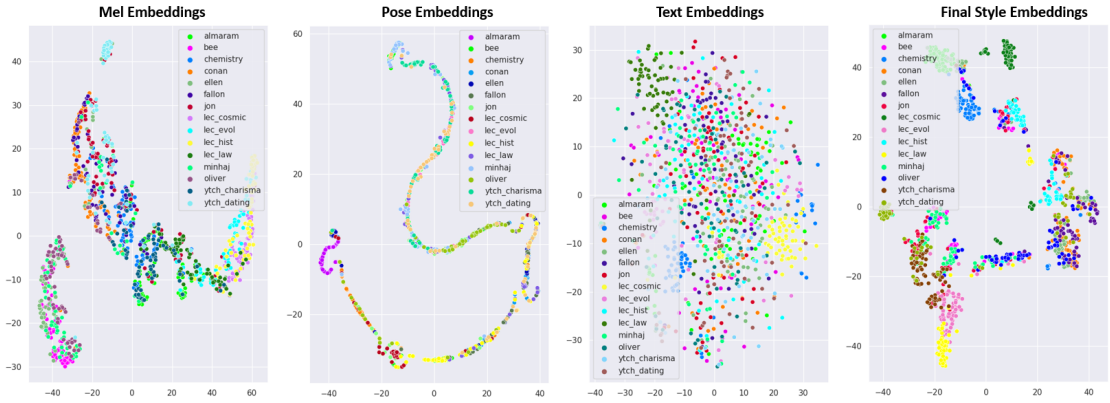


Figure 6.6 2D TSNE Analysis of the generated ***Mel Embeddings***, ***Pose Embeddings***, ***Text Embeddings***, and the final ***Style Embeddings***

In our study, we performed a post-hoc analysis using *t-SNE* (t-distributed Stochastic Neighbor Embedding) to visualize the distributions of vectors from different modalities. Specifically, we examined the *Mel Embeddings*, *Pose Embeddings*, *Text Embeddings*, and the final *Style Embeddings* generated by our ***ZS-MSTM 1.0*** model. The results of this analysis are illustrated in Figure 6.6, which displays 2D t-SNE plots. Based on our findings, we observed that the behavioral style exhibited the highest dependence on the pose modality. The pose modality captures the body posture, movements, and gestures, which strongly influence the overall behavioral style representation. Following the pose modality, we found that the speech modality contributed significantly to the motion style, as we can see some clustering in the 2D plot. This indicates that the Mel spectrogram have an impact on the encoded style vectors. The speech modality captures the acoustic properties and prosodic aspects of the spoken content, which can convey certain behavioral style cues. Lastly, we observed that the text semantics, represented by the Text Embeddings, had a relatively lower influence on the behavioral style compared to the pose and speech modalities. This suggests that while the textual content contributes to the overall style representation, it has a lesser impact on the motion style compared to the body pose and speech characteristics.

6.8 Human Perceptual Studies

We conduct three human perceptual studies.

1. **Study 1** - To investigate human perception of the stylized upper-body gestures produced by our model, we conduct a human perceptual study that aims to assess the style transfer of speakers *seen* during training - *Seen Speaker* condition.

2. **Study 2** - We conduct another human perceptual study that aims to assess the style transfer of speakers *unseen* during training - *Unseen Speaker* condition.

3. **Study 3** - We additionally conduct a third human perceptual study to compare **ZS-MSTM 1.0**'s produced stylized gestures in *Seen Speaker* and *Unseen Speaker* conditions, to *Mix-StAGE* (Ahuja et al. [2020b]) which we consider our baseline.

The evaluation studies are conducted with 35 participants that were recruited through the online crowd-sourcing website Prolific. Participants are selected such that they are fluent in English. Attention checks are added in the beginning and the middle of each study to filter out inattentive participants. All the animations presented in these studies are in the form of 2D stick figures.

Study 1 and 2. For Study 1 and 2, we presented 60 stimuli of 2D stick animations. Each study included 30 stimuli. A stimulus is a triplet of 2D animations composed of:

- A 2D animation with the *source style*,
- A 2D animation with the *target style*,
- A 2D animation of *ZS-MSTM 1.0*'s prediction after performing the style transfer.

Figure 6.7 illustrates the three animations we present for each set of questions. The animation of the target style is the ***Reference***. The animation of our model's predictions, and the source style is either ***Animation A*** or ***Animation B*** (randomly chosen).

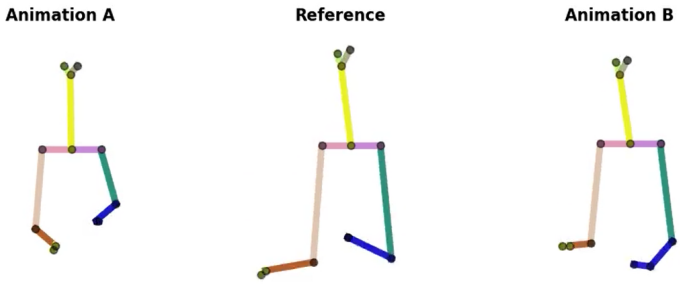


Figure 6.7 Three 2D stick animations: *Animation A*, the *Reference*, and *Animation B*. The target style is represented by *Reference*. **ZS-MSTM 1.0**'s predictions, and the **source style** are illustrated in Animation A or B.

For each triplet of animations, we asked 6 questions to evaluate 6 factors related to the *resemblance* of the produced gestures w.r.t the the *source style* and *target style*:

1. Please rate **the overall resemblance of the Reference** w.r.t A and B. (**Factor 1** - Overall resemblance)
2. Please rate the **resemblance of the Left (L) and Right (R) arms gesturing of the Reference** w.r.t the left and right arm gesturing of A and B. (**Factor 2** - Arms gesturing)
3. Please rate the **resemblance of the body orientation of the Reference** w.r.t the body orientation of A and B. (**Factor 3** - Body orientation)
4. Please rate the **resemblance of the gesture amplitude of the Reference** w.r.t the gesture amplitude of A and B. (**Factor 4** - Gesture amplitude)
5. Please rate the **resemblance of the gesture frequency of the Reference** w.r.t the gesture frequency of A and B. (**Factor 5** - Gesture frequency)
6. Please rate the **resemblance of the gesture velocity of Reference** w.r.t the gesture velocity of A and B. (**Factor 6** - Gesture velocity)

Each factor is rated on a *5 likert* scale, as follows:

1. Reference is very similar to A
2. Reference is mostly similar to A
3. Reference is in between A and B
4. Reference is mostly similar to B
5. Reference is very similar to B

Training. Each study includes a training at its beginning. The training provides an overview of the 2D upper-body skeleton of the virtual agent, its composition, and gesturing. The goal of the training is to get the participants familiarized with the 2D skeleton before starting the study. More specifically, the training included a description of how the motion of a speaker in a video is extracted by detecting his/her facial and body motion and extracting his/her 2D skeleton of joints, and stated that in a similar fashion, the eyes and upper-body movement of a virtual agent are represented by a 2D skeleton of joints, as depicted in Figure 6.8

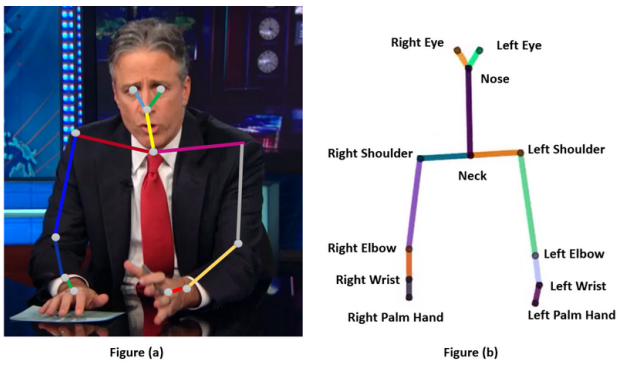


Figure 6.8 Upper-body 2D skeleton of a speaker Vs. a virtual agent

Moreover, we present and describe different shots of the 2D skeleton gesturing with *its right/left arms*, and with different *body orientation*, which is described as the orientation of the shoulders and neck.

Pre-tests. We conducted pre-tests to make sure that the 2D animations are comprehensible by participants, as well as the questions. Participants reported that the training, stimuli and questions are coherent and comprehensible, however each study was too long, as it lasted 30 minutes. For this reason, we divided each study to three, such that each study includes only 10 stimuli, and is conducted by different participants. Hence, 6 studies including a pre-training, and the evaluation of 10 stimuli were conducted by 35 participants that are different.

Study 3. For Study 3, we present 20 stimuli consisting of triplets of 2D stick animations. Similar to *Study 1* and *Study 2*, for each triplet, we present: *Animation A*, the *Reference*, and *Animation B*. The animation of the target style is the *Reference*. The animation of MixStAGE's predictions, and the source style is either Animation A or Animation B (randomly

chosen). We note that these stimuli include the same *source* and *target* styles that were used in *Study 1* and *Study 2*, and which were randomly chosen. Study 3 also included training at its beginning, which is the same as the one previously described.

6.8.1 Human Perceptual Studies Results

Study 1 - *Seen Speakers*.

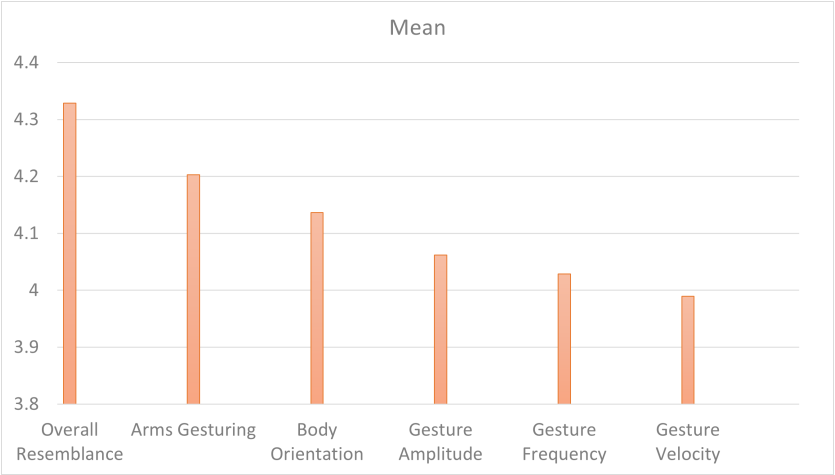


Figure 6.9 The mean scores of all the factors for *Seen Speakers* condition

Our first perceptive study (Study 1) aims to evaluate the style transfer of speakers *seen* during training. Figure 6.9 shows the mean scores obtained on the 6 factors for the condition "*seen speakers*". On a *5 likert scale*, the ***overall resemblance*** factor obtained a score of 4.32, which means that the **ZS-MSTM 1.0**'s 2D animations closely resemble the 2D animations of the *seen target style*. The resemblance is also reflected by the mean scores of ***arms gesturing***, ***body orientation***, ***gesture amplitude***, ***gesture frequency***, as well as ***gesture velocity***, which is between 3.99 and 4.2. We observed that for all factors, most of the participants gave a score between 3.8 and 5, as depicted in Figure 6.10.

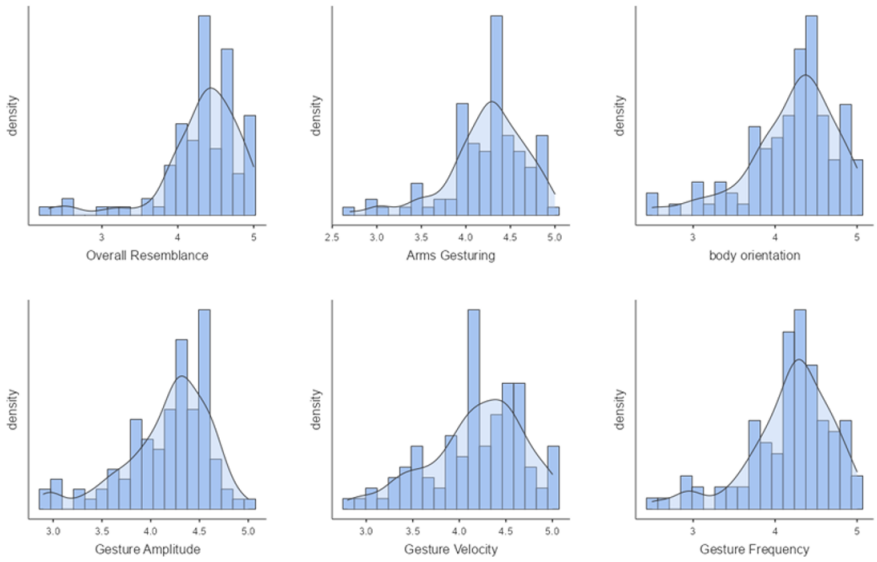


Figure 6.10 Density plots of ***Overall Resemblance***, ***Arms Gesturing***, ***Body Orientation***, ***Gesture Amplitude***, ***Gesture Frequency***, ***Gesture Velocity*** for the *Seen Speakers* condition

We additionally performed post-hoc paired samples t-tests between pairs of all the factors, and found significant results between ***overall resemblance*** and all the other factors ($p \leq 0.008$), which shows that each assessed factor contribute significantly to the *overall resemblance* of our model's predictions w.r.t. the target style.

Study 2 - *Unseen Speakers.*

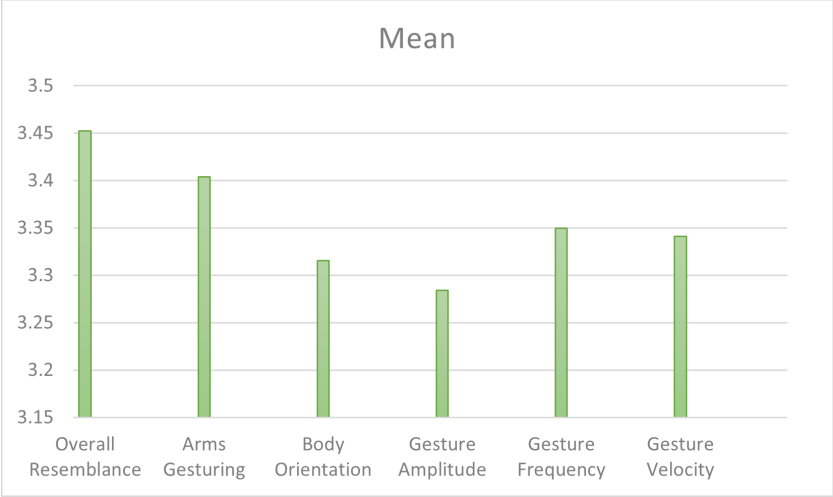


Figure 6.11 The mean scores of all the factors for *Unseen Speakers* condition

Our second perceptive study (Study 2) aims to evaluate the style transfer of speakers *unseen* during training. Figure 6.11 illustrates the mean scores obtained on the 6 factors for the condition "*unseen speakers*". On a *5 likert scale*, the ***overall resemblance*** factor obtained a score of 3.45, which means that there is an overall resemblance between **ZS-MSTM 1.0**'s 2D animations and the *unseen target style*. The resemblance is also reflected by the mean scores of ***arms gesturing***, ***body orientation***, ***gesture amplitude***, ***gesture***

frequency, as well as ***gesture velocity***, which is between 3.28 and 3.41. We observed that for all factors, most of the participants gave a score between 3 and 4, as depicted in Figure 6.12.

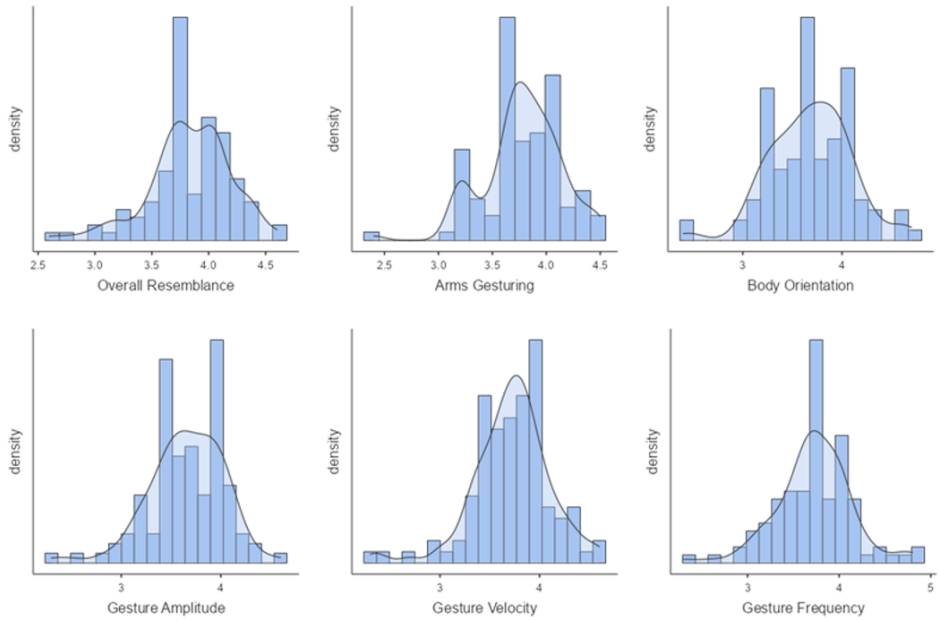


Figure 6.12 ***Body Orientation***, ***Gesture Amplitude***, ***Gesture Frequency***, ***Gesture Velocity*** for the *Unseen Speakers* condition

We additionally performed post-hoc paired samples t-tests between all the factors, and found significant results between ***overall resemblance*** and all the other factors ($p \leq 0.014$).

Study 3 - Comparing with Mix-StAGE. The third perceptive study aims to compare the performance of our model with respect to the State of the Art, *Mix-StAGE*. Figure 6.13 illustrates the mean scores obtained for the two conditions *Mix-StAGE* and *ZS-MSTM 1.0*, w.r.t the 6 factors.

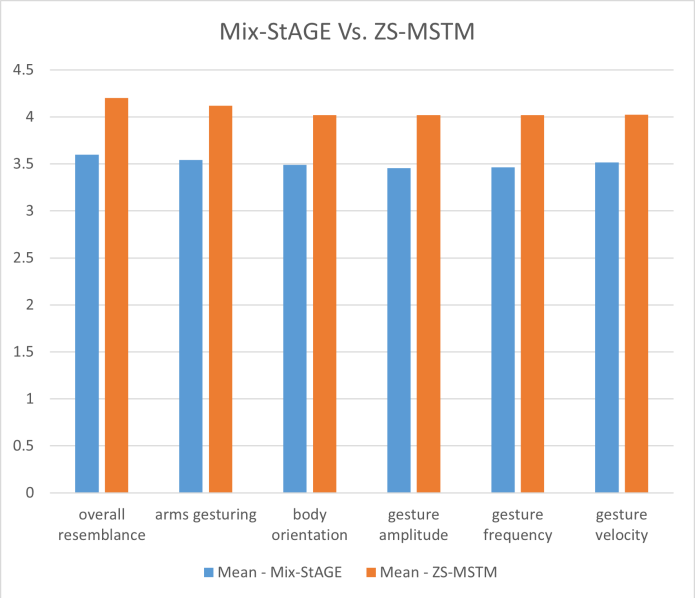


Figure 6.13 ***ZS-MSTM 1.0*** Vs. ***Mix-StAGE***

As shown in Figure 6.13, for all the factors, our model obtained higher mean scores than *Mix-StAGE*. Our model performs better than *Mix-StAGE* in terms of the ***overall resemblance*** of the generated gestures w.r.t the animations produced with the *target style* (mean score *ZS-MSTM 1.0* (4.2) $\geq$ mean score *Mix-StAGE* (3.6)). More specifically, the resemblance between the synthesized 2D gestures of *ZS-MSTM 1.0* and the target style is greater than the one between *Mix-StAGE* and the target style. This result is also reflected in the resemblance of the ***arms gesturing***, ***body orientation***, ***gesture amplitude***, ***gesture frequency*** and ***gesture velocity*** of our model's produced gestures w.r.t the *target style*. More specifically, our model obtained a mean score between 4 and 4.2 for all the factors, while *Mix-StAGE* obtained a mean score between 3.8 and 3.6 for all the factors. We additionally conducted post-hoc paired t-tests between the factors in condition *Mix-StAGE* and those in *ZS-MSTM 1.0*. We found significant results between all the factors in the condition *Mix-StAGE* and those in *ZS-MSTM 1.0* ($p < 0.001$ for all). These results show that the mean scores for all the factors in condition *ZS-MSTM 1.0* are significantly greater than those *Mix-StAGE*. Thus, we can conclude that our model *ZS-MSTM 1.0* can successfully render animations with the style of another speaker, going beyond the state of the art *Mix-StAGE*.

6.9 Conclusion

In this Chapter, we focus on addressing the research questions Q3 (*Multimodal style modeling.*), Q4 (*Style transfer.*) and Q5 (*Generalization*) discussed earlier in Chapter 1. More specifically, we have presented ***ZS-MSTM 1.0***, the first approach for zero-shot multimodal style transfer for 2D pose synthesis that allows the transfer of style from any speakers *seen* or *unseen* during the training phase. *Behavioral style* was never viewed as being *multimodal*; previous works limit behavior style to arm gestures only. However, both *text* and *speech* convey *style* information, and the embedding vector of *style* must consider the three modalities. Our assumption was confirmed by our post-hoc t-SNE analysis of the distributions of the style vectors at the output of each modality. We found that the motion style depends mainly on the body *pose modality*, followed by the *speech modality*, then the *text semantics modality*. We conducted an objective evaluation and three perceptive studies. The results of these studies show that our model produces stylized animations that are close to the target speakers style even for *unseen* speakers. ***ZS-MSTM 1.0*** can generalize style to new speakers without any fine-tuning or additional training, unlike *Mix-StAGE*. Its independence from the speaker's identity "ID" allows the generalization without being constrained and limited to the speakers used for training the model. DiffGAN was later on proposed by Ahuja et al. [2022] as an extension to *Mix-StAGE*, and an approach that performs *few-shot* style transfer strategy based on neural domain adaptation accounting for cross-modal grounding shift between source speaker and target style. However this adaptation still requires 2 minutes of the style to be transferred which is not required by our model. To sum up, we successfully answered the three research questions Q3, Q4, and Q5. We have presented ***ZS-MSTM 1.0*** an approach that can learn the style latent space of speakers, given their multimodal data, and independently from their identity. Our approach can synthesize body gestures of a source speaker, given the source speaker's mel spectrogram and text semantics, with the style of another target speaker given the target speaker's multimodal behavioral style that is encoded through the mel spectrogram, text semantics, and pose modalities. Moreover, our approach is *zero-shot*, thus is capable of transferring the style of unseen speakers, without the need of any additional training or fine-tuning.

The key points of this Chapter:

This Chapter addresses research questions Q3, Q4, and Q5:

- Q3 - How can we learn style latent space of given speakers, given their multimodal data, and independently from their identity?
- Q4 - How can we synthesize body gestures of a source speaker, given the source speaker multimodal data, but with the style of another speaker?
- Q5 - How can we render our approach able to perform zero-shot style transfer on new unseen speakers, without the need of any further training or fine-tuning?

ZS-MSTM 1.0

- We propose the first approach for zero-shot multimodal style transfer for 2D pose synthesis that allows the transfer of style from any speakers *seen* or *unseen* during the training phase.
- We consider *behavioral style* as being multimodal - present in pose, text and speech - unlike previous works which limit *behavioral style* to arm gestures only.
- We found that the motion style depends most on the *pose modality*, followed by the *speech modality*, then the *text semantics modality*, which was confirmed by the post-hoc T-SNE analysis.
- The results of the objective and subjective studies show that our model produces stylized animations that are close to the target speakers style even for *unseen* speakers.
- Unlike DiffGAN (Ahuja et al. [2022]) and *Mix-StAGE* (Ahuja et al. [2020b]), ***ZS-MSTM 1.0*** is *zero-shot*, thus is capable of transfer the style of unseen speakers, without the need of any additional training or fine-tuning.
- ***ZS-MSTM 1.0*** is independent from the speaker's identity "ID", which allows the generalization without being constrained and limited to the speakers used for training the model.

Chapter 7

ZS-MSTM 2.0: Zero-Shot Style Transfer for Facial and Body Gesture Animation

Contents

7.1 Introduction

The human face is an important "*organ of emotion*". It provides important clues by reacting in fractions of a second, often unconsciously, revealing a person's opinion, attitude and thoughts. It displays a large panel of communicative information manifested by complex variations of movements.

In the previous Chapter (6), we assumed that **behavioral style** is possibly encoded across the *pose*, *speech* and *text* modalities of communication. We used these factors to generate upper-body movement conditioned on style. We validated our assumption by demonstrating that *behavioral style* depends most on the *pose modality*, followed by the *speech* then *text semantics*. In this Chapter we want to generate *upper-body movements* and *facial expressions* conditioned on style. We present *ZS-MSTM 2.0*, an extended version of *ZS-MSTM 1.0* that was presented in Chapter 6. The goal of *ZS-MSTM 2.0* is to jointly synthesize *2D upper-body gestures* and *2D facial landmarks* of a *source speaker*, in the style of **any** *target speaker*. Similar to *ZS-MSTM 1.0*, *ZS-MSTM 2.0* is a Transformer-based architecture driven by the *content of* source speaker's speech - text semantics represented by *BERT Embeddings* and audio Mel Spectrogram -, and conditioned on a target speaker's multimodal style embeddding. *ZS-MSTM 2.0* also performs *zero-shot* style transfer, as the *stylized* generated gestures correspond to the *style* of target speakers that have been *seen* or were never seen ("*unseen*") by the model during training. To encode *behavioral style*, on top of the inputs we used for *ZS-MSTM 1.0* which are *2D poses*, *Mel spectrogram* and *BERT embeddings*; we have added two other inputs: *dialog tags*, *2D facial landmarks* of target speakers. We train *ZS-MSTM 2.0* on the extended version of the *PATS Corpus*, which includes the additional features - *2D Facial Landmarks* and *Dialog Tags* - as described in Chapter 4. The reason we have added *Dialog Tags* is to capture further semantic information in addition to *BERT embeddings*. Moreover, studies on communicative gestures have shown the link between dialog acts and gestures (Calbris [2011]).

To our knowledge and at the time of this research, *ZS-MSTM 2.0* is the first approach that synthesizes both, upper-body and facial gestures for ECAs, from speech and text inputs, in the style of *seen* or *unseen* speakers, in a *zero-shot* fashion.

In the following sections, we first review the two main *behavioral style* features which we added in *ZS-MSTM 2.0* to encode *behavioral style*.We then explain our model architecture, conduct an objective evaluation to assess it and discuss our results.

7.2 Additional Behavioral Style Features

In this Chapter, we make the assumption that *behavioral style* is also encoded in *2D Facial Landmarks* movements and *Dialog Tags*. As previously discussed in Chapter 4, we extracted 70 *2D Facial Landmarks* and 38 *Dialog Tags* from the *PATS Corpus* videos, which were initially collected by Ahuja et al. [2020b].

2D Facial Landmarks. We model 15 *2D Facial Keypoints* including keypoints related to *eyes*, *eyebrows*, and the *contour* of the face, which are illustrated in Figure 7.1.

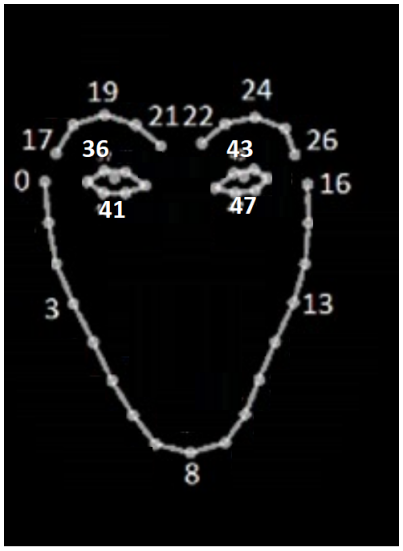


Figure 7.1 The 15 *2D Facial Keypoints* used for *ZS-MSTM 2.0* chosen amongst all the 70 facial landmarks extracted with OpenPose.

Dialog Tags. Dialog tags are tags that reflect additional text context information. An utterance is associated to one or more dialog acts. We consider all the 38 different tags that are listed in Table 4.3 in Chapter 4.

7.3 *ZS-MSTM 2.0* Architecture

ZS-MSTM 2.0 illustrated in Fig. 7.2 aims at mapping multimodal *speech* and *text* feature sequences into continuous *upper-body* gestures as well as *facial* gestures, conditioned on a speaker style embedding. The network operates on a segment-level $\boldsymbol{S}$ of 16 frames. In other words, the length of the segment-level $\boldsymbol{S}$ input and output features corresponds to $t = 16$ frames. We reduced the length of the segment to be given in input to our model from $t = 64$ frames (used in ***ZS-MSTM 1.0***), to $t = 16$ frames to allow us to generate smoother animations. The model produces a sequence of *facial* gestures and a sequence of *body* gestures corresponding to the same segment-level features given as inputs. Similar to *ZS-MSTM 1.0*, gestures are sequences of 2D poses represented by X and Y positions of the joints of the skeleton, and 2D facial gestures represented by X and Y positions of the facial landmarks. The network has an embedding dimension d_{model} equal to 64.

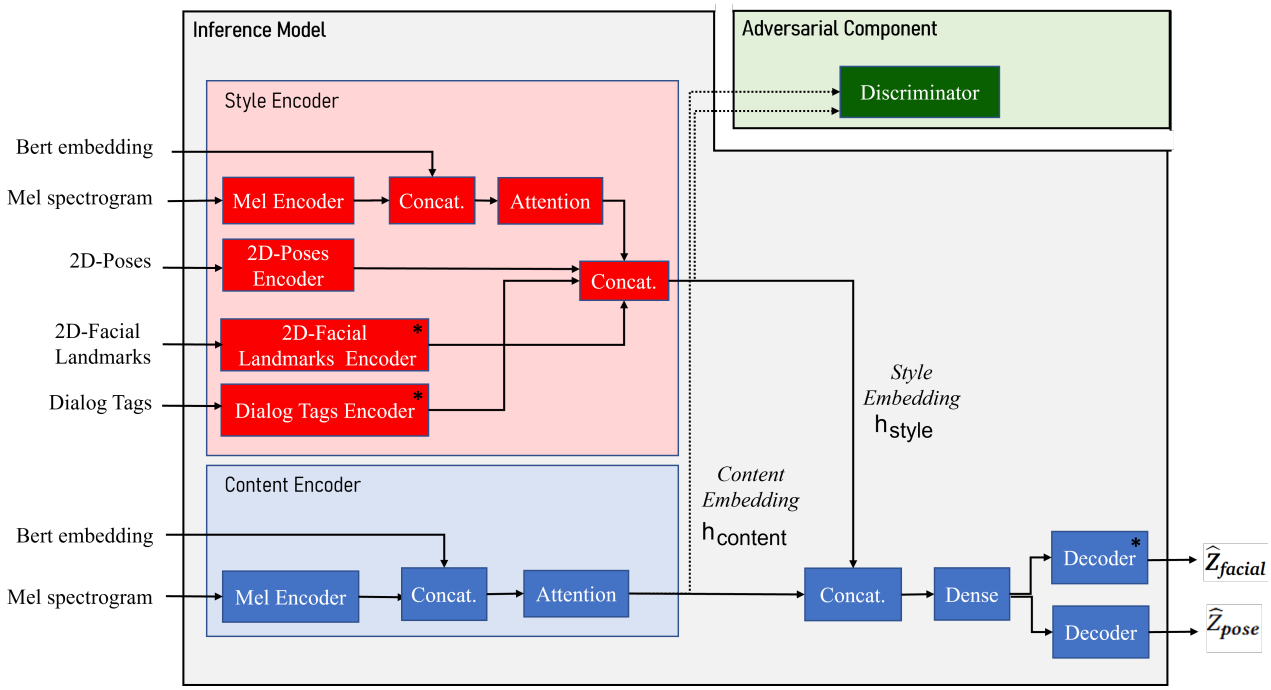


Figure 7.2 *ZS-MSTM 2.0* Overall Architecture. This architecture is similar to *ZS-MSTM 1.0* but with additional components for encoding and decoding *2D Facial Landmarks*, and encoding *Dialog Tags*. The newly added components are marked with an asterisk $(*)$.

Similar to *ZS-MSTM 1.0*, *ZS-MSTM 2.0* is composed of three main components:

1. A speaker **Style Encoder** E_{style} network that encodes target speaker's multimodal data and learns to produced a fixed-dimensional speaker embedding style from *Dialog Tags*, the 15 keypoints of *2D Facial Landmarks*, *2D Pose*, *BERT Embeddings* and *Mel Spectrogam*. The difference between E_{style} used in *ZS-MSTM 1.0* and the one used in *ZS-MSTM 2.0* is the inclusion of two additional main encoders in *ZS-MSTM 2.0*: *2D Facial Landmarks Encoder* and *Dialog Tags Encoder* which are referred to as E^{style}_{face} and E^{style}_{tags}, respectively.

2. A **sequence to sequence gesture synthesis** network that generates two sequences: (1) $\widehat{\mathbf{Z}}_{\boldsymbol{pose}}$, a sequence of 2D upper-body poses, and (2) $\widehat{\mathbf{Z}}_{\boldsymbol{face}}$, a sequence of 2D facial landmarks.

3. An **adversarial component** that functions similar to the one used in *ZS-MSTM 1.0*.

More specifically, the content encoder $E_{content}$ is used to encode content embedding $h_{content}$ from BERT text embeddings X_{text} and speech Mel-spectrograms X_{speech} using a speech encoder $E^{content}_{speech}$. The style encoder E_{style} is used to encode style embedding h_{style} from multimodal facial landmarks X_{face}, dialog tags X_{tags}, text X_{text}, speech X_{speech}, and pose X_{pose} using a 2D facial landmarks encoder E^{style}_{face}, dialog tags encoder E^{style}_{tags}, speech encoder E^{style}_{speech}, and 2D pose encoder E^{style}_{pose}. The generator G is a Transformer network that generates the sequence of poses $\widehat{Z}_{pose}$ and the sequence of facial landmarks $\widehat{Z}_{face}$ from the sequence of content embedding $h_{content}$ and the style embedding vector h_{style}. The adversarial module relying on the discriminator Dis is used to disentangle content and style embeddings $h_{content}$ and h_{style}.

7.3.1 2D Facial Landmarks Encoder

E^{style}_{face} takes as input X_{face}, which is a sequence of (x, y) joints positions that corresponds to a target speaker's 2D Facial Landmarks X_{face}. N_{lay} equals to 3 layers of LSTMs with a hidden-size equals to d_{model} are used to encode the vector representing the 2D Facial Landmarks. The last hidden layer is then concatenated with the remaining encoded modalities: X_{tags}, X_{text}, X_{speech}, and X_{pose}. The final style embedding h_{style} can therefore be written as follows:

$$h_{style} = \left[sa\left(\left[X_{text}, E^{style}_{speech}(X_{speech}) \right] \right), E^{style}_{pose}(X_{pose}), E^{style}_{face}(X_{face}), E^{style}_{tags}(X_{tags}) \right] \tag{7.1}$$

where: $sa(.)$ denotes self-attention.

7.3.2 Dialog Tags Encoder

E^{style}_{tags} is a *One Hot Encoder* that considers the 38 dialog tags as categorical features. The input features are encoded using a one-hot encoding scheme. The output is a sparse array containing binary values representing the presence or the absence of each tag in the segment S.

7.3.3 2D Facial Landmarks Decoder

The stylized 2D facial gestures are generated similar to the stylized 2D body gestures. First, $h_{content}$ - the content representation of the source speaker's Mel spectrogram and text embeddings obtained at S-level - is computed. $h_{content}$ is then conditioned on the style vector embedding h_{style} generated from a target speaker's multimodal data. More specifically, for decoding the stylized 2D-facial landmarks, the sequence of $h_{content}$ and the vector h_{style} are concatenated and given as input to a $Dense$ layer of size d_{model}. The resulting vector is then given as input to a *Transformer Decoder* which is composed of N_{dec} *= 1* decoding layer, with $N_h = 2$ attention heads, and an embedding size equals to d_{model}. The resulting final vector $\widehat{\boldsymbol{Z}}_{\boldsymbol{face}}$ is a sequence of 2D Facial Landmarks that corresponds to:

$$\widehat{Z}_{face} = G(h_{content}, h_{style}) \tag{7.2}$$

The loss used in ***ZS-MSTM 2.0*** is the same as the one used in ***ZS-MSTM 1.0*** by adding $\widehat{Z}_{face}$.

7.3.4 Hyperparameters

All ***ZS-MSTM 2.0*** hyperparameters were chosen empirically and are summarized in Table 7.1.

Component	Hyperparameter		Value
AST (base384 model)	Embedding size	d_{model}	64
	Encoding layers	N_{lay}	12
	Attention heads	N_h	12
Content Encoder	Attention heads	N_h	4
	Embedding size	d_{att}	832

Table 7.1 continued from previous page

Component	Hyperparameter		Value
Style Encoder - 2D Pose LSTMs	Encoding layers	N_{lay}	3
	Units	N_{hid}	64
Style Encoder - 2D Facial Landmarks LSTMs	Encoding layers	N_{lay}	3
	Units	N_{hid}	64
Style Encoder - Attention Layer	Attention heads	N_h	4
	Embedding size	d_{att}	832
Decoder - Pose	Decoding layers	N_{dec}	1
	Attention heads	N_h	2
	Embedding size	d_{model}	64
Decoder - Facial Landmarks	Decoding layers	N_{dec}	1
	Attention heads	N_h	2
	Embedding size	d_{model}	64

Table 7.1 ZS-MSTM 2.0 Hyperparameters

7.4 Training

We trained ***ZS-MSTM 2.0*** on the extension of the *PATS Corpus*, which includes 2D facial landmarks as well as dialog tags as discussed in Chapter 4. Similar to ***ZS-MSTM 1.0***, we do not model 2D finger joints since the extraction of finger data is very noisy and not accurate. For the facial landmarks, as previously stated, we only model the landmarks illustrated in Figure 7.1. We use less keypoints than those originally extracted. One reason is to have less input parameters and speed the training phase; another reason is to be aligned with the 3D MPEG-4 facial parameters used in our virtual agent platform, the Greta platform. We took out some keypoints from the face contour and 2 keypoints on each eyebrow, and we used only 2 keypoints for the eyelids. We can reconstruct the face from the remaining keypoints. In total, we model 11 body and arm joints, and 15 facial landmarks.

To evaluate ***ZS-MSTM 2.0***, we consider two test conditions:

1. The *Seen Speaker* condition which aims to assess the style transfer correctness that our model can achieve when presented with speakers that were *seen* during training as target style.

2. The *Unseen Speaker* condition which aims to assess the performance of our model when presented with *unseen* target speakers, to perform zero-shot style transfer.

We trained our model on the 16 speakers data that were used for training ***ZS-MSTM 1.0***. Our test set consists of the test sets of the same *Seen* and *Unseen* speakers that are listed in Table 6.2, which differ in their behavior style and lexical diversity.

Each training batch has $BS = 24$ pairs of word embeddings, Mel spectrogram, dialog acts, and their corresponding sequence of (x, y) joints of the skeleton of the upper-body pose and 2D facial landmarks. We use Adam optimizer with $\beta_1 = 0.95, \beta_2 = 0.999$, and a *Cyclical Learning Rate* (CLR) scheduler to render the learning balanced. The initial learning rate Lr_{init} of the CLR is equal to $1e-7$, the end learning rate Lr_e is equal to 0.1, and the step size St_{size} is equal to 196. We train the network for N_{it} equals to $78,400$ iterations.

All features values are normalized so that the dataset mean and standard deviation are 0 and 0.5, respectively. Table 7.2 summarizes all hyperparameters used for training.

Hyperparameter		Value
Batch Size	BS	24
Number of iterations	N_{it}	78,400
Adam Optimizer	β_1	0.95
	β_2	0.999
Cyclical Learning Rate Scheduler	Lr_{init}	1e-7
	St_{size}	196
	Lr_e	0.1

Table 7.2 Training Hyperparameters - *ZS-MSTM 2.0*

7.5 Objective Evaluation

We conduct an objective evaluation to assess ***ZS-MSTM 2.0*** in terms of *style transfer accuracy*, and *content preservation* for both conditions *Seen* and *Unseen*.

7.5.1 Metrics

We follow the recommendations of Fu et al. [2018] who propose two novel evaluation metrics to measure the characteristics of style transfer: *Transfer Strength Accuracy* and *Content Preservation*.

Transfer Strength Accuracy

Transfer Strength is a metric that assesses whether the style is transferred. As proposed by Fu et al. [2018], this metric is implemented using a classifier C. C performs a binary classification which helps to determine the style based on the output values generated by the model. More specifically, style is associated with a positive output if it is less than or equal to 0.5. In this case the style corresponds to the source style. Otherwise (if its greater than 0.5), it is associated with a negative output and it corresponds to the target style. We hence consider that *style* is defined as follows:

$$Style = \begin{cases} \textit{Source (positive) output} \leq 0.5 \\ \textit{Target (negative) output} > 0.5 \end{cases} \tag{7.3}$$

Transfer Strength Accuracy is defined as follows:

$$\textit{Transfer Strength Accuracy} = \frac{N_{right}}{N_{total}} \times 100 \tag{7.4}$$

where N_{right} is the number of correct cases which are transferred from target to source style, and N_{total} is total the number of test set data.

We train C on the train sets of the speakers that are in the test sets of SD and SI (the same train sets that are already defined in the *PATS Corpus*). C's overall architecture is depicted in Figure 7.3, which is more complex than the one used in Fu et al. [2018].

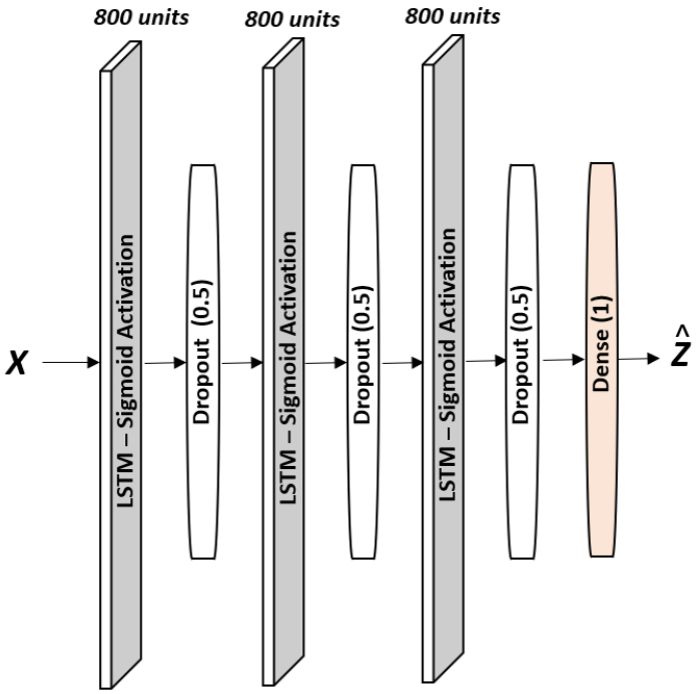


Figure 7.3 Classifier Architecture

The classifier C has an accuracy equals to *Accuracy* $= 96\%$. The hyperparameters of this network were chosen empirically and are summarized in Table 7.3.

Component	Hyperparameter	Value
3 LSTM layers	N_{hid}	800
3 Dropout layers	$Drop$	0.5
Dense layer	N_{hid}	1

Table 7.3 Classifier Hyperparameters

We trained C using a batch size BS equals to 256 for N_{ep} equals to $15,000$ epochs, using $Adam$ optimizer and *Binary Cross Entropy* loss (Table 7.4).

Hyperparameters		Value
Batch Size	BS	256
Number of epochs	N_{ep}	15,000
Optimizer	Adam	Default

Table 7.4 Classifier Training Hyperparameters

Content Preservation

Content Preservation is a metric that reflects the preservation of source content in predictions. It is defined as the cosine distance between predictions $\widehat{Z}_{gestures}$ and initial *source* gestures, as follows:

$$\textit{Cosine Distance} = \frac{X_{source}^{\intercal}\,\widehat{Z}_{gestures}}{\|X_{source}\| \cdot \left\|\widehat{Z}_{gestures}\right\|} \tag{7.5}$$

Distance Metrics

In addition to evaluating the *Transfer Strength Accuracy* and *Content Preservation*, we measure the *Minkowski distance* between the upper-body gestures and facial expressions produced by our model, and the ones of the *source* and *target* speakers. We tried with other distance metrics such as *cityblock*, *Chi2* distance, *euclidean* distance, and *cosine* distance. We kept only the *Minkowski distance* since all of them gave the same outcome.

More specifically, distances are calculated for both conditions *Seen* and *Unseen*, we define two sets of distances:

1. ***Dist.(****ZS-MSTM 2.0*, *Source*) - representing the average distance between all the body joints and facial landmarks generated by our model and those from the source data.
2. ***Dist.(****ZS-MSTM 2.0*, *Target*) - representing the average distance between all the body joints and facial landmarks generated by our model and those from the target data.

7.6 Objective Evaluation Results and Discussion

Condition	Transfer Strength Accuracy (%)	Content Preservation (%)
Seen	92.916	94.847
Unseen	84.583	91.017

Table 7.5 Style Transfer Evaluation Results

Condition	Minkowski ***Dist.(ZS-MSTM 2.0, Source)***	Minkowski ***Dist.(ZS-MSTM 2.0, Target)***
Seen	82.399	77.117
Unseen	81.545	75.597

Table 7.6 Minkowski Distances Results for both conditions *Seen* and *Unseen*

Objective evaluation results are presented in Table 7.5 and Table 7.6 for both *Seen* and *Unseen* conditions.

Transfer Strength Accuracy (%) and Content Preservation (%). For the *Seen* condition, we observe that ***ZS-MSTM 2.0*** has a style transfer strength accuracy equals to 92.9 % which means that ***ZS-MSTM 2.0*** transfers the style from target speakers to source speakers with a *high* accuracy. The accuracy is also high for the *Unseen* condition (84.6 %) however it is lower than the accuracy for the *Seen* condition, which was expected since the *target speakers* for the condition *Unseen* were not seen by ***ZS-MSTM 2.0*** during training. Yet the

accuracy is high, and our model is able to generalize the style on new *unseen* speakers. For both conditions *Seen* and *Unseen*, the model is capable of preserving 94.8 % (seen) and 91 % (unseen) of the source speakers' content.

Minkowski Distance. As shown in Table 7.6, the distance between our model's predictions and the source speakers' gestures - ***Dist.(*ZS-MSTM 2.0*, Source)*** - is higher than the one between our model's predictions and the target speakers' gestures - ***Dist.(*ZS-MSTM 2.0*, Target)***. These results confirm that the *behavioral style* is successfully transferred from *target* to *source* speakers for both conditions.

7.7 Conclusion

We have presented in this Chapter ***ZS-MSTM 2.0***, a model based on ***ZS-MSTM 1.0*** that produces stylized facial gestures in addition to upper-body gestures. Not only our approach is capable of transferring the style of target speakers to source speakers in a *zero-shot* fashion, but our approach is independent of speakers' identity "ID", which allows us to generalize *behavioral style* from speakers *seen* by our model, to new *unseen* ones. In this Chapter, we extended *behavioral style* to also encode *facial gestures* on top of *body pose*, *speech*, and *text*.To the best of our knowledge, our approach is the first to show that *behavioral style* is *multimodal* and encoded through this large panel of modalities. Our objective evaluation results show that our model is able to perform zero-shot style transfer with a high *Style Strength Accuracy* and *Content Preservation* for both conditions *Seen* and *Unseen*. Similar to ***ZS-MSTM 1.0***, ***ZS-MSTM 2.0*** performs well for *seen* and *unseen* target speakers to source speakers without any fine-tuning or additional training. For future work, we will conduct perceptive evaluations to validate it subjectively.

The key points of this Chapter:

This Chapter addresses research Questions Q3, Q4, Q5

- Q3 - How can we learn style latent space of given speakers, given their multimodal data, and independently from their identity?
- Q4 - How can we synthesize body gestures of a source speaker, given the source speaker multimodal data, but with the style of another speaker?
- Q5 - How can we render our approach able to perform zero-shot style transfer on new unseen speakers, without the need of any further training or fine-tuning?

ZS-MSTM 2.0

- Similar to *ZS-MSTM 1.0*, *ZS-MSTM 2.0* is a Transformer-based architecture driven by the *content of* source speaker's speech - text semantics represented by *BERT Embeddings*, *Dialog Tags*, and audio Mel Spectrogram -, and conditioned on a target speaker's multimodal style embeddding.
- To encode *behavioral style*, on top of the inputs we used for *ZS-MSTM 1.0*, namely *2D poses*, *Mel spectrogram* and *BERT embeddings*; we added two other inputs *dialog tags*, *2D facial landmarks* of target speakers. The reason we have added *Dialog Tags* is to capture further semantic information in addition to *BERT embeddings*.
- We train *ZS-MSTM 2.0* on the extended version of the *PATS Corpus*, which includes the additional features - *2D Facial Landmarks* and *Dialog Tags* - as described in Chapter 4.
- To our knowledge and at the time of this research, *ZS-MSTM 2.0* is the first approach that synthesizes both, upper-body and facial gestures for ECAs, from speech and text inputs, in the style of *seen* or *unseen* speakers, in a *zero-shot* fashion.
- We follow the recommendations of Fu et al. [2018] who propose two novel evaluation metrics - *Transfer Strength Accuracy* and *Content Preservation* - to measure the characteristics of ***ZS-MSTM 2.0***'s style transfer.
- For both conditions *seen* and *unseen*, ***ZS-MSTM 2.0*** transfers the style from target speakers to source speakers with a *high* accuracy, and preserves the source speakers' content.
- For both conditions *seen* and *unseen*, we measure the *Minkowski distance* between the upper-body gestures and facial expressions produced by our model, and the ones of the *source* and *target* speakers. For both conditions, results show that the distance between our model's predictions and the source speakers' gestures is higher than the one between our model's predictions and the target speakers' gestures.

Chapter 8

Conclusion

This Chapter concludes the present dissertation. It starts out with a summary of the Chapters, presented in Section 8.1. Then follows in Section 8.2 a discussion of the contributions of this thesis to current research and their impacts. The last section, Section 8.3, reviews open issues, limitations, and points out to future directions of research.

8.1 Summary

The aim of this dissertation is to generate expressive human-like gestures for conversational embodied agents, to leverage their *behavioral expressivity* and control their *behavioral style*.

We started by presenting in Chapter 5 a novel approach for synthesizing facial gestures by exploiting *language semantics*, *speech prosody*, and *visual prosody*. More specifically, we proposed two different networks:

1. **Model 1.** First, we presented an *LSTM-based sequence to sequence network* that translates the input sequences of f_0 and *Bert Embeddings* to sequences of AUs and head rotations. We conducted an objective evaluation to assess the quality of the gestures produced by this model. Results revealed that the generated gestures' errors were low for some output features. However, the predictions were not correlated with the *Ground Truth*. These results were explained by the low AUs activation produced by this model, and high AUs non-activation.

2. **Model 2.** To overcome the weaknesses of the first LSTM-based model (*Model 1*), we proposed a novel approach that makes use of *Transformers* and *Convolutions* to generate the sequences of *eyebrows*, *eyelids* and *head* motion based on the input sequences of f_0 and *Bert Embeddings*. We compared this approach to *Model 1*, which served as a *baseline*. Results showed that *Model 2* surpassed the *baseline* in terms of the generated low errors and the high correlation with the ground truth. These results were also reflected in the high percentage of AU activation produced by *Model 2*, which were higher than the ones generated by the *baseline*. In addition to that, the percentage of AU non-activation produced by *Model 2* was lower than the one produced by the *baseline*. Moreover, we conducted subjective evaluations to investigate human perception of the facial gestures produced by *Model 2*. Results

showed that when simulated with **SD** data, *Model 2* produces animations that are closer to the *Ground Truth* than those of the *baseline* and *error* condition, in terms of *naturalness*, *human-likeness*, and *coherence*, and while ensuring that speech and computed gestures are *aligned* and *synchronized*. Moreover, when simulated with **SI** data, objective and subjective evaluation results showed that *Model 2* is capable of generalising its predictions to new speakers.

Both networks were trained on the *TEDx Corpus*, a corpus we presented in Chapter 4, and which consists of a large amount of multimodal features - audio, text, and facial features - corresponding to different TEDx talks speakers.

In addition, we proposed ***ZS-MSTM 1.0*** and ***ZS-MSTM 2.0*** for modelling embodied agents with *behavioral style*. Our approach is a machine learning approach that can synthesize *stylized* upper-body gestures (in ***ZS-MSTM 1.0***) and facial gestures (in ***ZS-MSTM 2.0***) driven by *audio* and *text semantics*. More specifically, our approach allows the synthesis of stylized *upper-body* and *facial* gestures, driven by the *content* of a source speaker's speech (audio and text) and corresponding to the style of *any* target speakers, *seen* or *unseen* by ***ZS-MSTM 1.0*** and ***ZS-MSTM 2.0***. ***ZS-MSTM 1.0*** was trained on the *PATS Corpus* which includes multimodal data of speakers having different *behavioral style*. We proposed an extension of the *PATS Corpus* that includes additional facial features - *2D Facial Landmarks* - and text features - *Dialog Tags* - aligned with the other *multimodal* features. ***ZS-MSTM 2.0*** was trained on the latter corpus. Nevertheless, ***ZS-MSTM 1.0 and 2.0*** are not limited to *PATS* speakers, and can produce gesture in the style of any newly coming speaker without further training or fine-tuning, rendering our approaches *zero-shot*. *Behavioral style* is modelled based on multimodal speakers' data, and is *independent* from the *speaker's identity* ("ID"), which allows our model to generalize style to new *unseen* speakers. We validated our approach by conducting objective (for ***ZS-MSTM 1.0 and 2.0***) and subjective (for ***ZS-MSTM 1.0***) evaluations. The results of these studies showed that ***ZS-MSTM 1.0 and 2.0*** generate stylized animations that are close to the target style, for target speakers that are *seen* and *unseen* by our model. Moreover, we compared the performance of ***ZS-MSTM 1.0*** w.r.t the state of the art ***Mix-StAGE*** and results showed that ***ZS-MSTM 1.0*** performs better in terms of *overall resemblance* of the generated gestures w.r.t the animations produced with the *target style*.

8.2 Summary of Contributions

On a broad level, this thesis contributes to both ECAs and Signal Processing research communities by providing novel approaches for synthesizing ECA's gestures based on *multimodal data* including *speech*. On a more specific level, we discuss in the following the contributions made by this thesis.

Corpora.

We have gathered a corpus that aims to provide a large amount of data that could be used for studying the relationships governing *speech audio*, *text semantics* and *facial gestures*. This corpus was used to train, test, and validate our models that were developed to generate speech-driven and semantically-aware facial gestures (research questions Q1, and Q2). Moreover, we extended the *PATS corpus* to include 2D facial landmarks and dialog tags. This corpus was used in this thesis to train, test, and validate our models that

were developed to tackle the research questions related to *body visual prosody expressivity* (research questions Q1, and Q2) and *multimodal style modelling and transfer* (research questions Q3, Q4, and Q5). We intend to share our datasets (that include the extracted multimodal features) to facilitate further research.

Semantically-aware and speech-driven facial gestures

We proposed the first learning-based model for synthesizing facial gestures including eyebrows motion and head rotations based on *speech prosody* and a representation of *text semantics*. To the best of our knowledge, predicting facial movements based on both speech prosody and text semantics was never investigated. Our Transformer-based model, with the usage of its different *attention-mechanisms* that are applied in the Transformer encoder and decoders, as well as in-between the embedding vectors of the input modalities, has surpassed our LSTM-based baseline. The Transformer network does not rely on past hidden states to capture the dependencies with previous tokens in a sequence. Transformers instead process a sequence as a whole. Moreover, *multi-head attention* and *positional embeddings* both provide information about the relationship between different tokens in a sequence. In addition, we showed that predicting eyebrows and head motion based on multimodal data - speech prosody and text semantics - improved the quality of the results compared to predicting them given only one of the input modalities. To our knowledge, this work is the first one to synthesize gestures driven by both *speech* and *semantics* trained on a *multispeaker* dataset. We intend to share our code to facilitate further research.

Zero-shot style transfer for *upper-body* and *facial gestures* synthesis

Our third contribution is the development of ***ZS-MSTM* (1.0 and 2.0)**, an approach for synthesizing stylized gestures in a *zero-shot* fashion. Our models ***ZS-MSTM 1.0 and 2.0*** surpassed the state of the art ***Mix-StAGE*** model in four main aspects:

1. First, both of our models can generalize style to new speakers without any fine-tuning or additional training, unlike *Mix-StAGE*. Its independence from the speaker's identity "ID" allows the generalization without being constrained and limited to the speakers used for training the model. DiffGAN was later on proposed by Ahuja et al. [2022] as an extension to *Mix-StAGE*, and an approach that performs *few-shot* style transfer strategy based on neural domain adaptation accounting for cross-modal grounding shift between source speaker and target style. However this adaptation still requires 2 minutes of the style to be transferred.

2. Second, *behavioral style* was never viewed as being *multimodal*; previous works limit behavior style to arm gestures only. However, both text and speech convey *style* information, and the embedding vector of *style* must take into account the three modalities. Indeed, with our post-hoc t-SNE analysis of the distributions of the style vectors at the output of each modality, we found that the motion style depends most on the *pose modality*, followed by the *speech*, then the *text semantics*.

3. Third, both of our models have surpassed the state of the art in terms of the resemblance of the predicted gestures to the target speakers gestures. The resemblance of the ***arms gesturing***, ***body orientation***, ***gesture amplitude***, ***gesture frequency*** and ***gesture velocity*** of the generated gestures w.r.t the *target style* is greater than the resemblance of the behaviors obtained with *Mix-StAGE* with the *target style*.

4. Fourth, ***ZS-MSTM 2.0*** generates both facial and upper-body gestures.

8.3 Limitations and Future Work

This thesis revolves around modelling multimodal data and learning the correlations between the different modalities - language, speech, and gesturing - to leverage the *behavioral expressivity* of embodied conversational agents and control their *behavioral style*. While we have made some strides in these core challenges, there are still some limitations that we will highlight in this section, and we hope that these ideas could inspire researchers in the field.

Head motion synthesis quality. First, the speakers in the *TEDx Corpus* address their TEDx talk to an audience sitting in a semi-circular manner, which has affected the quality of the head rotations produced by our model. The ECA performs head movements as if it was speaking to an audience sitting in a semi-circular way. We believe that the quality of head movements could be leveraged by training our model on another Corpus where speakers do not necessarily speak to a large audience, but to a limited one seated right in front of the speaker.

Synthesizing gesture shapes. Gesture shapes convey different meanings. For example, a pointing index can indicate a direction. Hand shapes and arm movement can describe an object, an action, etc. Several attempts have looked at modelling metaphoric gestures (Ravenet et al. [2018]), or iconic gestures (Bergmann and Kopp [2009a]). Most generative models of gestures do not compute the gesture shapes and motions for those specific gesture types. Extending our models to capture their gesture shapes and motion would be an interesting direction for future work. That would require extending the Corpora we have collected, to include specific annotations related to gestures shapes and to identify better representations (such as image schemas Grady [2005] for metaphoric gestures).

Conducting subjective evaluation of ZS-MSTM 1.0 and 2.0 on ECAs. The main limitation of ***ZS-MSTM 1.0*** *and* ***2.0*** is that they were not evaluated on ECAs. The main reason is that they were trained on the original *PATS Corpus* and its extended version, which include 2D poses and 2D facial landmarks. The graphical representation of the data as 2D stick figure is not always readable, even when being projected on the video of a human speaker. The main reason behind this problem is that the animation is missing information on the body pose in the Z direction (the depth axis). As a first attempt to solve this problem, we convert the 2D poses generated by ***ZS-MSTM 1.0*** into 3D poses and we visualize the behavior animation resulting from our model on a 3D virtual agent (Appendix B 10). The generated 2D body poses correspond to incomplete skeleton joints; missing joints include lower body joints, as well as torso joints. To visualize the resulting animations of our model, we convert the 2D poses into 3D poses and use 3D human mesh. However, the quality of the animations is still not very natural and smooth, and we believe that this problem could be solved by training our model on a Corpus with 3D poses and 3D facial landmarks.

Chapter 9

Appendix A

We report here additional information regarding the architecture described in Chapter 5.

Data Autoencoders

In this work, we used autoencoders for dimensionality reduction. An autoencoder compresses the input data into a lower-dimensional representation, and then converts it back to a reconstruction of the original input.

We developed and trained two different autoencoders AE_{AU} and AE_{f_0} that compress AU and f_0 values into a lower-dimensional representation, and then convert them back to a reconstruction of the original input. For each autoencoder, the encoder and decoder were trained jointly and used independently as different components in the overall network architecture, which is described in the next section. The autoencoders AE_{AU} and AE_{f_0} are based on Long-Short Term Memory (LSTM).

Long-Short Term Memory (LSTM). The LSTM is an artificial neural network designed to learn patterns in sequences of data. The main advantage of LSTMs is that they learn the temporal dimension of sequences. LSTM architecture consists of a set of recurrently connected subnets, known as *memory blocks*. The *memory blocks* learn the important parts of sequences seen so far, and forget the less important ones. This is achieved by the self connected memory cells and the multiplicative units *memory blocks* contain. The multiplicative units are *input*, *output* and *forget* gates that are analogues of *write*, *read* and *reset* operations for the memory cells. More specifically, these gates serve to:

1. provide a compact representation of the sequence seen so far,
2. learn to fuse new input with the past representation of the sequence,
3. forget the information that are not important,
4. learn what to predict for the next time step.

AE_{AU} and AE_{f_0} architectures.

Figure 9.1 depicts two f_0 contours corresponding to a word. The contour in blue corresponds to the original f_0 values, the one in red corresponds to the output of the

AE, which is the reconstruction of f_0. Figures 9.2 and 9.3 also illustrate original and reconstructed f_0 contours but for two words.

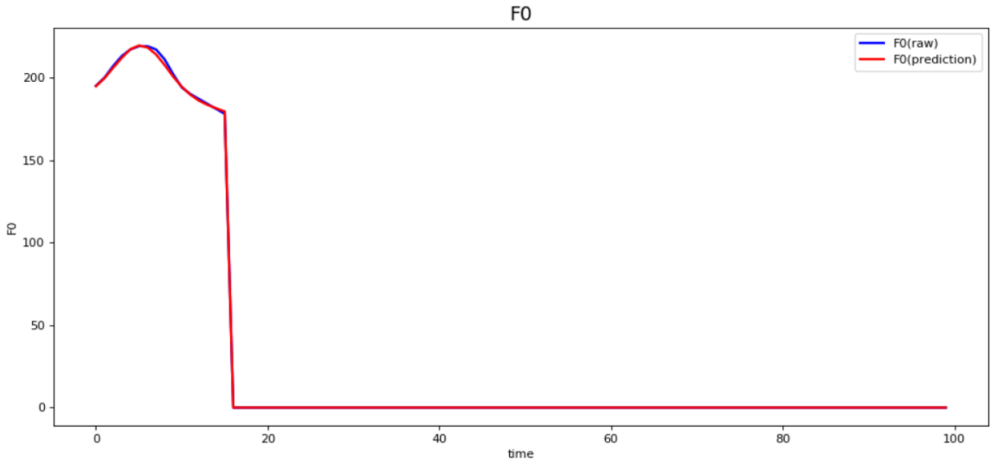


Figure 9.1 The f_0 contour in blue corresponds to the original f_0 values, the one in red corresponds to the output of the auto-encoder, which is the reconstruction of f_0.

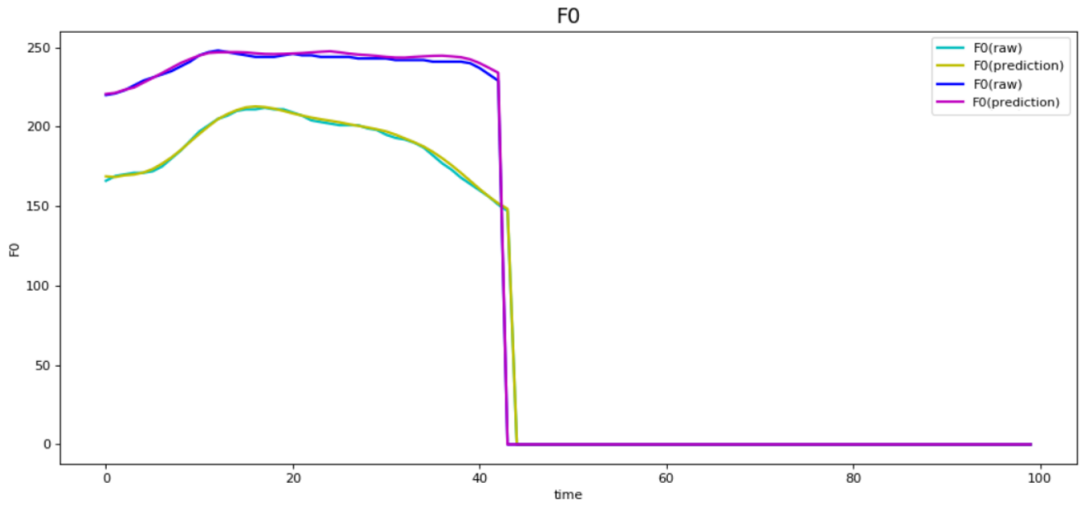


Figure 9.2 Original and predicted f_0 contours for two words

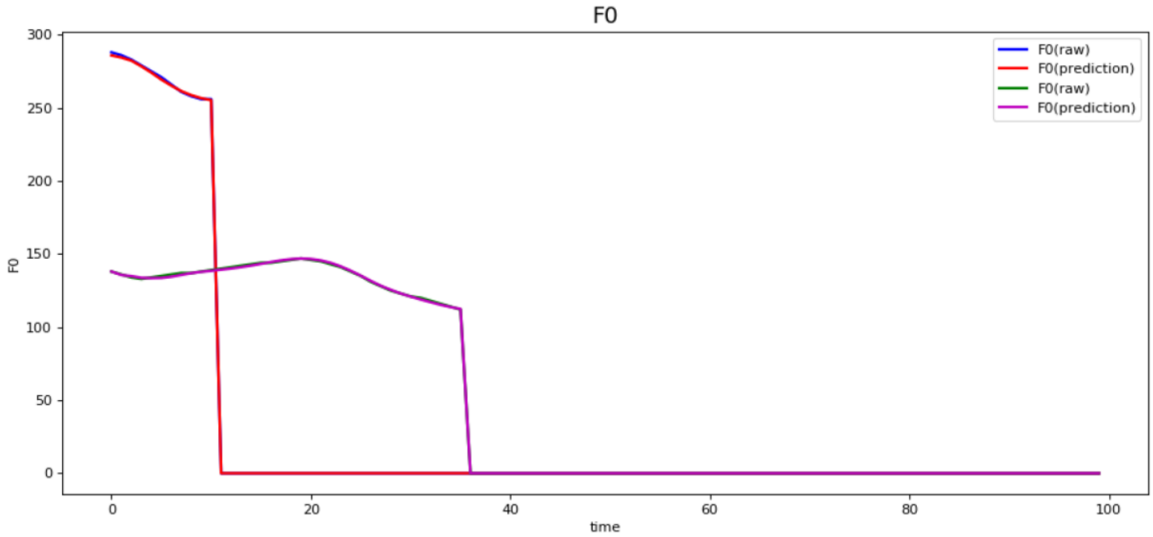


Figure 9.3 Original and predicted f_0 contours for two words

As Figures 9.1, 9.2, and 9.3 show, the autoencoder predicts f_0 contours with high accuracy.

Chapter 10

Appendix B

Human Perceptual Studies

This section provides additional tables describing the human perceptual studies results presented in Chapter 6, Section 6.8.1.

Study 1

	Overall Resemblance	Arms Gesturing	body orientation	Gesture Amplitude	Gesture Frequency	Gesture Velocity
Mean	4.32	4.24	4.21	4.17	4.18	4.18
Median	4.40	4.30	4.30	4.30	4.30	4.20
Standard deviation	0.535	0.441	0.534	0.448	0.520	0.511
Minimum	2.30	2.70	2.50	2.90	2.50	2.80
Maximum	5.00	5.00	5.00	5.00	5.00	5.00

Figure 10.1

Paired Samples T-Test

			statistic	df	p
Overall Resemblance	Arms Gesturing	Student's t	2.705	104	0.008
	body orientation	Student's t	3.333	104	0.001
	Gesture Amplitude	Student's t	4.266	104	< .001
	Gesture Frequency	Student's t	4.142	104	< .001
	Gesture Velocity	Student's t	3.479	104	< .001
Arms Gesturing	Gesture Amplitude	Student's t	2.606	104	0.010
	Gesture Velocity	Student's t	1.617	104	0.109
Gesture Amplitude	Gesture Velocity	Student's t	-0.336	104	0.737
Gesture Frequency		Student's t	0.165	104	0.869

Figure 10.2

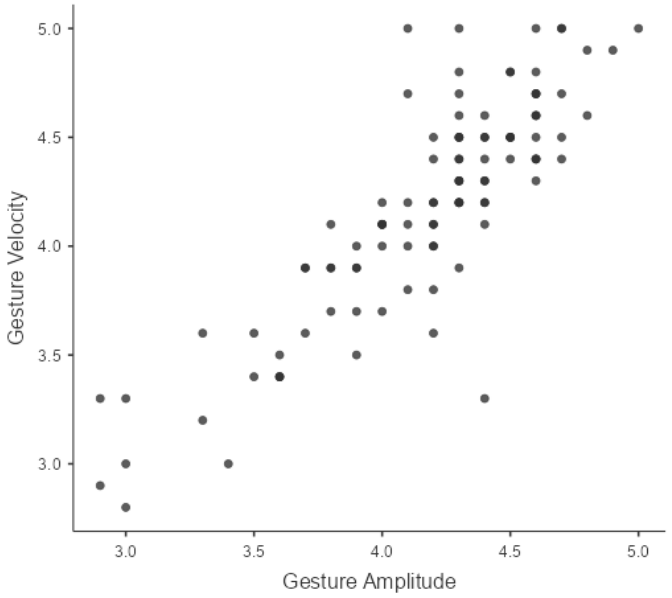


Figure 10.3

Study 2

	Overall Resemblance	Arms Gesturing	Body Orientation	Gesture Amplitude	Gesture Frequency	Gesture Velocity
Mean	3.83	3.78	3.68	3.66	3.73	3.72
Median	3.80	3.80	3.70	3.70	3.70	3.70
Standard deviation	0.364	0.359	0.406	0.378	0.411	0.390
Minimum	2.60	2.40	2.40	2.30	2.30	2.30
Maximum	4.60	4.50	4.70	4.60	4.80	4.60

Figure 10.4

			statistic	df	p
Overall Resemblance	Arms Gesturing	Student's t	2.487	105	0.014
	Body Orientation	Student's t	4.870	105	< .001
	Gesture Amplitude	Student's t	6.609	105	< .001
	Gesture Frequency	Student's t	3.798	105	< .001
	Gesture Velocity	Student's t	4.332	105	< .001
Arms Gesturing	Gesture Amplitude	Student's t	4.703	105	< .001
	Gesture Velocity	Student's t	2.153	105	0.034
Gesture Amplitude	Gesture Frequency	Student's t	-2.620	105	0.010
Gesture Frequency		Student's t	0.281	105	0.779

Figure 10.5

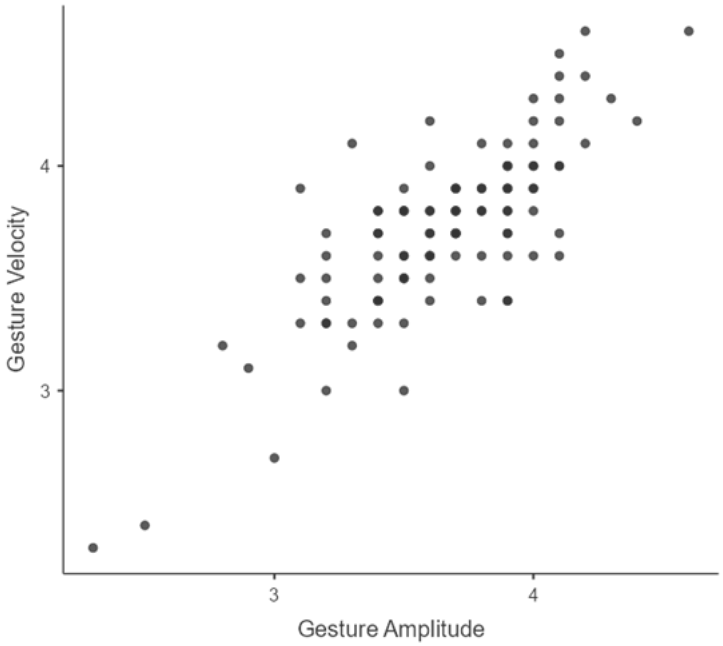


Figure 10.6

Study 3 - 3D Pose Generation and Simulation

Previous evaluation studies of models learned from video data have used 2D stick figures for their subjective evaluation (Ahuja et al. [2020b]). Even when the 2D stick figure is projected on the video of a human speaker, the animation is not always readable as, in particular, it is missing information on the body pose in the Z direction (the depth axis). So we choose to convert the 2D poses into 3D poses. We visualize the behavior animation resulting from our model on a 3D virtual agent. As in Ahuja et al. [2020b], we train our model on the database PATS, and therefore the generated 2D body poses correspond to incomplete skeleton joints; missing joints include lower body joints, as well as torso joints. To visualize the resulting animations of our model, we convert the 2D poses into 3D poses and use 3D human mesh.

We develop an approach that generates 3D poses from incomplete upper body 2D pose joints using MocapNET(**?**), an ensemble of SNN encoders that estimates the 3D human body pose based on 2D joint estimations extracted from monocular RGB images. It outputs skeletal information directly into the BVH format which can be rendered in real-time or imported without any additional processing in most popular 3D animation software. MocapNET operates on 2D joint input, received in the popular COCO(Cao et al. [2017]) or BODY25(Cao et al. [2017]) format. In order to be used, the file containing the predictions are formatted following the BODY25 format and the 2D joints are mapped to respect the BODY25 joints. The JSON files with 2D detections are subsequently converted to CSV files and then to 3D BVH files using the MocapNET. Finally we add zeros for the missing joints. MocapNET is trained using a 1920x1080 "virtual camera" to emulate a GoPRO Hero 4 running at the Full-HD mode. We adapted the output of our gesture generation model to such a configuration. We also set up the frames resolution to correspond to the original video stream size. Once the BVH file is created we use the 3D animation software

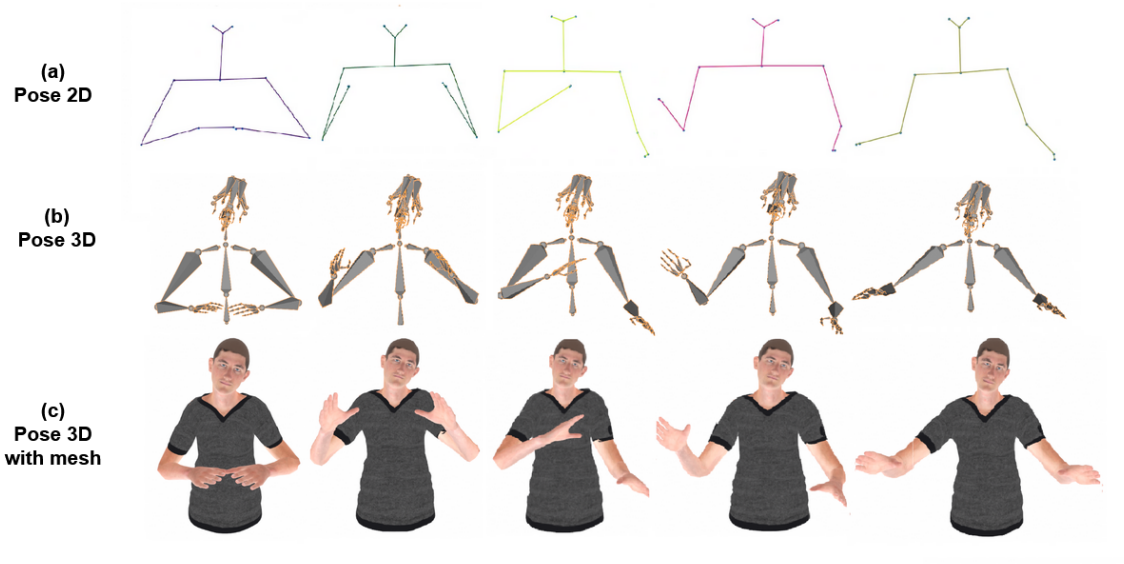


Figure 10.7 A sequence of gestures corresponding to a sequence of 2D poses. (a) 2D poses. (b) The corresponding sequence of 3D poses computed by MocapNet and simulated with Blender. (c) Resulting animation with a 3D human mesh.

Blender to simulate the animation. Finally, we apply a 3D human mesh to the skeleton to simulate a 3D human animation. The mesh is taken from Mixamo[1], an online database of characters and mocap animations used in art projects, movies and games. In order to fuse the mesh with the skeleton, we scale the mesh to fit the skeleton and we parent the skeleton and the mesh with automatic weights.

Additional Human Perceptual Studies

We additionally evaluate the 2D generated gestures by converting them to 3D poses, and simulating 3D animations of the generated gestures. The 3D poses generation is done from incomplete upper body 2D pose joints, using MocapNET, and are simulated on a 3D virtual agent. 3D poses estimation has never been done using 2D poses with such a large number of missing joints in the context of virtual agents animation.

We conduct three human perceptual studies.

As a pre-evaluation of our approach, we conduct a human perceptual study (**Study 1**) to validate the 2D to 3D pose conversion by measuring the *resemblance* of the 3D animations to the ground truth in terms of the expressivity of the style (gesture amplitude and dynamics), in addition to the quality of the produced 3D animations (such as naturalness and comprehensibility of movements).

Then, to investigate human perception of the stylized upper-body gestures produced by our model, we conduct another human perceptual study (**Study 2**) that aims to evaluate the gestures produced by our model and its capacity to perform "style preservation". A

[1]https://www.mixamo.com/

third study (**Study 3**) was conducted to evaluate the style transfer of speakers seen during training - *Seen Speaker* condition - , as well as speakers unseen during training - *Unseen Speaker* condition. In these studies, we present a virtual agent simulated with the converted 3D poses of the 2D poses synthesized by our model. **Study 3** aims to assess the *resemblance* of the produced stylized gestures to the target style. We additionally compare in **Study 3** our model's produced stylized gestures in *Seen Speaker* condition, to Mix-StAGE that we consider our baseline.

We used 7 factors linked to behavior expressivity to assess the quality of the 3D animation. We follow the recommendations proposed in Wolfert et al. [2022] and assess on a 1 to 7 likert scale the first 5 factors: *naturalness*, *coherence*, *human-likeness*, *appropriateness*, and *comprehensibility*. We add the 2 other factors *synchronization*, and *alignment* to evaluate the gestures' temporal property with speech.

In addition, 3 factors are used to evaluate the *resemblance*, *resemblance in terms of gestures amplitude*, and *resemblance in terms of gestures dynamics* between the human gestures and the virtual agent's gestures. We note that we distinguish between two types of factors that we want to assess in our studies: the first ones (7 expressivity factors) are related to evaluating the virtual agent's *behavioral expressivity* (for **Study 1** and **Study 2**), and the second ones (3 resemblance factors) are to assess the *resemblance* of our model's stylized produced gestures with the ground truth (for **Study 1**) and with the target style (for **Study 3**). Each factor is rated on a 7 likert scale. 30 participants are recruited for each study, including for the pre-evaluation study (**Study 1**), on Prolific, an online crowd-sourcing website. Participants are selected such that they are fluent in English and have a university degree. Attention checks are added in the beginning and the middle of each study to filter out inattentive participants. All the animations presented in these studies are produced on a 3D virtual agent.

Human Perceptual Studies

- **Study 1 - 3D Animation Pre-Evaluation**: The first human perceptual study we conduct aims to assess our approach for the 2D to 3D pose conversion. In this study, we present 4 pairs of videos: for each pair, the first video shows the generated 3D poses simulated on a virtual agent, and the second one is the video of the original speaker performing the same gestures. The 2D poses that we use for this 3D conversion are ground truth data extracted from the *PATS Corpus*.

- **Study 2 - Gesture Generation Evaluation**: To assess the quality of the 2D poses generated by our model, and its ability to perform "style preservation" and remember the unique style of each speaker, we conduct another human perceptual study. We use the 7 expressivity factors that are used in the pre-evaluation study to assess the quality of the produced virtual agent's gestures. This study consists of 8 videos: 4 videos show 3D animations of our model's predictions, and 4 other videos show the converted 2D to 3D poses animation of the original speaker's gestures which serve as ground truth. For each video, participants are asked to rate the 7 expressivity factors on a 1 to 7 likert scale (Wolfert et al. [2022]).

- **Study 3 - Style Transfer Evaluation**: The third perceptive study aims to assess the style transfer correctness performed by our model for both conditions: *Seen Speaker*

and *Unseen Speaker*. For each condition, participants watch 3 videos representing the ground truth (*video 1*), the target speaker (*video 2*) and our model (*video 3*), respectively. We ask the participants to answer questions related to the 3 resemblance factors, provided in a random order. For the *Seen Speaker* condition, we present 12 videos: 4 videos show the 3D animation of the source speaker gestures, 4 other videos show the 3D animation of the target speaker gestures, and the remaining 4 videos show the simulation of our model's predictions in 3D, after performing the style transfer from the target speaker to the source speaker. For the *Unseen Speaker* condition, we present 9 videos (different videos from the above ones): 3 videos with the source speaker gestures, 3 with the target speaker gestures, and the remaining 3 with our model's 3D simulated predictions after performing the style transfer from target speakers not seen during training, to the source speakers. We note that in this experimental study, the *resemblance* factors are the most important ones, since we want to assess the degree of resemblance of our model's stylized gestures to the target style. For each set of questions in each condition, the target 3D animation is presented to the participants as a "baseline". We ask the participants to choose one of the two video - the source speaker 3D animation, and the 3D simulation of our model's predictions - that resembles the most to the baseline in terms of the 3 resemblance factors. Participants are asked the following questions: (1) Which video resembles the most to the baseline video ?; (2) Which video resembles the most to the baseline video in terms of gestures dynamics ?; and (3) Which video resembles the most to the baseline video in terms of gestures amplitude ?

Comparing to the baseline **Mix-StAGE:** We additionally compare our stylized generated gestures in *Seen Speaker* condition with the predictions of *Mix-StAGE*(Ahuja et al. [2020b]), which serves as a baseline for this condition. We ask the participants to watch 3 videos representing the ground truth (*video 1*), the target speaker (*video 2*), and Mix-StAGE predictions after performing style transfer from target speakers to source speakers (*video 3*). We repeat this question 3 times (presenting 9 videos in total), and assess the *resemblance* of *Mix-StAGE*'s produced gestures with respect to the target speakers.

Human Perceptual Studies Results

Study 1 (*3D Animation Pre-Evaluation*): The first human perceptual study is the pre-evaluation to assess the 3D data animation and simulation on a virtual agent, which are converted from the 2D generated poses. We calculate the mean values obtained on the 7 expressivity factors and on the 3 resemblance factors.

Results show that all factors received a mean score above 3 on a likert-scale from 1 to 7. They reveal that the 2D to 3D conversion of the 2D-poses generated by our model tend to resemble the human's gestures which served as ground truth in this evaluation. We observe that the factor *Resemblance* gets the highest mean (above 4) and that the factor *Gestures Amplitude Resemblance* gets the highest second mean score, followed by the factor *Naturalness*. This indicates that the 3D animations show gestures that resemble the human's gestures, especially in terms of gestural amplitude resemblance. We obtain similar mean scores ($3.5 < mean < 3.6$) for the factors *Comprehensibility*, *Gestures Dynamics Resemblance*, *Likeness*, and *Alignment*. The mean score for the remaining factors is 3.1. While the 3D pose animation has not received the highest possible rate, its results are nevertheless good enough to be used as ground truth. In the remaining evaluations, all

Resemblance Metrics	ZS-MSTM - Seen Speaker		ZS-MSTM - Unseen Speaker		Mix-StAGE	
Resemblance to the target style	***Source Style***	***Prediction***	***Source Style***	***Prediction***	***Source Style***	***Prediction***
Globally	0.35 ± 0.02	0.65 ± 0.04	0.46 ± 0.01	0.54 ± 0.03	0.57 ± 0.03	0.43 ± 0.04
W.r.t. gesture dynamics	0.32 ± 0.05	0.68 ± 0.05	0.47 ± 0.02	0.53 ± 0.05	0.56 ± 0.03	0.44 ± 0.04
W.r.t. gesture amplitude	0.42 ± 0.03	0.58 ± 0.06	0.42 ± 0.04	0.58 ± 0.04	0.54 ± 0.05	0.46 ± 0.05

Table 10.1 Results of the perceptual study for the conditions ZS-MSTM (seen speakers), ZS-MSTM (unseen speakers), and baseline (Mix-StAGE). We also report the confidence intervals.

animations are obtained with this method, offering similar behavior quality.

Study 2 (*Gesture Generation Evaluation*): The second human perceptual study consists of assessing the quality of the generated poses and the ability of our model to perform "style preservation", thus its capacity of remembering the unique style of each speaker. We calculate the mean scores for the 7 behavioral expressivity factors.

We observe that our model's predictions (**P**) get mean values that are close to those of the ground truth (**GT**), especially for the factors *Appropriateness* (mean difference(**GT**, **P**)=0.1) and *Comprehensibility* (mean difference(**GT**, **P**)=0.3). The remaining factors have higher mean difference between the ground truth and predictions: *Coherence* (mean difference=0.4), *Human-likeness* (mean difference=0.44), *Synchronization* (mean difference=0.5), *Alignment* (mean difference=0.51), and *Synchronization* (mean difference=0.53). We additionally perform a Fisher's LSD Test to do pair-wise comparisons of the means of the 7 factors. Significant results ($p < 0.001$) are found for the factors *Appropriateness*, *Comprehensibility*, *Coherence* and *Human-Likeness* when comparing values for the Ground Truth gestures those of our model's generated gestures. This constitutes experimental validation that our model is perceived significantly close to the ground truth, and therefore allows "style preservation". Therefore, our model is able to remember the unique style of each speaker, even though it is trained on multiple ones. While our model is perceived significantly close to the ground truth, results show that we still need to leverage the synchronization of the produced gestures with the speech and its content.

Study 3 (*Style Transfer Evaluation*): The first four columns (**ZS-MSTM - Seen Speaker** and **ZS-MSTM - Unseen Speaker**) of Table 10.1 shows the results of the human perceptual study for assessing the stylized gestures generated by our model for both conditions *Seen Speaker* and *Unseen Speaker*. Results show that, on a scale from 0 to 1 representing the number of times our model is selected to resemble the target style, our model's predictions get values above 0.58 for condition *Seen Speaker*, and values between 0.53 and 0.58 for condition *Unseen Speaker*. Our model's generated style in condition *Unseen Speaker* is perceived as having quite high resemblance to the target style (score of 0.54), especially in terms of gesture amplitude (score of 0.53) and gesture dynamics (score of 0.58). We additionally performed t-test comparison between source style values and prediction style scores for the conditions *Unseen Speaker* and *Seen Speaker*. Significant results ($p < 0.001$) are found between the Source scores and the Prediction scores. These results reveal that our model's generated stylized gestures are significantly perceived as being closer to the target style than to the source style.

***Comparing to the baseline* Mix-StAGE:** The first two columns (**ZS-MSTM - Seen Speaker**)

and the last two columns (**Mix-StAGE**) of Table 10.1 present the results when comparing our generated gestures in condition *Seen Speaker* with the baseline *Mix-StAGE* which only operates in this condition and not in the condition *Unseen Speaker*.

On a scale from 0 to 1 representing the number of times our model is selected to resemble the target style, our model gets scores between 0.58 and 0.65, while Mix-StAGE gets lower scores, between 0.43 and 0.46. We additionally conduct a Fisher LSD test to do pair-wise comparisons of the means between the 3 factors of both conditions *Mix-StAGE* and *ZT-MSTM*, and identify the cases where the means are statistically different. We find a significant difference ($p < 0.003$) for the factor *Resemblance in terms of gesture dynamics*, and *Resemblance in terms of gesture amplitude*.

Bibliography

Facial Expression Recognition (Face Recognition Techniques) Part 1. http://what-when-how.com/face-recognition/facial-expression-recognition-face-recognition-techniques-part-1/.

GitHub. URL https://github.com/bhavitvyamalik/DialogTag.

Speech-to-text: Automatic speech recognition. URL \https://cloud.google.com/speech-to-text.

Tedx, ideas worth spreading. URL https://www.ted.com/.

Google code archive - long-term storage for google code project hosting. URL \https://code.google.com/archive/p/word2vec/.

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. In *2019 International Conference on Multimodal Interaction*, pages 74–84, 2019.

Chaitanya Ahuja, Dong Won Lee, Ryo Ishii, and Louis-Philippe Morency. No gestures left behind: Learning relationships between spoken language and freeform gestures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1884–1895, 2020a.

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. In *European Conference on Computer Vision*, pages 248–265. Springer, 2020b.

Chaitanya Ahuja, Dong Won Lee, and Louis-Philippe Morency. Low-resource adaptation for personalized co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Samer Al Moubayed, Jonas Beskow, and Gabriel Skantze. The furhat social companion talking head. In *INTERSPEECH*, pages 747–749, 2013.

Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. Style-controllable speech-driven gesture synthesis using normalising flows. In *Computer Graphics Forum*, volume 39, pages 487–496. Wiley Online Library, 2020.

Martha W Alibali, Julia L Evans, Autumn B Hostetter, Kristin Ryan, and Elina Mainela-Arnold. Gesture–speech integration in narrative: Are children less redundant than adults? *Gesture*, 9(3):290–311, 2009.

William Apple, Lynn A Streeter, and Robert M Krauss. Effects of pitch and speech rate on personal attributions. *Journal of personality and social psychology*, 37(5):page 715, 1979.

Michael Argyle. *Bodily communication*. Routledge, 2013.

David F Armstrong, William C Stokoe, and Sherman E Wilcox. *Gesture and the nature of language*. Cambridge University Press, 1995.

Ronald J Baken and Robert F Orlikoff. *Clinical measurement of speech and voice*. Cengage Learning, 2000.

G Ball and J Breese. Emotion and personality in a conversational agent, embodied conversational agents, edited by justine cassel, joseph sullivan, scott prevost and elizabeth churchill, 2000.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016.

Allan Bell. Language style as audience design. *Language in society*, 13(2):145–204, 1984.

Kirsten Bergmann and Stefan Kopp. Gnetic–using bayesian decision networks for iconic gesture generation. In *International workshop on intelligent virtual agents*, pages 76–89. Springer, 2009a.

Kirsten Bergmann and Stefan Kopp. Increasing the expressiveness of virtual agents: autonomous generation of speech and gesture for spatial description tasks. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 361–368, 2009b.

Emmanuel Biau, Luis Morís Fernández, Henning Holle, César Avila, and Salvador Soto-Faraco. Hand gestures as visual prosody: Bold responses to audio–visual alignment are modulated by the communicative nature of the stimuli. *Neuroimage*, 132:129–137, 2016.

RL Birdwhistell. The language of the body: The natural environment of words. hillsday: Ed. *A. Silverstein Hillsdale*, 1974.

Barry Bogin and Carlos Varea. Evolution of human life history. In *Evolutionary Neuroscience*, pages 753–767. Elsevier, 2020.

D Bolinger. *Intonation and its uses: Melody in grammar and discourse*. Stanford university press, 1989.

Sheryl Brahnam, Chao-Fa Chuang, Randall S Sexton, and Frank Y Shih. Machine assessment of neonatal facial expressions of acute pain. *Decision Support Systems*, 43(4): 1242–1254, 2007.

Matthew Brand and Aaron Hertzmann. Style machines. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 183–192, 2000.

Matthew Brand and Vera Kettnaker. Discovery and segmentation of activities in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):844–851, 2000.

Matthew Brand, Nuria Oliver, and Alex Pentland. Coupled hidden markov models for complex action recognition. In *Proceedings of IEEE computer society conference on Computer Vision and Pattern Recognition*, pages 994–999. IEEE, 1997.

Diane Brentari, Carolina González, Amanda Seidl, and Ronnie Wilbur. Sensitivity to visual prosodic cues in signers and nonsigners. *Language and Speech*, 54(1):49–72, 2011.

Peter Bull and Gerry Connelly. Body movement and emphasis in speech. *Journal of Nonverbal Behavior*, 9(3):169–187, 1985.

David B Buller and R Kelly Aune. The effects of speech rate similarity on compliance: Application of communication accommodation theory. *Western Journal of Communication*, 56(1):37–53, 1992.

Judee K Burgoon, Laura K Guerrero, and Valerie Manusov. *Nonverbal communication*. Routledge, 2016.

Judee K Burgoon, Valerie Manusov, and Laura K Guerrero. *Nonverbal communication*. Routledge, 2021.

Richard W Byrne, Erica Cartmill, Emilie Genty, Kirsty E Graham, Catherine Hobaiter, and Joanne Tanner. Great ape gestures: intentional communication with a rich set of innate signals. *Animal cognition*, 20(4):755–769, 2017.

Geneviève Calbris. *Elements of Meaning in Gesture*. John Benjamins Publishing Company, 2011.

Arturo Camacho and John G Harris. A sawtooth waveform inspired pitch estimator for speech and music. *The Journal of the Acoustical Society of America*, 124(3):1638–1652, 2008.

Kathryn Campbell-Kibler, Penelope Eckert, Norma Mendoza-Denton, and Emma Moore. The elements of style. In *Poster presented at New Ways of Analyzing Variation*, volume 35, 2006.

Yong Cao, Wen C Tien, Petros Faloutsos, and Frédéric Pighin. Expressive speech-driven facial animation. *ACM Transactions on Graphics (TOG)*, 24(4):1283–1302, 2005.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

J. Cassell. Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents. In S. Prevost J. Cassell, J. Sullivan and E. Churchill, editors, *Embodied Conversational Characters*. MITpress, Cambridge, MA, 2000.

Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. Beat: the behavior expression animation toolkit. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 477–486, 2001.

Brandon Castellano. Pyscenedetect. *Last accessed*. URL http://scenedetect.com.

Christian Cavé, Isabelle Guaïtella, Roxane Bertrand, Serge Santi, Françoise Harlay, and Robert Espesser. About the relationship between eyebrow movements and fo variations. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 4, pages 2175–2178. IEEE, 1996.

Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. Predicting co-verbal gestures: A deep and temporal modeling approach. In *International Conference on Intelligent Virtual Agents*, pages 152–166. Springer, 2015.

Nicole Chovil. Discourse-oriented facial displays in conversation. *Research on Language & Social Interaction*, 25(1-4):163–194, 1991.

A Cohen et al. Intonation by rule: a perceptual quest. *Journal of Phonetics*, 1(4):309–327, 1973.

Michael C Corballis. The gestural origins of language: Human language may have evolved from manual gestures, which survive today as a" behavioral fossil" coupled to speech. *American Scientist*, 87(2):138–145, 1999.

David Crystal. Prosodic development. *Language acquisition*, pages 33–48, 1986.

Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10101–10111, 2019.

Svetlana Dachkovsky and Wendy Sandler. Visual intonation in the prosody of a sign language. *Language and speech*, 52(2-3):287–314, 2009.

Charles Darwin. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

Jim Davies. Program good ethics into artificial intelligence. *Nature*, 2016.

Fiorella De Rosis, Catherine Pelachaud, Isabella Poggi, Valeria Carofiglio, and Berardina De Carolis. From greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent. *International journal of human-computer studies*, 59 (1-2):81–118, 2003.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of Association for Computational Linguistics: Human Language Technologies,*, pages 4171–4186, Minneapolis, Minnesota, June 2019.

James Price Dillard. *Persuasion*. 2010.

Chuang Ding, Lei Xie, and Pengcheng Zhu. Head motion synthesis from speech using deep neural networks. *Multimedia Tools and Applications*, 74(22):9871–9888, 2015.

Yu Ding, Catherine Pelachaud, and Thierry Artieres. Modeling multimodal behaviors from speech prosody. In *International Workshop on Intelligent Virtual Agents*, pages 217–228. Springer, 2013.

Amanda Cardoso Duarte, Francisco Roldan, Miquel Tubau, Janna Escur, Santiago Pascual, Amaia Salvador, Eva Mohedano, Kevin McGuinness, Jordi Torres, and Xavier Giro-i Nieto. Wav2pix: Speech-conditioned face generation using generative adversarial networks. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 8633–8637, 2019.

Susan D Duncan, Justine Cassell, and Elena T Levy. *Gesture and the dynamic dimension of language: Essays in honor of David McNeill*, volume 1. John Benjamins Publishing, 2007.

Funda Durupinar, Mubbasir Kapadia, Susan Deutsch, Michael Neff, and Norman I Badler. Perform: Perceptual approach for adding ocean personality to human motion using laban movement analysis. *ACM Transactions on Graphics (TOG)*, 36(1):1–16, 2016.

Paul Ekman. Facial expressions of emotion: an old controversy and new findings. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 335 (1273):63–69, 1992.

Paul Ekman. Emotional and conversational nonverbal signals. In *Language, knowledge, and representation*, pages 39–50. Springer, 2004.

Paul Ekman and Wallace V Friesen. Felt, false, and miserable smiles. *Journal of nonverbal behavior*, 6(4):238–252, 1982.

Paul Ekman and Harriet Oster. Facial expressions of emotion. *Annual review of psychology*, 30(1):527–554, 1979.

Paul Ekman, Wallace V Friesen, and Silvan S Tomkins. Facial affect scoring technique: A first validity study. *Walter de Gruyter*, 1971.

Núria Esteve-Gibert and Bahia Guellaï. Prosody in the auditory and visual domains: A developmental perspective. *Frontiers in Psychology*, 9:338, 2018.

Núria Esteve-Gibert and Pilar Prieto. Prosodic structure shapes the temporal realization of intonation and manual gesture movements. *ASHA*, 2013.

Núria Esteve-Gibert, Joan Borràs-Comes, Eli Asor, Marc Swerts, and Pilar Prieto. The timing of head movements: The role of prosodic heads and edges. *The Journal of the Acoustical Society of America*, 141(6):4727–4739, 2017.

Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838, 2013.

Mireille Fares. Towards multimodal human-like characteristics and expressive visual prosody in virtual agents. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 743–747, 2020.

Mireille Fares, Catherine Pelachaud, and Nicolas Obin. Multimodal generation of upper-facial and head gestures with a transformer network using speech and text. *arXiv preprint arXiv:2110.04527*, 2021.

Mireille Fares, Catherine Pelachaud, and Nicolas Obin. Transformer network for semantically-aware and speech-driven upper-face generation. In *European Signal Processing Conference (EUSIPCO)*, 2022.

Christiane Fellbaum. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer, 2010.

Ylva Ferstl, Michael Neff, and Rachel McDonnell. Multi-objective adversarial gesture generation. In *Motion, Interaction and Games*, pages 1–10. 2019.

Ylva Ferstl, Michael Neff, and Rachel McDonnell. Adversarial gesture generation with realistic gesture phasing. *Computers & Graphics*, 89:117–130, 2020.

Pierre Feyereisen and Isabelle Havard. Mental imagery and production of hand gestures while speaking in younger and older adults. *Journal of nonverbal behavior*, 23(2):153–171, 1999.

Pierre Feyereisen, Jacques-Dominique De Lannoy, et al. *Gestures and speech: Psychological investigations*. Cambridge University Press, 1991.

Jeff Forbes, Timothy Huang, Keiji Kanazawa, and Stuart Russell. The batmobile: Towards a bayesian automated taxi. In *International Joint Conference on Artificial Intelligence (IJCAI)*, volume 95, pages 1878–1885, 1995.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Aphrodite Galata, Neil Johnson, and David Hogg. Learning variable-length markov models of behavior. *Computer Vision and Image Understanding*, 81(3):398–413, 2001.

Pablo Garrido, Levi Valgaerts, Hamid Sarmadi, Ingmar Steiner, Kiran Varanasi, Patrick Perez, and Christian Theobalt. Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. In *Computer graphics forum*, volume 34, pages 193–204. Wiley Online Library, 2015.

Howard Giles. *The Nonverbal Communication Reader: Classic and Contemporary Readings*, page 425, 1999.

Howard Giles. *Communication accommodation theory: Negotiating personal relationships and social identities across contexts*. Cambridge University Press, 2016.

Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Frieda Goldman-Eisler. Segmentation of input in simultaneous translation. *Journal of psycholinguistic Research*, 1(2):127–140, 1972.

S Gong and H Buxton. Advanced visual surveillance using bayesian nets. In *IEEE Workshop on Context-Based Vision, Cambridge, MA*, 1995.

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

Joseph E Grady. Image schemas and perception: Refining a definition. *From perception to meaning: Image schemas in cognitive linguistics*, 29:35, 2005.

H. P. Graf, E. Cosatto, V. Strom, and Fu Jie Huang. Visual prosody: facial movements accompanying speech. In *IEEE Int Conf on Automatic Face Gesture Recognition*, pages 396–401, 2002.

Hans Peter Graf, Eric Cosatto, Volker Strom, and Fu Jie Huang. Visual prosody: Facial movements accompanying speech. In *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pages 396–401. IEEE, 2002.

DiAnne L Grieser and Patricia K Kuhl. Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. *Developmental psychology*, 24(1): 14, 1988.

Bahia Guellaï, Alan Langus, and Marina Nespor. Prosody in the hands of the speaker. *Frontiers in Psychology*, 5:700, 2014.

Yunhui Guo. A survey on methods and theories of quantized neural networks. *arXiv preprint arXiv:1808.04752*, 2018.

K Haag and H Shimodaira. Bidirectional LSTM networks employing stacked bottleneck features for expressive speech-driven head motion synthesis. In *Intelligent Virtual Agents*, pages 198–207, 2016.

Uri Hadar, Timothy J Steiner, Ewan C Grant, and Frank Clifford Rose. Head movement correlates of juncture and stress at sentence level. *Language and speech*, 26(2):117–129, 1983.

P EKMAN-WV FRIESEN-JC HAGER. Facial action coding system. the manual on cd rom, 2002.

Björn Hartmann, Maurizio Mancini, and Catherine Pelachaud. Implementing expressive gesture synthesis for embodied conversational agents. In *Gesture in Human-Computer Interaction and Simulation: 6th International Gesture Workshop, GW 2005, Berder Island, France, May 18-20, 2005, Revised Selected Papers 6*, pages 188–199. Springer, 2006.

Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. Evaluation of speech-to-gesture generation using bi-directional lstm network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 79–86, 2018.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, Jürgen Schmidhuber, et al. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, 2001.

G Hofer and H Shimodaira. Automatic head motion prediction from speech data. In *Interspeech*, 2007.

Somboon Hongeng, Francois Bremond, and Ramakant Nevatia. Representation and optimal recognition of human activities. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 1, pages 818–825. IEEE, 2000.

Autumn B Hostetter and Martha W Alibali. Raise your hand if you're spatial: Relations between verbal and spatial skills and gesture production. *Gesture*, 7(1):73–95, 2007.

Autumn B Hostetter and Andrea L Potthoff. Effects of personality and social situation on representational gesture production. *Gesture*, 12(1):62–83, 2012.

O Hrinchuk, M Popova, and B Ginsburg. Correction of automatic speech recognition with transformer sequence-to-sequence model. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7074–7078. IEEE, 2020.

Carlos T Ishi, Daichi Machiyashiki, Ryusuke Mikata, and Hiroshi Ishiguro. A speech-driven hand gesture generation method and evaluation in android robots. *IEEE Robotics and Automation Letters*, 3(4):3757–3764, 2018.

Jana M Iverson and Susan Goldin-Meadow. Why people gesture when they speak. *Nature*, 396(6708):228–228, 1998.

Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. You said that?: Synthesising talking faces from audio. *International Journal of Computer Vision*, 127(11):1767–1779, 2019.

Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, 31, 2018.

Patrik Jonell, Taras Kucherenko, Gustav Eje Henter, and Jonas Beskow. Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2020.

Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017.

Adam Kendon. Some relationships between body motion and speech. *Studies in dyadic communication*, 7(177):90, 1972.

Adam Kendon. *Gesture: Visible action as utterance*. Cambridge University Press, 2004.

Michael Kipp. Anvil-a generic annotation tool for multimodal dialogue. In *Seventh European conference on speech communication and technology*, 2001.

Michael Kipp. *Gesture generation by imitation: From human behavior to computer character animation*. Universal-Publishers, 2005.

Michael Kipp, Michael Neff, and Irene Albrecht. An annotation scheme for conversational gestures: how to economically capture timing and form. *Language Resources and Evaluation*, 41(3):325–339, 2007.

Sotaro Kita. Cross-cultural variation of speech-accompanying gesture: A review. *Language and cognitive processes*, 24(2):145–167, 2009.

Mark L Knapp, Judith A Hall, and Terrence G Horgan. *Nonverbal communication in human interaction*. Cengage Learning, 2013.

Stefan Kopp, Bernhard Jung, Nadine Lessmann, and Ipke Wachsmuth. Max-a multimodal assistant in virtual reality construction. *KI*, 17(4):11, 2003.

Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. Towards a common framework for multimodal generation: The behavior markup language. In *International workshop on intelligent virtual agents*, pages 205–217. Springer, 2006.

Jelena Krivokapić, Mark K Tiede, and Martha E Tyrone. A kinematic study of prosodic structure in articulatory and manual gestures: Results from a novel method of data collection. *Laboratory phonology*, 8(1), 2017.

Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 97–104, 2019.

Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexanderson, Iolanda Leite, and Hedvig Kjellström. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the ACM International Conference on Multimodal Interaction*, 2020.

Nadezhda Nikolaevna Ladygina-Kohts, Frans de Waal, and Boris Trans Wekker. *Infant chimpanzee and human child: A classic 1935 comparative study of ape emotions and intelligence.* Oxford University Press, 2002.

Robin Lakoff and Robin Tolmach Lakoff. *Language and woman's place: Text and commentaries*, volume 3. Oxford University Press, USA, 2004.

Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. *Advances in neural information processing systems*, 30, 2017.

Pierre Lanchantin, Andrew C Morris, Xavier Rodet, and Christophe Veaux. Automatic phoneme segmentation with relaxed textual constraints. In *International Conference on Language Resources and Evaluation (LREC)*, 2008.

Willem JM Levelt, Graham Richardson, and Wido La Heij. Pointing and voicing in deictic expressions. *Journal of Memory and Language*, 24(2):133–164, 1985.

Sergey Levine, Christian Theobalt, and Vladlen Koltun. Real-time prosody-driven synthesis of body language. In *ACM SIGGRAPH Asia*, pages 1–10. 2009.

Stephen C Levinson and Judith Holler. The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651), 2014.

Stephen C Levinson, Stephen C Levinson, and S Levinson. *Pragmatics*. Cambridge university press, 1983.

Lin Liao, Donald J Patterson, Dieter Fox, and Henry Kautz. Learning and inferring transportation routines. *Artificial intelligence*, 171(5-6):311–331, 2007.

Scott K Liddell and Melanie Metzger. Gesture in sign language discourse. *Journal of pragmatics*, 30(6):657–697, 1998.

Jacqueline Lindenfeld. Verbal and non-verbal elements in discourse. *Walter de Gruyter*, 1971.

J Lu and H Shimodaira. Prediction of head motion from speech waveforms with a canonical-correlation-constrained autoencoder. *arXiv preprint arXiv:2002.01869*, 2020.

Birgit Lugrin. Introduction to socially interactive agents. In *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition*, pages 1–20. 2021.

S Mariooryad and C Busso. Generating human-like behaviors using joint, speech-driven models for conversational agents. *IEEE Trans on Audio, Speech, & Language Processing*, 20(8), 2012.

Stacy Marsella, Ari Shapiro, Andrew Feng, Yuyu Xu, Margaux Lhommet, and Stefan Scherer. Towards higher quality character performance in previz. In *Proceedings of the Symposium on Digital Production*, pages 31–35, 2013a.

Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics symposium on computer animation*, pages 25–35, 2013b.

Robert R McCrae and Paul T Costa Jr. Personality trait structure as a human universal. *American psychologist*, 52(5):509, 1997.

David McNeill. Hand and mind: What gestures reveal about thought. *Leonardo*, 27(4): 358–358, 1994.

David McNeill and Elena Levy. Conceptual representations in language activity and gesture. *Speech, place, and action*, pages 271–295, 1982.

David McNeill, Bennett Bertenthal, Jonathan Cole, and Shaun Gallagher. Gesture-first, but no gestures? *Behavioral and Brain Sciences*, 28(2):138–139, 2005.

Norma Mendoza-Denton. Style. *Journal of Linguistic Anthropology*, 9(1/2):238–240, 1999.

Ineke Mennen, Felix Schaeffler, and Gerard Docherty. Pitching it differently: A comparison of the pitch ranges of german and english speakers. In *16th International Congress of Phonetic Sciences*, 2007.

A Mohamed, D Okhonko, and L Zettlemoyer. Transformers with convolutional context for asr. *arXiv preprint arXiv:1904.11660*, 2019.

Carlos Monzo, Ignasi Iriondo, and Joan Claudi Socoró. Voice quality modelling for expressive speech synthesis. *The Scientific World Journal*, 2014, 2014.

Sungwoo Moon, Sunghyun Kim, and Yong-Hoon Choi. Mist-tacotron: End-to-end emotional speech synthesis using mel-spectrogram image style transfer. *IEEE Access*, 10: 25455–25463, 2022.

Kevin G Munhall, Jeffery A Jones, Daniel E Callan, Takaaki Kuratate, and Eric Vatikiotis-Bateson. Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological science*, 15(2):133–137, 2004.

Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics (TOG)*, 27(1):1–24, 2008.

Marina Nespor and Wendy Sandler. Prosody in israeli sign language. *Language and speech*, 42(2-3):143–176, 1999.

Graham Neubig. Neural machine translation and sequence-to-sequence models: A tutorial. *arXiv preprint arXiv:1703.01619*, 2017.

Sigrid Norris. *Analyzing multimodal interaction: A methodological framework*. Routledge, 2004.

Christian Obermeier, Spencer D Kelly, and Thomas C Gunter. A speaker's gesture style can affect language comprehension: Erp evidence from gesture-speech integration. *Social cognitive and affective neuroscience*, 10(9):1236–1243, 2015.

Nicolas Obin. *MeLos: Analysis and modelling of speech prosody and speaking style*. PhD thesis, 2011.

Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T Freeman, Michael Rubinstein, and Wojciech Matusik. Speech2face: Learning the face behind a voice. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7539–7548, 2019.

W Q Ong, A W C Tan, V V Vengadasalam, C H Tan, and T H Ooi. Real-time robust voice activity detection using the upper envelope weighted entropy measure and the dual-rate adaptive nonlinear filter. *Entropy*, 19(11), 2017.

C Pelachaud. Greta: a conversing socio-emotional agent. In *ACM SIGCHI Int WS on ISIAA*, pages 9–10, 2017.

Catherine Pelachaud. Studies on gesture expressivity for a virtual agent. *Speech Communication*, 51(7):630–639, 2009.

Catherine Pelachaud. Greta: an interactive expressive embodied conversational agent. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 5–5, 2015.

Catherine Pelachaud, Norman I Badler, and Mark Steedman. Generating facial expressions for speech. *Cognitive science*, 20(1):1–46, 1996.

Catherine Pelachaud, Valeria Carofiglio, Berardina De Carolis, Fiorella de Rosis, and Isabella Poggi. Embodied contextual agent in information delivering application. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2*, pages 758–765, 2002.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

Janet Pierrehumbert. The meaning of intonational contours in the interpretation of discourse janet pierrehumbert and julia hirschberg. *Intentions in communication*, 271, 1990.

Fernando Poyatos. Paralinguistic qualifiers: Our many voices. *Language & Communication*, 11(3):181–195, 1991.

Brian Ravenet, Catherine Pelachaud, Chloé Clavel, and Stacy Marsella. Automating the production of communicative gestures in embodied characters. *Frontiers in psychology*, 9:1144, 2018.

Benjamin Roustan and Marion Dohen. Co-production of contrastive prosodic focus and manual gestures: Temporal coordination and effects on the acoustic and articulatory correlates of focus. In *5th International Conference on Speech Prosody*, 2010.

J de Ruiter. *Gesture and speech production*. 1998.

Zsófia Ruttkay and Catherine Pelachaud. *From brows to trust: Evaluating embodied conversational agents*, volume 7. Springer Science & Business Media, 2004.

N Sadoughi and C Busso. Speech-driven animation with meaningful behaviors. *Speech Communication*, 110:90–100, 2019.

Najmeh Sadoughi and Carlos Busso. Novel realizations of speech-driven head movements with generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6169–6173. IEEE, 2018.

M Salem, K Rohlfing, S Kopp, and F Joublin. A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction. In *RoMan*, pages 247–252. IEEE, 2011.

Mehmet E Sargin, Yucel Yemez, Engin Erzin, and Ahmet M Tekalp. Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1330–1345, 2008.

Albert E Scheflen. The significance of posture in communication systems. *Psychiatry*, 27 (4):316–331, 1964.

Syamimi Shamsuddin, Luthffi Idzhar Ismail, Hanafiah Yussof, Nur Ismarrubie Zahari, Saiful Bahari, Hafizan Hashim, and Ahmed Jaffar. Humanoid robot nao: Review of control and motion exploration. In *2011 IEEE international conference on Control System, Computing and Engineering*, pages 511–516. IEEE, 2011.

Ren Shaozeng. Culture, discourse, and choice of structure. *Educational Linguistics, Cross-cultural Communication, and Global Interdependence*, page 150, 1995.

Lihong Shen. Context and text. *Theory and practice in language studies*, 2(12):2663, 2012.

Chilin Shih. A declination model of mandarin chinese. In *Intonation*, pages 243–268. Springer, 2000.

Chilin Shih and Greg Kochanski. Prosody and prosodic models, 2002. URL http://www.cs.columbia.edu/~julia/cs4706/chilin.htm.

Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. Audio to body dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7574–7583, 2018.

Harrison Jesse Smith and Michael Neff. Understanding the impact of animated gesture performance on personality perceptions. *ACM Transactions on Graphics (TOG)*, 36(4): 1–12, 2017.

Catherine E Snow and Charles A Ferguson. Talking to children: Language input and acquisition. 1977.

Yang Song, Jingwen Zhu, Dawei Li, Xiaolong Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. *arXiv preprint arXiv:1804.04786*, 2018.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)*, 36(4):1–11, 2017.

IR Titze. *Principles of Voice Production*. Prentice-Hall Inc., 1994.

Els Van der Kooij, Onno Crasborn, and Wim Emmerik. Explaining prosodic body leans in sign language of the netherlands: Pragmatics required. *Journal of Pragmatics*, 38(10): 1598–1614, 2006.

A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A N Gomez, L Kaiser, and I Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

K Vougioukas, S Petridis, and M Pantic. Realistic speech-driven facial animation with gans. *Int Journal of Computer Vision*, pages 1–16, 2019.

Dirk Nicolas Wagner et al. Augmented human-centered management. human resource development for highly automated business environments. *Journal of Human Resource Management*, 23(1):13–27, 2020.

Petra Wagner, Zofia Malisz, and Stefan Kopp. Gesture and speech in interaction: An overview, 2014.

H.G. Wallbott. Bodily expression of emotion. *European Journal of Social Psychology*, 28: 879–896, 1998.

S Wang, L Li, Y Ding, C Fan, and X Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *preprint arXiv:2107.09293*, 2021.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning*, pages 5180–5189. Proceedings of Machine Learning Research (PMLR), 2018.

Ronnie B Wilbur. The use of asl to support the development of english and literacy. *Journal of deaf studies and deaf education*, 5(1):81–104, 2000.

Andrew D Wilson and Aaron F Bobick. Recognition and interpretation of parametric gesture. In *Sixth International Conference on Computer Vision*, pages 329–336. IEEE, 1998.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Pieter Wolfert, Nicole Robinson, and Tony Belpaeme. A review of evaluation practices of gesture generation in embodied conversational agents. *IEEE Transactions on Human-Machine Systems*, 2022.

Yi Xu. Prosody, tone and intonation. *The Routledge handbook of phonetics*, pages 314–356, 2019.

Shaowei Yao and Xiaojun Wan. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4346–4350, 2020.

Hani C Yehia, Takaaki Kuratate, and Eric Vatikiotis-Bateson. Linking facial animation, head motion and speech acoustics. *Journal of phonetics*, 30(3):555–568, 2002.

Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4303–4309. IEEE, 2019.

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 (TOG)*, 39(6):1–16, 2020.

Heiga Ze, Andrew Senior, and Mike Schuster. Statistical parametric speech synthesis using deep neural networks. In *2013 International Conference on Acoustics, Speech and Signal Processing*, pages 7962–7966. IEEE, 2013.

Goranka Zoric, Karlo Smid, and Igor Pandzic. Automated gesturing for embodied animated agent: Speech-driven and text-driven approaches. *Journal of Multimedia*, 1(1), 2006.

Goranka Zoric, Karlo Smid, and Igor S Pandzic. Facial gestures: taxonomy and application of non-verbal, non-emotional facial displays for embodied conversational agents. *Conversational Informatics: An Engineering Approach*, pages 161–182, 2007.