Journal of Universal Computer Science, vol. 27, no. 3 (2021), 254-284
submitted: 20/2/2020, accepted: 2/2/2021, appeared: 28/3/2021

MuBeFE: Multimodal Behavioural Features Extraction Method

Alessia D'Andrea
(Istituto di Ricerca sulla Popolazione e le Politiche Sociali, Consiglio Nazionale delle Ricerche
Rome, Italy
https://orcid.org/0000-0002-3724-0222, alessia.dandrea@irpps.cnr.it)

Maria Chiara Caschera
(Istituto di Ricerca sulla Popolazione e le Politiche Sociali, Consiglio Nazionale delle Ricerche
Rome, Italy
https://orcid.org/0000-0002-3580-0505, mc.caschera@irpps.cnr.it)

Fernando Ferri
(Istituto di Ricerca sulla Popolazione e le Politiche Sociali, Consiglio Nazionale delle Ricerche
Rome, Italy,
https://orcid.org/0000-0001-9963-3315, fernando.ferri@irpps.cnr.it)

Patrizia Grifoni*
(Istituto di Ricerca sulla Popolazione e le Politiche Sociali, Consiglio Nazionale delle Ricerche
Rome, Italy
*Corresponding author
https://orcid.org/0000-0003-4298-5437, patrizia.grifoni@irpps.cnr.it)

Abstract: The paper aims to provide a method to analyse and observe the characteristics that distinguish the individual communication style such as the voice intonation, the size and slant used in handwriting and the trait, pressure and dimension used for sketching. These features are referred to as Communication Extensional Features. Observing from the Communication Extensional Features, the user's behavioural features, such as the communicative intention, the social style and personality traits can be extracted. These behavioural features are referred to as Communication Intentional Features. For the extraction of Communication Intentional Features, a method based on Hidden Markov Models is provided in the paper. The Communication Intentional Features have been extracted at the modal and multimodal level; this represents an important novelty provided by the paper. The accuracy of the method was tested both at modal and multimodal levels. The evaluation process results indicate an accuracy of 93.3% for the Modal layer (handwriting layer) and 95.3% for the Multimodal layer.

Keywords: Language, Behavioural communication features, Multimodal communication, personality traits
Categories: E.0
DOI: 10.3897/jucs.66375

1 Introduction

The dynamic exchange of information through different modalities, such as speech, handwriting, sketching etc., characterises human communication. During the

communication process, the features of the different modalities, such as the acoustic or sound intensity of the voice, the handwriting and sketching style are conceived as variables related to individual differences. This means that each person has a unique communication style. Suppose the communication style is closely linked to individual characteristics. In that case, this means that its analysis can allow extracting some users' information (referred in this paper as behavioural features) such as the communicative intention, emotions, personality traits, social style etc.

The extraction of communication styles allows achieving several aims in various contexts, such as:

- monitoring people with affect-related personality trait disorders within a healthcare context;
- preventing risk by detecting the drowsiness of a driver in an automotive context;
- developing personalised teaching practices to improve the effectiveness of learning processes in an educational context;
- evaluating candidates in working environments;
- detecting aggressive or dangerous behaviour in a security context.

Two research questions have been addressed by the paper: which characteristics distinguish the individual communication styles? How to extract the users' behavioural features?

The literature addresses these questions providing methods for extracting CIFs primarily on a single modality, which is frequently speech [Vicsi, 2008; Ya, 2011; Huang, 2019; Schuller, 2019]; however, human communication is intrinsically multimodal. According to Martin [Martin, 2002], "the mechanisms that underlie this multimodality of human communication are neither completely identified nor understood. Similarly, we do not know completely the behaviour a subject might have when facing a system, which allows him/her to use these different modalities of communication". Moreover, technological devices (mobile devices, sensors etc.) are increasingly designed to use numerous modalities. This fact is creating an interest in multimodality within human-machine communication processes and multimodal interaction systems [Caschera, 2007a; Caschera, 2007b]. Vigliocco [Vigliocco, 2014] highlighted the multimodal nature of language, making it clear that all modalities "are part and parcel of the same system and together constitute a tightly integrated processing unit, thus underscoring the need for a multimodal approach to the study of language". This suggests that a shift from specific applications addressing the extraction of behavioural features to a higher level of abstraction will allow the formalisation of human behaviour at a multimodal level (in both human-human and human-machine interaction processes).

To address that shift and to answer the previously cited research questions, this paper analyses, both at the modal and multimodal level, the characteristics of the individual communication style (such as the voice intonation, the size and slant used in handwriting and the trait, pressure and dimension used for sketching). In this paper, these individual characteristics are referred to as communication Extensional Features (CEFs). Starting from the identified CEFs, the user' behavioural features (e.g., the communicative intention, the social style and personality traits) have been extracted. The behavioural features are referred to as Communication Intentional Features (CIFs). The CIFs have been extracted both at the modal and multimodal level; this represents an important novelty provided by the paper. For the extraction of CIFs, a method based on Hidden Markov Models (HMMs) is provided in the paper. The method is used

because of its proven effectiveness in extraction and classification process [Grifoni, 2020a]. Besides, as stated in [Grifoni, 2020b], HMMs can represent differences in the whole structure of multimodal sentences managing multimodal features and incorporating temporal frequent pattern analysis for baseball event classification.

The current paper is structured as follows. In Section 2, this study's motivation is described and a scenario is detailed which aims to outline the research problem. In Section 3, the various methods used in the literature to extract multimodal behavioural features are described. Section 4 defines the attributes used to model CEFs and CIFs. Section 5 describes the MuBeFE method and Section 6 shows its training and testing at the modal and multimodal levels. Finally, Section 7 presents a discussion and conclusion for the paper.

2 Objectives

The extraction of CIFs which characterise communication processes allows the achievement of several purposes within different contexts, such as the detection of affect-related personality trait disorders in healthcare, safety in automotive applications, the effectiveness of learning within education, the evaluation of candidates in working environments and the detection of dangerous behaviour in a security context. As an example, we consider the educational context in this paper.

During research activities, the authors frequently collaborate with schools. Thus, it was in an educational context that they first began to discuss the problems and solutions provided by the MuBeFE method. In particular, during a collaboration with a secondary school, Maria, a professor of Italian literature, was interested in extracting information on students' personality traits and social styles to improve the students' learning processes effectiveness using personalised teaching practices.

To achieve this purpose has been configured an experimental setting in which students discussed a specific topic (the phenomenon of immigration in Europe).

During the discussion, each student was invited to talk about the topic using a touchscreen computer for handwriting and sketching. The student's behavioural features were acquired during the debate. During the discussion, one student used a multimodal sentence (shown in Fig. 1) consisting of the following spoken, sketched and handwritten components:

- spoken: "Questo grafico rappresenta le percentuali di Italiani e stranieri in Italia" (in English: "This pie chart represents the percentages of Italians and foreigners in Italy");
- handwritten: the labels of the pie chart represented the values of the percentages of Italians and foreigners within Italy (about 93% Italians and 7% foreigners);
- sketched: the student sketched a pie chart to represent the percentages of Italians and foreigners within Italy.

The first column of Figure 1 contains the modality used, and the other columns aggregate the elements of the multimodal sentence by the concept.

In this scenario, the features related to the different modalities in Figure 1 are examples of CEFs, as follows:

- the tone of voice used to articulate the sentence;

- the size and slant of the characters used in writing the words “italiani”, “stranieri”, “7%” and “93%”;
- The pie chart's traits and dimensions sketched to represent the percentages of Italians and foreigners in Italy.

Considering the CEFs described above, we aim to represent and extract features related to the student (such as the kind of sentence s/he pronounced, and his/her social style and personality). These features represent the CIFs.

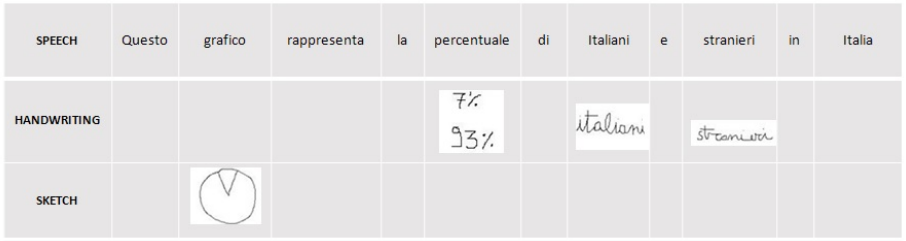


Figure 1: Multimodal sentence input elements

It should be noted that the touchscreen computer currently available on the market and used in this study do not allow the pressure values of a sketch to be acquired; therefore, we did not measure the pressure used in the sketch in this study. For this reason, the personality features, which are usually obtained by observing the pressure, traits and dimensions of the sketch, result in this case from analysis of only the traits and dimensions of the sketch. The speech modality features were acquired using the voice recorder of the device.

Let us consider the CEFs through an analysis of the multimodal sentence defined in Figure 1. Starting with the speech modality we consider the tone of voice used to articulate the sentence, as suggested in [D’Andrea, 2014].

Note that the syntactic structure of spoken Italian sentences does not contain enough significant information to identify the type of sentence (unlike in the English language); thus, in this case, it is important to analyse the tone of voice. LepschyLepschy, 1978] classified five-tone types which apply specifically to the Italian language: fall, rise, level, fall-rise and rise-fall. To identify the tone, it is necessary to individuate the segment of a sentence in which the accent is located. Halliday [Halliday, 1967] defined this segment, distinguishing the pre-tonic, i.e., “the part before the last sentence accent” concerning the tonic, i.e., “the remainder part” (Figure 2).

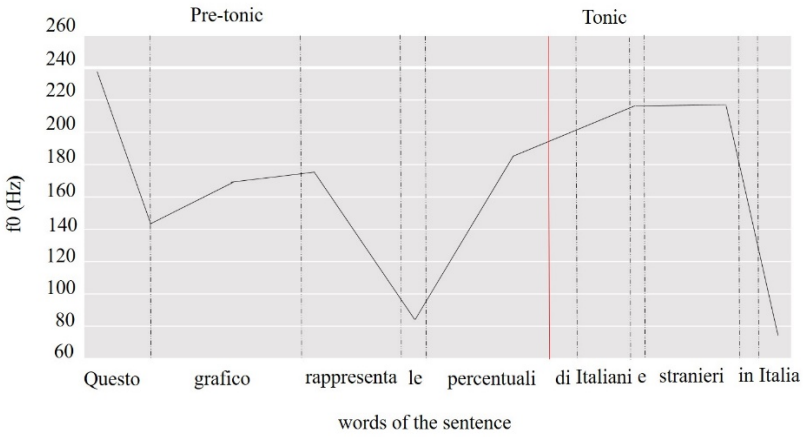


Figure 2: Pre-tonic and tonic parts of the spoken part of the sentence

The different tone types that characterise the tonic and pre-tonic part of each Italian sentence used for the analysis have been identified to define the intonation pattern (intpat) that characterises the different kinds of sentences. Intpat is defined as the sequence of the tone types that characterise both the tonic and pre-tonic part of the sentence. For instance, lets us suppose that a sentence has a pre-tonic part with a falling tone and a tonic part with a rising tone, the intapt of the sentence will be: (falling, rising).

To identify the dominant accent characterising the tonic part of the sentence, we used the PRAAT system and a plug-in that adds the INTSINT model (http://www.fon.hum.uva.nl/praat/) for analysing the spoken part. The PRAAT system allowed the dominant accent to be identified between the words "percentuali" and "di", where the maximum fundamental frequency, f0, and the maximum intensity are both located. The PRAAT system allows the pre-tonic part to be distinguished as a fall-rise tone characterises it, while the tonic part has a rise-fall tone.

For the handwriting modality, we consider two different parameters: size and slant. This is similar to the studies described by Merrill and Reid [Merrill,1981] and Rosario [Rosario, 2004].

The authors defined five different size classes: (i) tiny, (ii) small, (iii) average, (iv) large and (v) huge. With respect to the slant the authors identified six slants: (i) slant "A" (far left), (ii) slant "B" (less left), (iii) slant "C" (vertical), (iv) slant "D" (normal), (v) slant "E" (forward), (vi) slant "F" (far forward).

The handwritten part of the sentence was analysed using a handwriting recognition system developed by the Italian National Research Council. Four words of the handwriting were recognised: (i) the first word is "7%", with a value of "large" for the size attribute and a value of "vertical" for the slant attribute; (ii) the second word is "93%" with a value of "large" for the size attribute and a value of "vertical" for the

slant attribute; (iii) the third word is “italiani”, with a value of “large” for the size attribute and a value of “vertical” for the slant attribute; (iv) the fourth word is “stranieri”, with a value of “large” for the size attribute and a value of “vertical” for the slant attribute.

Finally, in the analysis of the sketched part of the multimodal sentence (see Figure 1), following Federici [Federici, 2005], we start from a consideration of the three most significant parameters: (i) traits, (ii) pressure, and (iii) dimensions.

The traits of a sketch can be classified as regular, irregular and point. The pressure can be classified as weak, very weak, strong, very strong, decreasing, and discontinuous. Finally, the dimensions can be classified as very large, large, small and very small. As stated above, the touchscreen computer does not allow pressure values to be acquired; therefore, we do not consider this parameter. The sketched part of the sentence was analysed using a sketch recognition system developed by the Italian National Research Council [Avola, 2010; Avola, 2007; Avola, 2006]. This system recognised both the object as a pie chart and its attributes (i.e., regular trait and large dimension).

The educational context represents only one of the potential applications of the MuBeFE method. As described above in the introduction, this method can also be applied to other contexts (healthcare, automotive and security) using behavioural features extraction for different purposes.

In the healthcare context, the extraction of behavioural features can be important for monitoring individuals' health conditions with personality trait disorders. These people often present changes in vocal acoustics and facial movements associated with psychomotor problems, which are behaviorally expressed as altered coordination and timing across motor-based properties [Williamson, 2014]. Clinically, attention paid to these behavioural changes can help monitor the disorder's course and responses to treatment, with a relatively low computational cost. The inclusion of these aspects in assessment improves measurement reliability and enables more finely-tuned interventions [Yang, 2013].

Concerning the automotive context, the modelling and monitoring of behavioural features allow the vehicle's security to be enhanced. A major cause of accidents is drowsiness while driving [Karchani, 2015]. An analysis of changes in facial expression, head movements, eye closure or constant yawning is important for real-time detection of drowsiness. The extraction of these behavioural changes allows driver fatigue detection systems to be improved to prevent many accidents [Tadesse, 2013].

Finally, in terms of the security context, the extraction of behavioural features, along with other information, is important in detecting aggressive behaviours such as those shown by terrorists. In this context, airports' security systems can use behavioural features to detect potential criminal actions [Ma, 2012].

In the following section, these various contexts and the methods used to extract behavioural features are described in detail.

3 Related works

Any user interaction activity (mouse pointing and clicking, keyboard usage, digital pen input, eye-gaze tracking, gesture input or any other kind of interactive input) may offer several behavioural features. These behavioural features allow the retrieval of an

appropriate and fine-grained user profile, providing personalised content, as well as recognition of its current status (e.g., aggressive behaviour) and different reactions according to an understanding of the user's emotional state, personality, social style and so on.

Many related studies have been carried out over the years to identify user behavioural features in specific contexts, such as healthcare, automotive applications, education, working environment and security

In a healthcare context, speech features (such as prosody) facial expression and body gestures are often associated with particular clinical disorders. Several works have found links between children with autism spectrum disorder (ASD) and atypicality in their prosody and facial expression such as the study provided in [Scheerer, 2020]in which faces and emotion word are analysed. To examine the relationship between the mean prosody accuracy values, the age, IQ, and social competence of the children, a hierarchical linear regression has been used. The correlation between autism spectrum disorder and body emotion identification has been analysed in [Metcalfe, 2019]. In the study children with and without autism spectrum disorder completed an emotion recognition task, that used dynamic stimuli. Processing style bias, autistic-like-traits and empathy have been measured. A multilevel logistic model was created with emotion recognition as the outcome variable. A model-based on machine learning, support vector machine and deep convolution neural network model is provided by [Zhao, 2020] to complete the facial expression recognition. The study involved normal and autistic children in testing the accuracy of the information system and the diagnostic effect of autism. Therefore, many attempts have been made to capture abnormal variations in prosodic, voice quality, and pronunciation characteristics in pathological speech. In [Almeida, 2019] the processing of voice signals has been investigated for detecting Parkinson's disease. The approach evaluates the use of eighteen feature extraction techniques and four machine learning methods to classify data obtained from sustained phonation and speech tasks. Attempts to capture abnormal variations in prosodic, voice quality and pronunciation characteristics in pathological speech have also been made in [Połap, 2019] where authors propose a method based on neural networks to evaluate voice problems. A method based on fuzzy inference systems to classify the Parkinson's patients as healthy or unhealthy analysing the voice features has been provided in [Sujatha, 2018]. Information theory, time-series modelling and statistical analysis have been used to analyse the differences in facial dynamics between children with autism spectrum disorder and their typically developing peers in [Guha, 2015]. While in Metallinou et al. [Metallinou, 2013] functional data analysis of facial motion to quantify the atypical characteristics of expression is provided. The authors uncovered patterns of expression evolution in both i) typically developing children and ii) children with high-functioning autism. A k-nearest neighbour algorithm is indeed provided by Koné et al. [Koné, 2015] to provide a multimodal emotion recognition based on data extracted from physiological signals, facial expressions and speech.

Modelling and monitoring human emotion have been also addressed within the automotive context. Many researchers have shown interest in monitoring driver behavioural characteristics within the automotive context for detecting fatigue during recent years. In [Bani, 2019] a method based on a Bayesian network that integrates the most relevant causes and effects of fatigue: sleep quality, road environment, and driving duration has been provided. As consequences, real-time facial expressions, such as

blinking, yawning, gaze, and head position have been analysed. A dynamic fatigue detection model based on HMM has been proposed in [Govardhan, 2018] for analysing static aspects of fatigue, integrated with relevant contextual information and spatially available sensory data. Also [Yan, 2018] use an HMM to estimate the driver's fatigue state reasonably. For the analysis, the eyes, mouth, and head posture under different mental conditions have been considered. In [Wang, 2018] a real-time fatigued detection has been carried out using an Active Shape Model (ASM) and a Support Vector Machine (SVM). The ASM has been used to detect the face and extract the Histogram of Orientation Gradient (HOG) features of mouth and eyes. Support Vector Machine (SVM) has been used for estimating the poses of the head. Based on the states of face, a fatigue decision index has been calculated.

A further relevant real-world setting (an example of which is provided in Section 2) is education. Several attempts have been made to address the extraction of behavioural features for learning purposes. These studies have focused on enabling the system to be more aware of the students' emotional and attentional expressions. In [Kapoor, 2005] dan HMM-based approach has been used to classify different levels of interest in children. For the analysis, students' postures facial features and head gesture information have been considered. Pivec et al. [Pivec, 2006] proposed a semantic-based solution for an adaptive e-learning framework. To observe students' learning activities in real-time eye movements for adaptive learning purposes have been monitored. More recent studies such as that provided by Minaee&Abdolrashidi [Minaee, 2019] proposed a system based on the attentional convolutional network to focus on important parts of face regions for detecting students' emotions. At the same time, a system for providing adaptive feedback based on the presence of students' confusion is developed in [Tiam-Lee, 2018]. Confusion is detected on students' compilations, typing activity, and facial expressions using a Hidden Markov Model trained.

At working environment context, the extraction of behavioural features is important for evaluating candidates' personality traits for a job position. An example is provided in [Güçlütürk, 2018] that consider audio-only, visual only, language only, audio-visual, and combination of audio-visual and language for predicting apparent personality traits of people. For the analysis, the machine-learning method has been used. In [Okada, 2019] a novel feature-extraction framework for inferring impression personality traits, emergent leadership skills, communicative competence, and hiring decisions is provided. To capture intermodal and interpersonal relationships explicitly as features, and efficient co-occurrence mining method is provided. In [Khalifa, 2018] body gestures in comparison to facial expressions have been analysed. An in-depth spatial-temporal approach that merges the temporal normalisation method with deep learning based on stacked auto-encoder for emotional body gesture recognition is used for the analysis.

Finally, in the context of security, the voice, facial expressions and posture give relevant information that facilitates the establishment of a suspect's guilt. In particular, posture plays an important role in security purposes, as underlined Piana et al. [Piana, 2014], who proposed an automatic emotions recognition based on a Support Vector Machine (SVM) classifier. Furthermore, the human voice is a source of information for understanding emotional states, as it allows for the detection of stress and supports automatic surveillance, emergency call centres and pilot/troop communications. In Calix et al. [Calix, 2012] an automated system methodology for detecting emotion from text and speech features has been developed. For emotion detection, corpora and

machine learning classification models are used to train and test the methodology. Lefter et al. [Lefter, 2011] defined a method to detect negative emotions, stress and aggression in speech data based on SVM, while Lefter [Lefter, 2014] used a Bayesian Network classifier in the automatic assessment of stress combined with an analysis of speech and gestures extracted from audio and video signals. Emotion analysis from text, audio and video has been addressed also in [Caschera, 2016] reviewing and evaluating the various techniques used for sentiment analysis and emotion recognition, and proposing an HMM-based approach to extract emotion from multimodal data.

Table 1 summarises the works cited above, classified according to the different contexts used, the input signal analysed, the methods applied, and the extracted multimodal behavioural features. From the analysis, it appears that all the revised methods can discriminate between and classify multimodal behavioural features by analysing data provided by different signals. Of these revised methods, HMMs and BNs appear to be the most appropriate methods for extracting behavioural features and dealing with the uncertainty inherent in the extraction of CEFs from CIFs. Indeed, HMMs and BNs can generate language items by listing sequences that fall into the group of sequences to be modelled. More specifically, both of these allow diagnostic reasoning to be carried out from effects (features related to the modalities used to communicate) to causes (features related to the users), which is the purpose of this paper. Both HMMs and BNs require a training process; however, a large dataset is not always available for the training phase. Unlike BNs, HMMs are simpler to train using a lower computational burden [Oliver, 2005]. HMMs require a smaller set of data for training and have proven flexibility.

Moreover, a multimodal input contains heterogeneous information, and the effectiveness of HMMs in managing flexible, complex, dynamic and heterogeneous (acquired from different modalities) information for stochastic processes has been demonstrated, for example in the case of interpretation and disambiguation of multimodal sentences [Caschera, 2013a; Caschera, 2007c; Caschera, 2008; Caschera, 2009]. HMMs allow sequences of structured data to be modelled since they can represent differences in the entire structure of multimodal sentences [Caschera, 2013b]. For example, they have been frequently applied to model and classify dialogue patterns [Twitchell, 2004].

Context	Signal	Method	Multimodal behavioural features	
			Emotions	*User's affective state*
Healthcare	*Video*	*Machine learning model support vector machine and deep convolution neural network*	*n/a*	[Zhao,2020]
		Functional data analysis	*n/a*	[Metallinou, 2013]
		Information theory, time-series modelling and statistical analysis	*n/a*	[Guha, 2015]
		Multilevel logistic model	[Metcalfe, 2019]	*n/a*
	Audio, video and physiological data	*K-nearest neighbour algorithm*	[Koné, 2015]	*n/a*
	Audio and video	*Convolution Neural Network*	[Hossain, 2019]	*n/a*
		Hierarchical linear regressions	[Scheerer, 2020]	*n/a*
	Audio	*Fuzzy inference system*	[Sujatha, 2018]	*n/a*
		Machine learning methods	*n/a*	[Almeida et al., 2019]
		Neural Networks	*n/a*	[Połap, 2019]
Automotive	*Video, audio and sensor data*	*Hidden Markov Model*	*n/a*	[Govardhan, 2018]
	Video	*Hidden Markov Model*	*n/a*	[Yan, 2018]
		Bayesian Network	*n/a*	[Bani, 2019]
		Support Vector Machine	*n/a*	[Wang, 2018]

Education	*Video*	*Semantic-based approach*	*n/a*	[Pivec, 2006]
		Hidden Markov Model	*n/a*	[Kapoor, 2005]
			n/a	[Tiam-Lee, 2018]
		Attentional convolutional network	[Minaee, 2019]	*n/a*
Working environment	*Video*	*Spatio-temporal approach*	*n/a*	[Khalifa, 2018]
	Audio and video	*Machine learning*	*n/a*	[Güçlütürk et al., 2018]
		Co-occurrence mining method	*n/a*	[Okada et al., 2019]
Security	*Video*	*Machine Learning*	[Calix, 2012	*n/a*
		Support Vector Machine	[Piana, 2014]	*n/a*
	Audio	*Support Vector Machine*	[Lefter, 2011]	*n/a*
	Audio and video	*Bayesian Network*	[Lefter, 2014]	*n/a*

Table 1: Examples of studies provided in the literature, classified according to the context used

Since the purpose of our paper is to define a model which is capable of being trained by both small and large datasets and heterogeneous datasets and to deal the complex structure of multimodal sentences, we apply HMMs in extracting the associations between CEFs (e.g., voice intonation in speech, size and slant of handwriting etc.) and CIFs (e.g., the user's personality, social style and types of sentences).

Before a description of the proposed HMM-based method is given, the following section provides the preliminary concepts needed to define the MuBeFE method.

4 Preliminary concepts

This section defines the sets of attributes that allow CEFs and CIFs (introduced in the previous sections) to be represented and instantiated. The proposed method uses attributes to represent the various characteristics of multimodal dialogue. CEFs and CIFs. In particular, the CEFs and CIFs are represented as attributes that update a multimodal attribute grammar according to the representation proposed by D'Andrea [D'Andrea, 2017]; this extends the multimodal attribute grammar defined in D'Ulizia [D'Ulizia, 2010].

CEFs and CIFs are represented respectively by qualitative synthesised attributes S(X) and qualitative inherited attributes I(X) as shown in Table 2.

CEFs			
S(X)	$S_{mod}(X)$	$S_{speech}(X)$	$S_{speech}(X)$ = {intpat, pre-tonic, tonic} = {fall-rise/rise-fall, fall-rise, rise-fall}
		$S_{handwriting}(X)$	$S_{handwriting}(X)$ = {size, slant} = {large, vertical}
		$S_{sketch}(X)$	$S_{sketch}(X)$ = {trait, dimension} = {regular, large}
	S_{mm} (X) = {fall-rise/rise-fall, fall-rise, rise-fall, large, vertical, regular, large}		
CIFs			
I(X)	I_{mod} (X)	$I_{speech}(X)$	$I_{speech}(X)$ = {typsen}
		$I_{handwriting}(X)$	$I_{handwriting}(X)$ = {socsty}
		$I_{sketch}(X)$	$I_{sketch}(X)$ = {perstype}
	$I_{mm}(X)$ = {typsen, socsty, perstype}		

Table 2: Qualitative synthesised attributes S(X) and qualitative inherited attributes I(X) representing Communication Extensional Features. (CEFs) and Communication Intentional Features (CIFs)

The set S(X) consists of two subsets: $S_{mod}(X)$ and $S_{mm}(X)$.

The set $S_{mod}(X)$ contains all the attributes necessary for analysis of the individual's communication features related to the specific modality. In the example provided in Section 3, three different modalities are considered: speech, handwriting and a sketch. In this case, three different $S_{mod}(X)$ are provided:

- $S_{speech}(X)$
  - where: $S_{speech}(X)$ = {intpat, pre-tonic, tonic} = {fall-rise/rise-fall, fall-rise, rise-fall};
- $S_{handwriting}(X)$
  - where: $S_{handwriting}(X)$ = {size, slant} = {large, vertical}
- $S_{sketch}(X)$
  - where: $S_{sketch}(X)$ = {trait, dimension} = {regular, large}

The set $S_{mm}(X)$ contains all the attributes needed to represent the CEFs:

S_{mm} (X) = {fall-rise/rise-fall, fall-rise, rise-fall, large, vertical, regular, large}

The set I(X) consists of two subsets: I_{mod} (X) and $I_{mm}(X)$.

I_{mod} (X) contains all the attributes needed to analyse the features of the individual's communication in terms of the specific modality (i.e., speech, handwriting and sketching):

- $I_{speech}(X)$
  - where: $I_{speech}(X)$ = {typsen}
- $I_{handwriting}(X)$
  - where: $I_{handwriting}(X)$ = {socsty}
- $I_{sketch}(X)$
  - where: $I_{sketch}(X)$ = {perstype}

where "typsen" identifies the type of sentence, "socsty" defines the social style and "perstype" the personality type.

Since the set I_{mm} (X) involves all three modalities (i.e., speech, handwriting and sketching), it contains all the attributes needed to represent the CIFs (i.e., the type of sentence, social style and personality type) that will be extracted by the MuBeFE method:

- $I_{mm}(X)$ = {typsen, socsty, perstype}

A detailed description of the qualitative synthesised attributes S(X) and the qualitative inherited attributes I(X) is provided in Section 6.
In the following section, the MuBeFE model is described.

5 Method

This section describes the MuBeFE method (see Figure 3), which allows the extraction of I(X), ($I_{mod1}(X)$, $I_{mod2}(X)$, $I_{mod3}(X)$... $I_{modn}(X)$), introduced in the previous section, from the S(X) ($S_{mod1}(X)$, $S_{mod2}(X)$, $S_{mod3}(X)$... $S_{modn}(X)$).

The MuBeFE method uses the following modules:

- a S(X) extraction module (CEFs EXTRACTION) that provides the functionality to extract the different features (e.g., tone, size, slant, trait etc.) from the modal and multimodal inputs (sentences); and
- an HMM-based module (HMM) that allows the I(X) (the type of sentence, user's personality etc.) to be extracted from the S(X).

In particular, the HMM-based module uses a method based on HMMs consisting of a Modal Layer (mod) and a Multimodal Layer (see Figure 3).

The Modal layer of the HMM takes into account the input coming from the different modalities (mod_1, mod_2, mod_3... mod_n) to extract the modal features $S_{mod1}(X)$, $S_{mod2}(X)$, $S_{mod3}(X)$... $S_{modn}(X)$, where the number of its feature's extraction modules (features $extraction_{mod1}$, features $extraction_{mod2}$, features $extraction_{mod3}$... features $extraction_{mode}$) is equal to the number *n* of modalities. The set of the modal features ($S_{mod1}(X)$, $S_{mod2}(X)$, $S_{mod3}(X)$... $S_{modn}(X)$) defines the set of the observation sequence of the HMM.

The extracted modal features are elaborated by the HMMs of the Modal Layer (HMM_{mod1}, HMM_{mod2}, HMM_{mod3}... HMM_{modn}) to extract the $I_{mod1}(X)$, $I_{mod2}(X)$, $I_{mod3}(X)$... $I_{modn}(X)$. The set of $I_{mod1}(X)$, $I_{mod2}(X)$, $I_{mod3}(X)$... $I_{modn}(X)$ defines the set of the hidden states of the HMM.

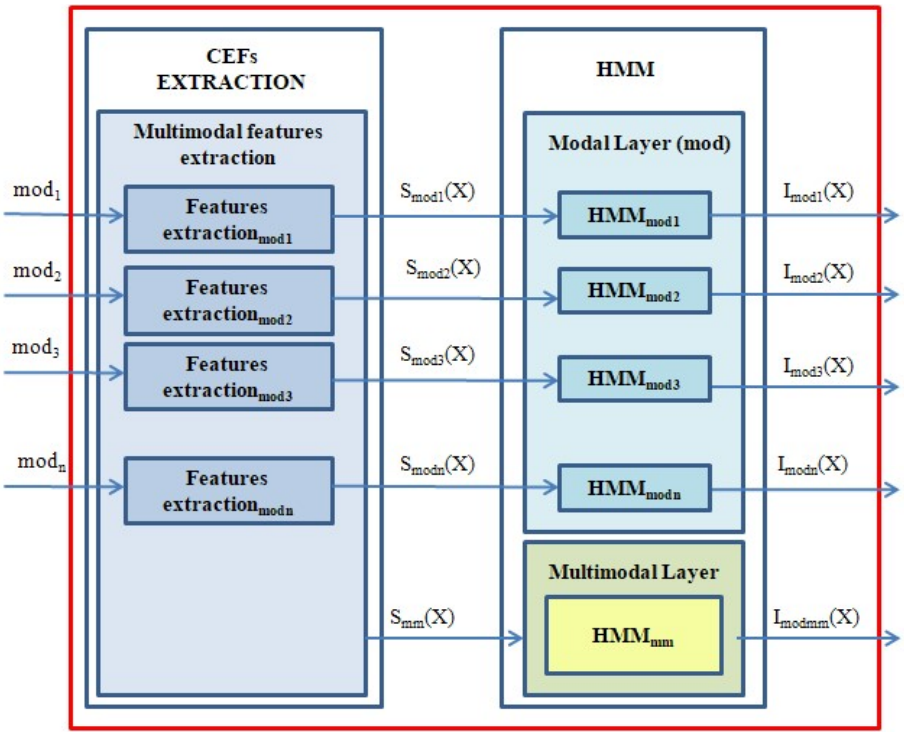


Figure 3: The MuBeFE method

Each HMM is characterised by transition probability matrix A, the output probability matrix B and the initial distribution vector π. The transition probability matrix contains the probability to have transitions from two hidden states q_{i_i} and q_{i_j} as defined in the following formula:

$$A=[a_{ij}] \text{with } a_{ij}=prob(q_{t+1}=s_j/q_t=s_i)\ \forall s_i,s_j \in I(X)=\{ I_{mod1}(X), I_{mod2}(X), I_{mod3}(X)... I_{modn}(X)\}$$

The output probability matrix B represents the output probability matrix that defines the probability that each state $s_i \in I(X)$ produces the observation FV_j:

$$B=[b_i(j)] \text{ with } b_i(j)=prob(o_t=qs_j/qi_t=s_i)\ \forall s_i \in Q_m, \forall\ qs_j \in S(X) =\{S_{mod1}(X), S_{mod2}(X), S_{mod3}(X)... S_{modn}(X)\}$$

Finally, π_I represents the initial distribution vector giving the probability that the state $i_i \in I\ (X)$ is the initial state of the sequence:

$$\pi=[\pi_i] \text{ with } \pi_i=prob(q_1=s_j)\ \forall s_i \in I(X)$$

To determine how well each HMM accounts for the observations by computing $Prob(S(X)/HMM_{modn})$, we use the Forward algorithm [Rabiner, 1989]. The transition probability matrix, the output probability matrix and the initial distribution vector are estimated by using the Baum-Welch algorithm [Benesch, 2001]. This training procedure is used to tune the parameter of the model to obtain probabilities, considering the observation sequences S(X). The purpose of the HMM is to identify the model having the highest probability to generate the observed output, given the parameters of the model using the Viterbi algorithm [Viterbi, 1967] for calculating the best sequence of HMM states that generates S(X).

For the sake of clarity, we apply this method to the different modalities which compose the multimodal sentence in Figure 1 and the S(X) and I(X) introduced in the example described in Section 3 and detailed in Section 6.

Figure 4 describes the method applied to the speech modality, in which the $I_{speech}(X)$ extraction module (i.e., the PRAAT system) is applied to extract pre-tonic and tonic features from the spoken part of the multimodal sentence. The HMM-based module categorises the features extracted from the different types of the sentence (i.e., statements, questions, exclamations, commands etc.) as described in Section 6.

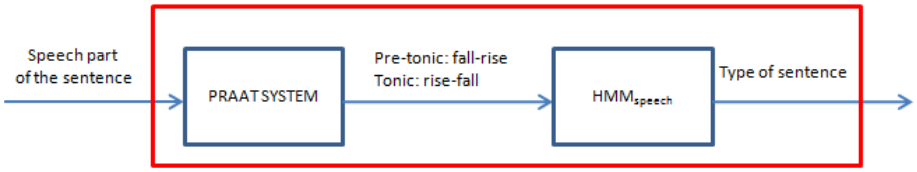


Figure 4: The MuBeFE method applied to the speech modality

Figure 5 shows the MuBeFE method applied to the handwriting modality. The $S_{handwriting}(X)$ extraction module (i.e., the handwriting recogniser) returns the size and slant features of the four handwritten phrases within the multimodal sentence. The HMM-based module links these extracted features to the social styles (expressive, driver, analyst and amiable) described in Section 6.

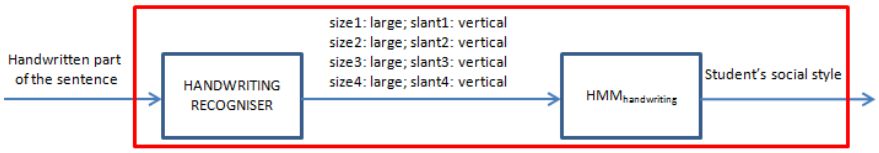


Figure 5: The MuBeFE method applied to the handwriting modality

Finally, Figure 6 shows the method applied to the sketch modality. In this case, the $S_{sketch}(X)$ extraction module (i.e., the sketch recogniser) returns the traits and dimensions of the sketched part of the multimodal sentence, while the HMM-based

module associates these features with the various types of students' personalities (i.e., rational, emotional, sociable, aggressive, etc.) as described below in Section 6.

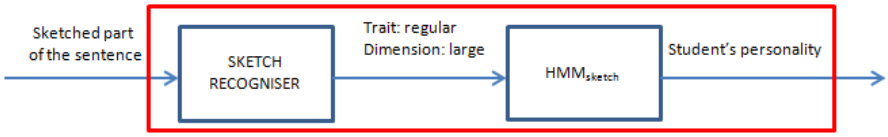


Figure 6: The MuBeFE method applied to the sketch modality

The Multimodal layer elaborates the combination of the input coming from the different modalities (mod_1, mod_2, mod_3... mod_n); the multimodal features extraction module (Multimodal features extraction) extracts the $S_{mm}(X)$ (i.e., $S_{speech}(X)$, $S_{handwriting}(X)$, $S_{sketch}(X)$), which is taken as input in the Multimodal HMM to extract the $I_{mm}(X)$ at the multimodal level (as shown in Figure 7).

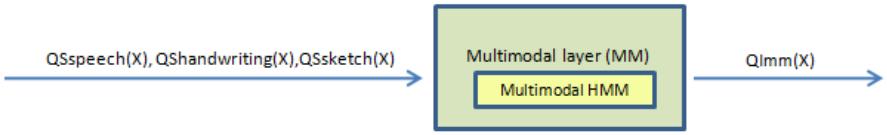


Figure 7: Instance of the Multimodal layer

The Multimodal Layer takes advantages from the associations between the $S_{modi}(X)$ and the $I_{modi}(X)$ for i=1,2,3..n (defined at Modal layer). When a synchronous multimodal input occurs, and when, for example, the handwriting input is $S_{handwriting}(X)$ (size and slant), we know that the output will be $I_{handwriting}(X)$ (the user's social style). In this situation, we can collect information on all the other input modalities, as an example, $S_{speech}(X)$ and $S_{sketch}(X)$ and their coordination features, associating their features with $I_{handwriting}(X)$ (the user's social style). Thus, unlike the Modal Layer, the Multimodal Layer allows social styles to be obtained not only from users' handwriting, but from any other modality or combination of modalities (such as speech, sketch, or speech and sketch, etc.), thereby improving the probability of obtaining a correct association when the number of observations increases. The differences between the two layers will be clarified in the following sections by describing the training and testing processes of the method at the Modal Layer (i.e., for handwriting) and the Multimodal Layer.

6 Training and testing of the MuBeFE method at the Modal and Multimodal Layers

The MuBeFE method described in the previous section was trained and tested on a dataset at both the Modal (for handwriting) and Multimodal layers, as described in the following sections.

6.1 Data collection

A convenience sample of fifty Italian individuals (thirty-two males and eighteen females between twenty-five and sixty years old) was considered to build the dataset. The average age of people was 40.08 and the standard deviation was 12.27.

The behavioural features of these individuals were recorded by acquiring video/voice signals, handwriting and sketch inputs by a touchscreen computer in sessions of 30 minutes. The speech modality features were acquired using the voice recorder of the computer. The spoken part has been analysed using the PRAAT system together with and a plug-in that adds the INTSINT model (http://www.fon.hum.uva.nl/praat/). The handwriting and sketched parts of the sentence was analysed using a sketch recognition system developed by the Italian National Research Council [Avola, 2010; Avola, 2007; Avola, 2006]."

In the first stage, each participant was provided with a list of sentences to be spoken aloud. Each participant, one at a time, spoke aloud twelve sentences (three for each type of class: statements, questions, exclamations and commands).

In the second stage, each participant was asked to choose the social style that best represented him/her from a list prepared by the researcher. The list contained several types of social styles extracted from the work conducted by Merrill and Reid [Merrill, 1999] and Rosario [Rosario, 2004]. Participants were then asked to write six sentences using a touchscreen computer. A recorded voice was used to convey to the users the sentences they were asked to write, to ensure identical conditions for all participants. Finally, each participant was asked to choose the personality type that best described his/her personality from a list prepared by the researcher. The list contained various types of personalities extracted from Federici's work [Federici, 2005]. Participants were then asked to draw ten simple objects using the touchscreen computer. A recorded voice was also used at this stage to convey to the users the objects that were to be drawn, thus recreating the same conditions for all users.

A total of 600 spoken sentences (twelve sentences pronounced by fifty people), 300 handwritten sentences (six sentences written by fifty people) and 500 sketches (ten sketches by fifty people) were recorded. To perform the cross-validation, we organised the 3 datasets of collected data for the three modalities used as follows:

- 300 samples for training (performing an incremental training from 30 to 300 samples) and 300 samples for testing of the MuBeFE in terms of the speech modality;
- 150 samples for the training (performing an incremental training from 15 to 150 samples) and 150 samples for the testing of the MuBeFE in terms of the handwriting modality;
- 250 samples for the training (performing an incremental training form25 to 250 samples) and 250 samples for the testing of the MuBeFE in terms of the sketching modality.

Participants were provided with a set of multimodal sentences for the multimodal data collection. Fifty Italian individuals were involved; each participant was provided with a list of multimodal sentences, and participants were asked to choose features concerning different kinds of sentences, social styles and personalities. Six multimodal sentences were selected from the samples used by Caschera et al. [Caschera, 2013a]. Three hundred samples were considered (six multimodal sentences for each of the 50

individuals) of which 150 samples were used for training (performing an incremental training form15 to 150 samples) and 150 samples for testing.

To clarify the process of data collection, we provide an example at the modal level that concerns the social styles of the participants associated with the handwriting modality (as shown in Table 3).

SOCIAL STYLE	%	MALES	FEMALES
amiable	42	8	13
analyst	26	8	5
expressive	18	5	4
driver	14	4	3

Table 3: Social styles selected by users

During the collection process, 42% of users selected the "amiable" social style. Another significant proportion of the users (26%) selected the "analyst" social style, while 18% of the users selected the "expressive" social style. Only 14 % selected the "driver" social style. Starting from the values of size and slant provided by the handwriting recogniser, various associations between the four selected social styles and the handwriting features (size/slant) were detected (as shown in Table 4). Concerning the "amiable" social style, four different associations resulted from the analysis: (i) small/forward (43% of users) (ii) small/normal (29% of the users) (iii) tiny/normal (19% of the users) and (iv) tiny/far-forward (9% of the users).

TYPE OF SOCIAL STYLE	FIRST		SECOND		THIRD		FOURTH	
amaible	**size**	**slant**	**size**	**slant**	**size**	**slant**	**size**	**slant**
	Small	Forward	Small	Normal	Tiny	Normal	Tiny	Far-forward
	43% of the users		29% of the users		19% of the users		9% of the users	
analyst	Tiny	Less-left	Tiny	Far-left	Small	Forward		
	54% of the users		31% of the users		15% of the users			
expressive	Huge	Far-Forward	Large	Forward	Small	Normal		
	56% of the users		33% of the users		11% of the users			
driver	Large	Vertical	Large	Less-left				
	71% of the users		29% of the users					

Table 4: Associations between social styles and modal features

For the "analyst" and "expressive" social styles, three different associations were extracted: for the "analyst" (i) tiny/less-left (54 % of the users), (ii) tiny/far-left (31 %

of the users), and (iii) small/forward (15 % of the users); for the "Expressive" (i) huge/far-forward (56 % of the users), (ii) large/forward (33 % of the users), (iii) small/normal (11 % of the users). Finally, two different associations for the "Driver" social style resulted from the analysis: (i) large/vertical (71 % of the users), and (ii) large/less-left (29 % of the users).

6.2 Training of the Modal Layer of the MuBeFE method

As described in the previous section, data collected at the modal level were used to train the Modal Layer.

In particular, the training set considered both the S_{mod} (X) and the I_{mod}(X). The association between the values of the set of S_{mod} (X) and the classes defined for the set I_{mod}(X) is formalised using the following expression:

$$S_{mod}(X) / I_{mod}(X)$$

As an example, we describe the training process of the Mod layer for the handwriting modality. In this case, the purpose of the MuBeFEmethodis to recognise the user's social style starting from handwriting features (size and slant), and 150 samples were used to train the handwriting layer.

The attributes of the set $I_{handwriting}$(X), which refers to the social styles attribute, were extracted from the handwriting samples. In particular, the domain of these attributes is the following:

$$I_{handwriting}(X) = \{\text{expressive, amiable, driver, analyst}\}$$

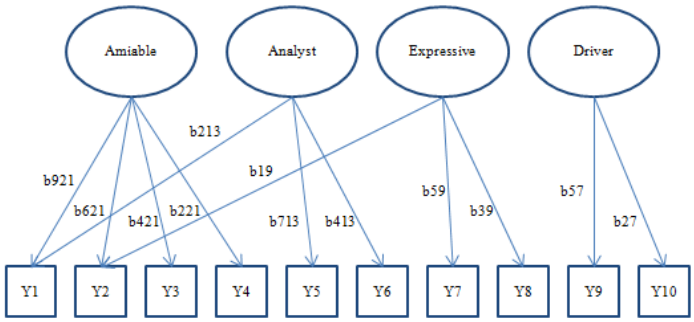
Moreover, the set $S_{handwriting}$ (X) attributes, extracted from the handwriting samples, represent the set of observation symbols for the handwriting layer. More precisely, the $S_{handwriting}$ (X) set involves the values of the *size* attribute, which expresses the dimensions of the handwriting traits, and the *slant* attribute, which refers to the writing direction:

$$D_{size} = \{\text{tiny, small, average, large, huge}\}$$

$$D_{slant} = \{\text{far left, less left, vertical, normal, forward, far forward}\}$$

Each observation is composed of a sequence of symbols which refers to all possible combinations of the instances belonging to the D_{size} and D_{slant}sets.

$S_{handwriting}$ (X)= {tiny, small, average, large, huge, far left, less left, vertical, normal, forward, far forward}



Legend

Y1 indicates the couple: (small, forward)
Y2 indicates the triple: (small, normal)
Y3 indicates the triple: (tyni, normal)
Y4 indicates the triple: (tyni, far-forward)
Y5 indicates the triple: (tyni, less-left)
Y6 indicates the couple: (tyni, far-letf)
Y7 indicates the couple: (huge, far-forward)
Y8 indicates the couple: (large, forward)
Y9 indicates the couple: (large, vertical)
Y10 indicates the couple: (large, less-left)

Figure 8: Example of observation sequences and hidden states in the HMM for handwriting

Following this, the associations between the different social styles ($I_{handwriting}(X)$) and the captured handwriting features ($S_{handwriting}(X)$) (a pair composed of size and slant) were determined (as shown in Table 3).

In detail, the training process has had the purpose of capturing the associations between a couple of values (size, slant) and the four classes of the *socsty* attribute (amiable, analyst, expressive, driver) as described in the previous section and illustrated in Figure 5.

		1	**2**	**3**	**4**
		amiable	**analyst**	**expressive**	**driver**
1	**amiable**	0.25	0.25	0.25	0.25
2	**analyst**	0.25	0.25	0.25	0.25
3	**expressive**	0.25	0.25	0.25	0.25
4	**driver**	0.25	0.25	0.25	0.25

Table 5: Matrix A of the HMM-based method

The starting values assigned to the different matrices of the HMM are shown in Tables 5, 6 and 7. Matrix A has all values equal to 0.25 since the probability at the beginning is equally distributed throughout the hidden states.

		1	2	3	4
		amiable	**analyst**	**expressive**	**driver**
1	**(small, forward)**	0.43	0.15	0	0
2	**(small, normal)**	0.29	0	0.11	0
3	**(tiny, normal)**	0.19	0	0	0
4	**(tiny, far-forward)**	0.9	0	0	0
5	**(tiny, less-left)**	0	0.54	0	0
6	**(tiny, far-left)**	0	0.31	0	0
7	**(huge, far-forward)**	0	0	0.56	0
8	**(large, forward)**	0	0	0.33	0
9	**(large, vertical)**	0	0	0	0.71
10	**(large, less-left)**	0	0	0	0.29

Table 6: Matrix B of the HMM-based method

The values b_{ij} of matrix B are instantiated according to experimentally obtained data. As in the case of matrix A in Table 6, the probability is equally distributed throughout the hidden states.

	1	2	3	4
	amiable	**analyst**	**expressive**	**driver**
π	0.25	0.25	0.25	0.25

Table 7: Matrix Π of the HMM-based method

Matrices A, B and π were updated by training the method using the dataset of 150 handwriting sentences collected from fifty Italian individuals (as described above).

This dataset contained associations between all the 150 handwriting sentences and the four social styles (expressive, amiable, driver and analyst).

6.3 Training of the multimodal layer of the MuBeFE method

The same process was carried out at the Multimodal layer, were for the training set, the associations between the values of the set S_{mm} (X) and the classes defined for the set I_{mm}(X) are considered and formalised using the following expression:

$$S_{mm}\ (X)/\ I_{mm}(X).$$

As an example, we consider the multimodal sentence (X) composed of the input elements shown in Figure 1 where the student said: *"Questograficorappresenta le percentuali di Italiani e stranieri in Italia*" (in English: "This pie chart represents the percentages of Italians and foreigners in Italy"). In handwriting, s/he produced labels for the pie chart representing the values of the percentages of Italians and foreigners in Italy (about 93% Italians and 7% foreigners), and by sketching, s/he drew a pie chart representing these percentages of Italians and foreigners in Italy.
The observation sequence S_{speech} (X) therefore contains the following:

$$S_{speech}\ (X) = \{\text{fall-rise, rise-fall}\}.$$

For the handwritten part of the sentence, the recognition system recognised (i) the first word as "7%" and is assigned the value *large* to the *size* attribute and the value *vertical* to the slant attribute; (ii) the second word as "93%" and assigned the value *large* to the *size* attribute and the value *vertical* to the slant attribute; (iii) the third world as "Italiani" and assigned the value *large* to the *size* attribute and the value *vertical* to the slant attribute; (iv) the fourth word as "stranieri" and assigned the value *large* to the *size* attribute and the value *vertical* to the slant attribute.
Therefore, the observation sequence $S_{handwriting}$ (X) is the following:

$$S_{handwriting}\ (X) = \{\text{large, vertical, large, vertical, large, vertical, large, vertical}\}$$

Finally, in the sketching part, the sketch recogniser returned the object as a pie chart and attributed the value *regular* to the trait feature, no value to the pressure (as described in Section 3) and the value *large* to the dimension feature.

Therefore, the observation sequence S_{sketch} (X) is the following:

$$S_{sketch}\ (X) = \{\text{regular, not\_available, large}\}$$

The multimodal behavioural features (S_{speech} (X), $S_{handwriting}$ (X), S_{sketch} (X)) extracted from each modality are used to train the method in terms of the associations between these and the classes of the *different kinds of the sentence*, *social style s*and *personalities* attributes (i.e., statement, question, exclamation, command, amiable, analyst, expressive, driver, rational, emotional, sociable, aggressive), where the sets S_{mm} (X) and I_{mm} (X) are respectively:

$$S_{mm}\ (X) = \{D_{pre\text{-}tonic}, D_{tonic,}\ D_{size}, D_{slant,}\ D_{trait}, D_{dimension}\}$$

$$I_{mm}(X) = \{\text{statement, question, exclamation, command, amiable, analyst expressive, driver, rational, emotional, sociable, aggressive}\}$$

The values of the features comprising the S_{mm} (X) were extracted from the collected data, as described in Section 6.1.
As described in the example of the handwriting modality, the estimation of the HMM parameters A, B and π was performed using the Baum-Welch algorithm [Rabiner,

1989] by using as a training set the 150 samples of correct association between the values belonging to the set S_{mm} (X) and the values of the set I_{mm} (X) (i.e., statements, questions, exclamations, commands, amiable, analyst expressive, driver, rational, emotional, sociable, aggressive). In this method, the observation sequences (S_{mm} (X)) are composed of all the possible combinations of values that the parameters can assume from the reference domains (i.e., pre-tonic, tonic, fall, rise, level, fall-rise, rise-fall, tiny, small, average, large, huge, far left, less left, vertical, normal, forward, far forward, regular, irregular, point, weak, very weak, strong, very strong, decreasing, discontinuous, very large, large, very small, small).

In this layer, features connected to the speech modality (types of sentence, i.e. statements, questions, exclamations, commands) can also be associated to the handwriting and sketch modalities, as well as features connected to the handwriting modality (social styles, i.e. expressive, driver, analyst and amiable) to the speech and sketch, and features connected to the sketch modality (types of students' personalities, i.e. rational, emotional, sociable, aggressive) to the speech and handwriting,

To clarify these cross associations, e we evaluate the multimodal layer using as observation sequence S_{mm} (X) the observation sequences associated with the multimodal sentence (X) in the example given in Figure 1.
Therefore, the observation sequence S_{mm} (X) is the following:

S_{mm} (X) = {fall-rise, rise-fall, large, vertical, large, vertical, large, vertical, large, vertical, regular, not_available, large}

The purpose of the evaluation process is to extract a subset of the social styles, types of sentence and types of students' personalities features, therefore, the set of considered hidden states is composed as follows:

I_{mm}(X) = {statement, driver, aggressive}.

6.4 Testing of the MuBeFE method at the modal and multimodal layers

Following the training process, the MuBeFE method was tested at both Modal and Multimodal layers. The testing process aimed to obtain the correct association between the qualitative synthesised attributes (S_{mod} (X) for the Modal layer and S_{mm}(X) for the multimodal one) and the qualitative inherited attributes for both the Modal (I_{mod} (X)) and the Multimodal (I_{mm}(X)) layers.
The considered performance evaluation measure is accuracy (A_i) that measures the fraction of the correctly classified qualitative synthesised attributes in the considered qualitative inherited attribute among the retrieved qualitative inherited attributes.
For the handwriting modality, 150 samples were used among the values extracted from the 300 case study samples to create the test set.
The performance of the MuBeFE method for the handwriting modality was evaluated in terms of accuracy to evaluate the correct association between the $S_{handwriting}$ (X) and the $I_{handwriting}$ (X), which is measured as:
For the Modal layer, we consider the handwriting modality, and the evaluation measure is defined as follows [Manliguez, 2016]:

$$A_{handwriting} = \frac{\sum_{k=expressive}^{analyst} x_{kk}}{T_{Nh}}$$

with k ϵ KandK={expressive, amiable, driver, analyst} and with T_{Nh} as the total number of samples

Table 8 provides the summary of the experiments and, in particular, presents the normalised handwriting confusion matrix performed on the 150samples associated to the four *social styles* attribute (expressive, amiable, driver, analyst).

		Output			
		expressive	amiable	driver	analyst
Actual	expressive	0,97	0,00	0,03	0,00
	amiable	0,03	0,92	0,03	0,03
	driver	0,00	0,05	0,92	0,03
	analyst	0,03	0,00	0,03	0,95

Table. 8: Confusion matrix of the handwriting layer for the social style

Figure 9 provides the incremental accuracy rate for different amounts, from 15 to 150 samples, of data for the handwriting layer. This figure shows that the learning rate improves when the amount of data in the training set increases, and more precisely, the accuracy rate increases from 66.6% for the first fifteen pairs of values of the training set (size, slant) to 93.3% for all 150 pairs of values (size, slant) of the training set.

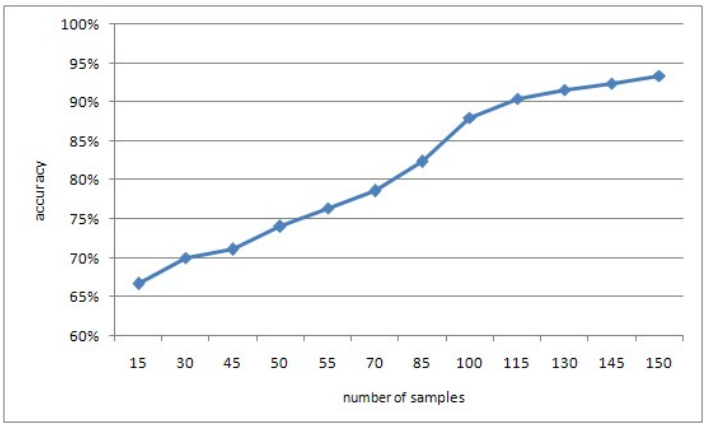


Figure 9: Accuracy rates for different amounts of data for the handwriting layer

The MuBeFE method was also evaluated in terms of accuracy at the Multimodal layer. The test set was composed of features extracted from multimodal sentences of the 150 samples extracted from the test set used by Caschera*et al*. [Caschera, 2013a]. For the Multimodal layer, the evaluation measure is defined as follows:

$$A_{multimodal} = \frac{\sum_{z=statment}^{aggressive} x_{zz}}{T_N}$$

with $z \epsilon Z$ and $Z=\{statement, driver, aggressive\}$ and with T_N as the total number of samples

Table 9 provides the summary of the experiments for the multimodal layer and, in particular, presents the normalised confusion matrix performed on the 150 multimodal samples associated to the subset of the social styles, types of sentence and types of students' personalities features (statement, driver, aggressive).

		Actual		
		statement	driver	aggressive
Output	statement	0,96	0,04	0,00
	driver	0,02	0,94	0,04
	aggressive	0,02	0,02	0,96

Table 9: Confusion-matrix of the multimodal layer

Figure 10 provides the incremental accuracy rate for different amounts, from 15 to 150 samples, of data for the Multimodal layer.

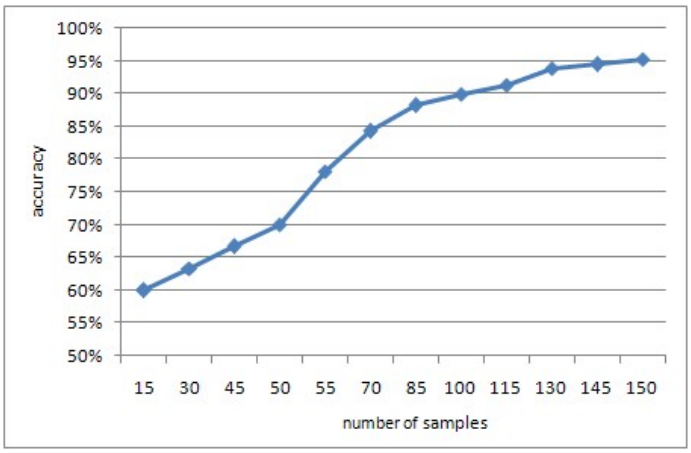


Figure 10: Accuracy rates for different amounts of data for the Multimodal layer

The method for the Multimodal layer can incrementally learn when the amount of data in the training set increases from fifteen to 150 samples. In this case, the accuracy rate improves from 60% to 95.3% (see Figure 10).

In summary, when the number of training samples is increased, the rate of improvement for the Modal layer becomes higher than that for the Multimodal layer. This is because the Multimodal layer has to manage a set of heterogeneous data that includes the different features of the considered modalities and to learn a greater number of connections than the Mod layer. Thus, achieving greater accuracy requires

a larger dataset to train the method to associate sequences of more complex data correctly.

7 Conclusions

The proposed MuBeFE method meets the need for a shift from specific applications to a higher level of abstraction during the extraction of specific behavioural features from messages exchanged in the communication process. This leads to the extraction of higher quality and more reliable behavioural information than information obtained from a single modality, allowing the formalisation of human behaviour at a multimodal level. The advantage is two-fold. First, since the modalities are usually complementary, the result of multimodal behaviour extraction is more informative than for each of the modalities individually. The second advantage is that since modalities are not always reliable if one modality becomes corrupted it is possible to extract the missing behavioural information from another. Moreover, the MuBeFE method allows the extraction of multimodal behavioural features at a higher level of abstraction, responding to the need for extraction of multimodal behavioural features in different contexts.

The management of complex and heterogeneous information in a flexible way has been addressed using a linguistic approach combined with a method based on HMMs, which has proven effectiveness in managing flexible, complex, dynamic and heterogeneous information for stochastic processes.

We decided to use HMMs because they achieve good classification accuracy on multi-dimensions and discrete or categorical features; therefore, they allow dealing the sequence of structured data of the multimodal sentences we need to consider. Therefore, HMMs are well suited for the purposes of this paper, which are to build a method aiming to automatically characterise communication processes and to progressively learn the dynamic features of the communication processes. In addition, HMM obtains a good level of accuracy even with small datasets; therefore, we chose to use this method as it is more widely applicable. When larger datasets are available, the method achieves a lower accuracy level compared to other methods.

The accuracy of the method was tested at both modal and multimodal levels. The results of the evaluation process indicate an accuracy of 93.3% for the Modal layer (handwriting layer) and 95.3% for the Multimodal layer.

In this paper, the method has been applied to the educational context to extract information on students' social styles and personality traits, to develop personalised teaching practices to improve the effectiveness of learning processes in education. In future work, a large-scale test of the environment will be developed for other contexts, to validate the efficiency of the MuBeFE method for other purposes such as the monitoring of the mental conditions of people with affect-related personality trait disorders in the healthcare context, enhancing the security of a vehicle by detecting drowsiness in the driver in the automotive context, and detecting aggressive behaviour in the security context.

References

[Almeida, 2019] Almeida, J. S., Rebouças Filho, P. P., Carneiro, T., Wei, W., Damaševičius, R., Maskeliūnas, R., & de Albuquerque, V. H. C. (2019). Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques. Pattern Recognition Letters, 125, 55-62.

[Avola, 2006] Avola D., Caschera M. C. and Grifoni P. (2006). Solving Ambiguities for Sketch-Based Interaction in Mobile Environments, OTM 2006 Ws, Part I, LNCS 4277, Springer-Verlag, 904-915.

[Avola, 2007] Avola D., Caschera M. C., Ferri F. and Grifoni P. (2007). Ambiguities in Sketch-Based Interfaces, 40th Annual Hawaii International Conference on System Sciences (HICSS'07). IEEE Computer Society. pp. 290.

[Avola, 2010] Avola D., Caschera M. C., Ferri F. and Grifoni P. (2010). Classifying and Resolving Ambiguities in Sketch-Based Interaction, International Journal of Virtual Technology and Multimedia. 1(2): 104-139.

[Bani, 2019] Bani, I., Akrout, B., & Mahdi, W. (2019). Real-Time Driver Fatigue Monitoring with a Dynamic Bayesian Network Model. In Digital Health Approach for Predictive, Preventive, Personalised and Participatory Medicine (pp. 69-77). Springer, Cham

[Benesch, 2001] Benesch, T.: The Baum-Welch algorithm for parameter estimation of Gaussian autoregressive mixture models. J. Math. Sci. (New York), 105, 2001, pp. 2515-2518.

[Calix, 2012] Calix, R. A., Javadpour, L., & Knapp, G. M. (2012). Detection of affective states from text and speech for real-time human–computer interaction. Human factors, 54(4), 530-545.

[Caschera, 2007a] Caschera M. C., Ferri F. and Grifoni P. (2007a): Multimodal Interaction Systems: Information and Time Features, International Journal of Web and Grid Services (IJWGS), 3(1): 82-99.

[Caschera, 2007b] Caschera M. C., Ferri F. and Grifoni P. (2007b). Multimodality in Mobile Applications and Services, Encyclopedia of Mobile Computing and Commerce (2 Volumes), ed. David Taniar, Monash University, Australia, pp. 675-681.

[Caschera, 2007c] Caschera M. C., Ferri F. and Grifoni P. (2007c). An Approach for Managing Ambiguities in Multimodal Interaction, OTM 2007 Ws, Part I, Springer Publishing, LNCS 4805, 387–397.

[Caschera,2008] Caschera M. C., Ferri F. and Grifoni P. (2008). Ambiguity Detection in Multimodal Systems. Advanced Visual Interfaces, AVI 2008. ACM Press 2008. pp. 331-334.

[Caschera,2009] Caschera M.C. (2009). Interpretation Methods and Ambiguity Management in Multimodal Systems, In: Handbook of Research on Multimodal Human Computer Interaction and Pervasive Services: Evolutionary Techniques for Improving Accessibility. P. Grifoni, editor. IGI Global (USA). pp. 87-102.

[Caschera, 2013a] Caschera, M.C., Ferri, F. and Grifoni, P. (2013a). InteSe: An Integrated Model for Resolving Ambiguities in Multimodal Sentences, IEEE Transactions on Systems, Man, and Cybernetics: Systems, Volume: 43, Issue: 4, pp. 911 - 931.

[Caschera, 2013b] Caschera, M. C., Ferri, F. and Grifoni, P. (2013b). From Modal to Multimodal Ambiguities: a Classification Approach, JNIT, 4(5): 87-109. arXiv preprint arXiv:1704.02841

[Caschera, 2016] Caschera, M. C., Ferri, F. and Grifoni, P. (2016). Sentiment analysis from textual to multimodal features in digital environments. Proceedings of the 8th International Conference on Management of Digital EcoSystems (MEDES) November 2016. Pp. 137-144. https://doi.org/10.1145/3012071.3012089

[D'Andrea, 2014] D'Andrea, A., Ferri, F. and Grifoni, P. (2014). Prosodic Analysis: An Italian Case Study, International Journal of Language Studies, 8(4): 107-126.

[D'Andrea, 2017] D'Andrea, A., D'Ulizia, A., Ferri, F. and Grifoni, P. (2017). EMAG: An Extended Multimodal Attribute Grammar for Behavioural Features. Digital Scholarship in the Humanities, 32 (2), pp. 251-275. http://dsh.oxfordjournals.org/content/by/year/2015 doi: 10.1093/llc/fqv064

[Datcu, 2009] Datcu, D. and Rothkrantz, L. (2009). Multimodal Recognition of Emotions in Car Environments, In Driver Car Interaction and Interface 2009. Praag and Czech Republic.

[Denmark, 2019] Denmark, T., Atkinson, J., Campbell, R., & Swettenham, J. (2019). Signing with the face: Emotional expression in narrative production in deaf children with autism spectrum disorder. Journal of autism and developmental disorders, 49(1), 294-306

[D'Ulizia, 2010] D'Ulizia A., Ferri F. and Grifoni P. (2010). Generating Multimodal Grammars for Multimodal Dialogue Processing, IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 40(6): 1130-1145.

[Federici, 2005] Federici, P. (2005). GliAdulti di Fronte ai Disegnidei Bambini, Manuale di Interpretazione del Disegno per Educatori e Operatori. Franco Angeli, Milano.

[Govardhan, 2018] Govardhan, S.D., Raghul, S., Selvakumar, S., Pavithra, M., Niranjana, B. (2018). Driver Monitoring System Usi Ng Hidden Markov Model. International Journal of Recent Trends in Engineering & Research (IJRTER). Special Issue; March - 2018 [ISSN: 2455-1457] DOI : 10.23883/IJRTER.CONF.02180328.049.TY8AS

[Grifoni, 2020a] Grifoni, P., Caschera, M.C. &Ferri, F. Evaluation of a dynamic classification method for multimodal ambiguities based on Hidden Markov Models. Evolving Systems (2020). https://doi.org/10.1007/s12530-020-09344-3

[Grifoni, 2020b] Grifoni P, Caschera MC, Ferri F (2020) DAMA: a dynamic classification of multimodal ambiguities. Int J ComputIntellSyst 13(1):178–192. https://doi.org/10.2991/ijcis.d.200208.001

[Güçlütürk, 2017] Güçlütürk, Y., Güçlü, U., Baro, X., Escalante, H. J., Guyon, I., Escalera, S., ... & Van Lier, R. (2017). Multimodal first impression analysis with deep residual networks. IEEE Transactions on Affective Computing, 9(3), 316-329.

[Guha, 2015] Guha T., Yang Z., Ramakrishna A., Grossman R. B., Hedley D., Lee S. and Narayanan S. S.: On Quantifying Facial Expression-Related Atypicality of Children with Autism Spectrum Disorder, Proc IEEE Int Conf Acoust Speech Signal Process. 2015 Apr; 2015: 803–807. doi: 10.1109/ICASSP.2015.7178080

[Halliday, 1967] Halliday, M.A.K. (1967). Notes on Transitivity and Theme in English, Part II, Journal of linguistics, 3: 199–244.

[Hossain, 2019] Hossain, M. S., & Muhammad, G. (2019). An audio-visual emotion recognition system using deep learning fusion for a cognitive wireless framework. IEEE Wireless Communications, 26(3), 62-68.

[Huang, 2019] Huang, K. Y., Wu, C. H., Hong, Q. B., Su, M. H., & Chen, Y. H. (2019). Speech Emotion Recognition Using Deep Neural Network Considering Verbal and Nonverbal Speech Sounds. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5866-5870). IEEE

[Kapoor, 2005] Kapoor, A and Picard, R. W. (2005). Multimodal Affect Recognition in Learning Environments, In ACM International Conference on Multimedia.

[Karchani, 2015] Karchani, M., Mazloumi, A., Saraji, G. N., Gharagozlou, F., Nahvi, A., Haghighi, K. S.,= and Foroshani, A. R. (2015). Presenting a Model for Dynamic Facial Expression Changes in Detecting Drivers' Drowsiness, Electronic Physician, 7(2): 1073.

[Khalifa, 208] Khalifa, I., Ejbali, R., &Zaied, M. (2018). Body Gesture Modeling for Psychology Analysis in Job Interview Based on Deep Spatio-Temporal Approach. In International Conference on Parallel and Distributed Computing: Applications and Technologies (pp. 274-284). Springer, Singapore.

[Koné, 2015] KonéC. ,Tayari I. M., Le-Thanh N. and Belleudy C. (2015) Multimodal Recognition of Emotions Using Physiological Signals with the Method of Decision-Level Fusion for Healthcare Applications, Inclusive Smart Cities and e-Health, Volume 9102 of the Series Lecture Notes in Computer Science 301-306.

[Lefter, 2014] Lefter I. (2014). Multimodal Surveillance: Behavior Analysis for Recognizing Stress and Aggression. PhD dissertation. Doi: 10.4233/uuid:d6b8a31e-71f5-4509-adca-9cd672432c1e.

[Lefter, 2011] Lefter, I., Rothkrantz, L. J. M., Van Leeuwen, D. A., and Wiggers P. (2011). Automatic Stress Detection in Emergency (Telephone) Calls, International Journal of Intelligent Defence Support Systems 4(2), 148-168.

[Lepschy, 1978] Lepschy, G. C. (1978). AppuntiSull'intonazioneItaliana, 127-142.

[Ma, 2012] Ma, M. D. (2012). Methods of Detecting Potential Terrorists at Airports, Security Dimensions and Socio-Legal Studies, 7, 33-46.

[Manliguez, 2016] Manliguez, C. (2016). Generalised Confusion Matrix for Multiple Classes. 10.13140/RG.2.2.31150.51523.[Martin, 2002] Martin, J. C. (2002). On the Use of Multimodal Clues in Human Behaviour for the Modelling of Agent Co-Operative Behaviour, Connection Science, 14(4): 297-309.

[Merrill, 1981] Merrill, D. W. and Reid, R. H. (1981). Personal Styles and Effective Performance. New York: CRC Press.

[Metallinou, 2013] Metallinou, A., Grossman, R. B. and Narayanan, S. (2013). Quantifying Atypicality in Affective Facial Expressions of Children with Autism Spectrum Disorders, In Multimedia and Expo (ICME), 2013a IEEE International Conference on. IEEE, 2013, 1-6.

[Metcalfe, 2019] Metcalfe, D., McKenzie, K., McCarty, K., &Pollet, T. V. (2019). Emotion recognition from body movement and gesture in children with Autism Spectrum Disorder is improved by situational cues. Research in developmental disabilities, 86, 1-10.

[Minaee, 2019] Minaee, S., &Abdolrashidi, A. (2019). Deep-emotion: Facial expression recognition using attentional convolutional network. arXiv preprint arXiv:1902.01019

[Nguwi, 2010] Nguwi, Y. Y. and Cho, S. Y. (2010). Support-Vector-Based Emergent Self-Organising Approach for Emotional Understanding, Connection Science, 22(4): 355-371.

[Oliver, 2005] Oliver, N. and Horvitz, E. (2005). A Comparison of HMMs and Dynamic Bayesian Networks for Recognising Office Activities, In Proceedings of the 10th International Conference on User Modeling (UM'05), Liliana Ardissono, Paul Brna, and Antonija Mitrovic (Eds.). Springer-Verlag, Berlin, Heidelberg, 199-209.
DOI=http://dx.doi.org/10.1007/11527886_26.

[Okada, 2019] Okada, S., Nguyen, L. S., Aran, O., &Gatica-Perez, D. (2019). Modeling dyadic and group impressions with intermodal and interperson features. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 15(1s), 1-30.

[Piana, 2014] Piana, S, Staglianó, A., Odone, F., Verri A. and Camurri A. (2014). Real-Time Automatic Emotion Recognition from Body Gestures, arXiv:1402.5047 (2014).

[Połap, 2019] Połap, D., Woźniak, M., Damaševičius, R., &Maskeliūnas, R. (2019). Bio-inspired voice evaluation mechanism. Applied Soft Computing, 80, 342-357.

[Rabiner, 1989] Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE 77(2): 257-286.

[Rosario, 2004] Rosario, KJ. (2004) Quick identification of social style, aptitudes, and motivation, http://www.docdatabase.net/more-quick-identification-of-social-style-aptitudes-and-motivation-392238.html (accessed 25 January 2021).

[Scheerer, 2020] Scheerer, N. E., Shafai, F., Stevenson, R. A., &Iarocci, G. (2020). Affective Prosody Perception and the Relation to Social Competence in Autistic and Typically Developing Children. Journal of abnormal child psychology, 1-11

[Schuller, 2019] Schuller, B., Weninger, F., Zhang, Y., Ringeval, F., Batliner, A., Steidl, S., ... &Chetouani, M. (2019). Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge. Computer Speech & Language, 53, 156-180.

[Sujatha, 2018] Sujatha, J., & Rajagopalan, S. P. (2018). Classification Of Parkinson Disease With The Voice Attributes Using Fuzzy Inference System. International Journal of Pure and Applied Mathematics, 118(9), 253-257.

[Tadesse, 2013] Tadesse, E. (2013). Drowsiness Detection for Driver Assistance (Doctoral dissertation, Oklahoma State University).

[Tiam-Lee, 2018] Tiam-Lee, T. J., & Sumi, K. (2018). Adaptive feedback based on student emotion in a system for programming practice. In International Conference on Intelligent Tutoring Systems (pp. 243-255). Springer, Cham

[Twitchell, 2004] Twitchell, D. P., Adkins, M., Nunamaker, J. F. and Burgoon, J. K. (2004). Using Speech Act Theory to Model Conversations for Automated Classification and Retrieval. In: Procs of the 9th International Working Conference on the Language-Action Perspective on Communication Modeling: 121-130.

[Vicsi, 2008] Vicsi, K. and Szaszak, G. (2008): Using Prosody for the Improvement of ASR - Sentence Modality Recognition, In Interspeech-2008, 2877-2880.

[Vigliocco, 2014] Vigliocco, G., Perniss, P. and Vinson, D. (2014). Language as a Multimodal Phenomenon: Implications for Language Learning, Processing, and Evolution. Introduction to Theme Issue Philosophical Transactions of the Royal Society B 369(1651), 20130292, 1-7. DOI:10.1098/rstb.2013.0292

[Viterbi, 1967] Viterbi, A.: Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory, 1967, pp. 260-269.

[Williamson, 2014] Williamson, J. R., Quatieri, T. F., Helfer, B. S., Ciccarelli, G. and Mehta, D. D. (2014). Vocal and Facial Biomarkers of Depression Based on Motor Incoordination and Timing, In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (65-72), ACM.

[Wang, 2018] Wang, C., Tian, Y., & Jia, H. (2018). Driving fatigue detection based on feature fusion of information entropy. Journal of Computational Methods in Sciences and Engineering, 18(4), 977-988.

[Wollmer, 2010] Wollmer, M., Metallinou, A., Eyben, F., Schuller, B. and Narayanan, S. S. (2010). Context-Sensitive Multimodal Emotion Recognition from Speech and Facial Expression using Bidirectional LstmModeling, INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010.

[Yang, 2013] Yang, Y., Fairbairn, C. and Cohn, J. F. (2013). Detecting Depression Severity from Vocal Prosody, Affective Computing, IEEE Transactions on, 4(2): 142-150.

[Yan, 2018] Yan, T., Wang, C., & Jia, H. (2018, July). Fatigue Detection Based on Facial Features with CNN-HMM. In International Conference on Frontier Computing (pp. 1516-1524). Springer, Singapore.

[Zhao, 2020] Zhao, W., & Lu, L. (2020). Research and development of autism diagnosis information system based on deep convolution neural network and facial expression data. Library Hi Tech.