



**HAL**  
open science

# Machine Learning-Based Multimodal integration for Short Utterance-Based Biometrics Identification and Engagement Detection

Abderrazzaq Moufidi

► **To cite this version:**

Abderrazzaq Moufidi. Machine Learning-Based Multimodal integration for Short Utterance-Based Biometrics Identification and Engagement Detection. Other. Université d'Angers, 2024. English. NNT: 2024ANGE0026 . tel-04910166

**HAL Id: tel-04910166**

**<https://theses.hal.science/tel-04910166v1>**

Submitted on 24 Jan 2025

**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.

# THÈSE DE DOCTORAT DE

L'UNIVERSITÉ D'ANGERS

ÉCOLE DOCTORALE N° 641  
*Mathématiques et Sciences et Technologies  
de l'Information et de la Communication*  
Spécialité : « *Signal, Image, Vision* »

Par

« **Abderrazzaq MOUFIDI** »

« **Machine Learning-Based Multimodal integration for Short Utterance-Based Biometrics Identification and Engagement Detection** »

Thèse présentée et soutenue à « INRAE - Angers », le « 25/10/2024 »

Unité de recherche : « LARIS - Laboratoire Angevin de Recherche en Ingénierie des Systèmes »

## Rapporteurs avant soutenance :

Pr. Kurosh MADANI, Université Paris-Est Créteil, France  
Pr. Hasan DEMIREL, Eastern Mediterranean University, Chypre

## Composition du Jury :

Président : Pr. Jean-Hugh THOMAS, Le Mans Université, France  
Examinatrice : Dr. Delphine GUEDAT-BITTIGHOFFER, Université d'Angers, France  
Dir. de thèse : Pr. David ROUSSEAU, Université d'Angers, France.  
Co-Dir. de thèse : Pr. Pejman RASTI, Université d'Angers, France.



# ACKNOWLEDGEMENT

---

I am very grateful to Professor David Rousseau, my supervisor, for his exceptional support and guidance during my PhD journey. Under his expert mentorship, I was able to enhance my skills in critical analysis and scientific judgment. His patience, unwavering support, and enthusiasm were crucial to my development as a researcher, and it was a privilege to learn from him. I would also like to extend my heartfelt thanks to Professor Pejman Rasti for his extensive support and invaluable guidance both within and beyond my research. His professional experience and calm, thoughtful advice have been instrumental in making this academic journey both enjoyable and enlightening. Additionally, I would like to express my sincere gratitude to Dr. Delphine Guedat-Bittighoffer for her immense help and expert feedback, which were essential for the development of the NeuroCam dataset designed for the multimodal detection of student engagement.

I would like to express my sincere gratitude to my thesis committee members, Professor Kurosh Madani, Professor Jean-Hugh Thomas, and Professor Hasan Demirel, for their kind agreement to evaluate my work and for their meaningful participation in my PhD defense.

This PhD is dedicated to the soul of my father, Mohammed Moufidi, and to my family, whose love and support have been my greatest sources of strength throughout this journey. I am deeply thankful to my mother, Barri El Kafazi, for her tireless support and dedication to my education and well-being. Despite her own lack of access to formal education, she was always deeply involved in our academic lives and constantly sought our comfort. I also want to thank my sister, Chadia, and my two brothers, Jean and Joseph, for their emotional, financial, and educational support and advices throughout my life; being their little brother has been a profound blessing and chance for me. I would like to express my gratitude to my two nieces, for their arrival into our lives and for filling the Moufidi family's space with joy and light.

I am so grateful for all my friends. A special thank you goes to my friend Alexis Pécheux. His friendship and the discussions about our PhD experiences, history and sports were a great comfort and a delightful distraction during the length of my PhD journey. I wish him a great success in his PhD defense next year.

---

I also appreciate the quality time in discussion and exchanging ideas with all colleagues in INRAe, especially all members of the ImHorPhen team.

# TABLE OF CONTENTS

---

<b>1</b>	<b>Introduction</b>	<b>15</b>
1.1	Research Context . . . . .	15
1.2	Multimodal Fusion Architectures . . . . .	16
1.3	The Core Research Problem . . . . .	18
1.4	Research Objectives . . . . .	18
1.5	Methodological Challenges in Multimodal Fusion . . . . .	19
1.6	Research Questions . . . . .	19
1.7	Contributions and Organization of the Document . . . . .	20
1.8	Publications . . . . .	21
<b>2</b>	<b>From Supervised Unimodal to Multimodal Biometrics Identification</b>	<b>23</b>
2.1	Introduction . . . . .	23
2.2	Speaker Identification based on Wavelet Scattering Transform . . . . .	26
2.2.1	Materials and methods . . . . .	26
2.2.2	Results and Discussion . . . . .	29
2.2.3	Conclusion and perspectives . . . . .	33
2.3	Attention-based Fusion of Ultra-short Voice Utterances and Depth Videos for Multimodal Person Identification . . . . .	34
2.3.1	Methodology . . . . .	34
2.3.2	Experimental results and Discussion . . . . .	40
2.3.3	Conclusion . . . . .	47
<b>3</b>	<b>Toward Comprehensive Short Utterances Manipulations Detection in Audio-Visual Videos</b>	<b>49</b>
3.1	Introduction . . . . .	49
3.2	Related Works . . . . .	52
3.3	Methodology . . . . .	53
3.3.1	Pre-processing: lips part selection . . . . .	53
3.3.2	Rationale for Hand-crafted Techniques . . . . .	54

TABLE OF CONTENTS

---

3.3.3	Deep learning Methodology . . . . .	58
3.4	Experimental Strategy and Materials . . . . .	60
3.4.1	Datasets . . . . .	60
3.4.2	Experimental Strategy . . . . .	63
3.5	Experimental Results . . . . .	64
3.5.1	Hand-crafted methods . . . . .	64
3.5.2	Audio - Visual late fusion deep learning method . . . . .	66
3.6	Discussion . . . . .	69
3.6.1	Hand-crafted methods . . . . .	69
3.6.2	Audio-Visual late fusion deep learning method . . . . .	70
3.7	Conclusion and perspectives . . . . .	71
<b>4</b>	<b>Unsupervised Multimodal Student’s Engagement Detection</b>	<b>73</b>
4.1	Introduction . . . . .	73
4.2	Related Works . . . . .	74
4.2.1	Multidimensionality of emotions in FL courses . . . . .	75
4.2.2	Anxiety, boredom, and enjoyment in FL classes . . . . .	77
4.3	Materials and methods . . . . .	81
4.3.1	Study design and participants . . . . .	81
4.3.2	Instruments . . . . .	81
4.3.3	Preprocessing Methods . . . . .	84
4.3.4	Data analysis . . . . .	87
4.4	Results . . . . .	92
4.4.1	Necessary clarifications about the experimental results . . . . .	92
4.4.2	Convergences between the cross-referenced results of Components 1 and 2 . . . . .	94
4.4.3	Convergences between the cross-referenced results of Components 1 ( $C_1$ ), 2 ( $C_2$ ) and their fusion and the observations of the courses (OBS) . . . . .	96
4.4.4	Detailed analysis of triangulated results with component 3: focus on each student . . . . .	99
4.5	Discussion . . . . .	102
4.6	Conclusion . . . . .	106

<b>5</b>	<b>Conclusion and Perspectives</b>	<b>107</b>
5.1	Methodological Contributions to Multimodal Pattern Recognition . . . . .	107
5.2	Future Perspectives . . . . .	108
5.2.1	Biometrics Identification . . . . .	108
5.2.2	Student Engagement Detection . . . . .	109
5.3	Publications . . . . .	110
	<b>Bibliography</b>	<b>111</b>
<b>A</b>	<b>Wavelet Scattering Transform</b>	<b>111</b>
<b>B</b>	<b>Technical Description of the oximeter used in our experiments</b>	<b>113</b>
B.1	Fingertip Pulse Oximeter Features: . . . . .	113
B.2	Fingertip Pulse Oximeter Specification: . . . . .	114



# LIST OF FIGURES

---

1.1	Types of fusion, (a) represents the early fusion, (b) late fusion and (c) hybrid fusion (joint feature representation projects features from different modalities into a common latent space). Red arrows represent the inter-connections between models in hybrid fusion. . . . .	17
2.1	Fusion of 1D-CNN architecture with WST ( $C_{out}$ is the number of output channels from a 1D temporal convolution, here it was set to 128, $C_d$ is the number of WST coefficients of depth $\geq 2$ ). . . . .	29
2.2	Classification accuracy per sentence for different invariance scales when the audios are sampled at 16 kHz (c) and at 8 kHz (a), and for different depths of the best invariance scale 64 ms sampled at 16 kHz (d) and at 8 kHz (b). . . . .	30
2.3	The classification accuracy % per sentence (a-c) and the classification accuracy % per frame (b-d) for TIMIT sampled at 8 kHz (a-b) and 16 kHz (c-d) when using WST + 1D-CNN for different depths of invariance scales < the optimal invariance scale 64 ms. . . . .	32
2.4	Spectrogram of the WST coefficients at the first depth (first speaker (a-1) and the second speaker (b-1)) and the second depth (first speaker (a-2) and the second speaker (b-2)) (x-axis time and y-axis is frequencies). . . . .	32
2.5	Spectrogram of the WST coefficients at the third depth (first speaker (a-3) and the second speaker (b-3)) and the fourth depth (first speaker (a-4) and the second speaker (b-4)) (x-axis time and y-axis is frequencies). . . . .	33
2.6	Multi-view Video CNN architecture used on lip depth videos (The red dashed line represent the extraction of the features vector from the view projection of the video). . . . .	38
2.7	Architecture of late fusion of the two modalities (audio and depth video). . . . .	39

2.8	General Informations on the three benchmarks used during the study ( $F_s$ represents the frequency sampling before the re-sampling to $16\text{ kHz}$ , the audio curves are full utterance or a part of a word spoken by a speaker in each dataset). Real depth refers to the depth map obtained directly from a depth camera. In contrast, estimated depth stands to depth map generated from RGB images using the method outlined in paragraph 2.3.2.1. . . . .	42
3.1	Proposed pipeline for short utterances manipulations detection in videos. .	55
3.2	(a) The difference between the WST of positive and negative sample of Real audio. (b) The difference between the WST of positive and negative sample of the fake audio represented in (a). (c) The WST of the same real signal. (d) The WST of the fake signal. Only the zero-th and the first order WST are presented and the fake signal was generated using the method described in [66] (i.e: the WST index refers to the mel-scaled frequencies between 0 Hz and 8 kHz). . . . .	56
3.3	Motion jitters problem illustration from the article [62]. Left and middle are two deepfake generation methods, and the right image is taken from a real video. For each video, the vertical cuts (the vertical red/green/blue line) are made in each frame along time, and then show their concatenation at the bottom. . . . .	59
3.4	Multi-view CNN late fusion architecture for audio-lips correlation [45]. For multimodal deepfake detection, the number of classes are set to 2 or 4, in addition, we only tune the fusion layers (all layers except 2D ResNet-18 and x-vectors). . . . .	60
3.5	Samples from the 4 classes of FakeAVCeleb dataset [80]. . . . .	61
3.6	Two frames samples from DeepfakeTIMIT dataset ((a) Real, (b) Fake). . .	62
3.7	Analysis of low-quality 20 fake and genuine videos split into $600\text{ms}$ with an overlap of 50% using our method. Top-left: Mean and standard deviation without spatial derivation. Top-right: After second-level spatial derivation. Bottom-left: After fourth-level spatial derivation. Bottom-right: Temporal inverse coefficient of variation of videos from DeepfakeTIMIT and VidTIMIT.	67
3.8	Pearson coefficient value (y-axis) described in Eq. 3.2 per each channel (x-axis) (orange circles refers to the real audio and the blue star stands to the fake audio). . . . .	69

4.1 Left: Oximeter used to record the heart rate beats of the students. Right: ViHealth application that receive the data from the oximeter. . . . . 82

4.2 Camera used to record RGB videos (top left) and a sample of a teaching session from NeuroCam dataset. . . . . 83

4.3 Example of the output image from our oximeter. (a) is the name of the participant, (b) the date and the hour of the beginning of the measurement, (c) test duration, (d) remarks, (e) the birthday date and the age, (f) the signal curve of the oxygen level during the session (y-axis: oxygen level % and x-axis: time instant of the session), (g) the heart rate variation of the student during the teaching session (y-axis: heart rate *bpm* and x-axis: time instant of the session), (h) gender, (i) table representing the statistics (max, mean, min) of the oxygen level % and the heart rate *bpm*, (j) the end date and hour of the recording. . . . . 85

4.4 Heart Rate signal of a student during a session. The green curve corresponds to the real and the blue one is its approximation using our method described in 4.3.3.1 (x-axis: time instant of the session, y-axis: the heart rate of the student (bpm)). . . . . 86

4.5 Architecture used for emotional facial expressions recognition taken from [142].  $L$  is set to 12 in the pre-trained model. . . . . 89

4.6 Our proposed pipeline method to detect significant moments during a teaching session. . . . . 90

4.7 The pipeline designed for tracking the student engagement based on the HR signal.  $T$  is the duration of the framing. . . . . 91

4.8 GMM applied on the HR features space for Oksana.  $\mu, \sigma$  refer respectively to the mean and the standard deviation of the HR signal,  $\mu_v, \mu_a$  refer respectively to the mean of the speed and the absolute acceleration of the HR signal. . . . . 93

4.9 GMM applied on the HR features space for Mitch.  $\mu, \sigma$  refer respectively to the mean and the standard deviation of the HR signal,  $\mu_v, \mu_a$  refer respectively to the mean of the speed and the absolute acceleration of the HR signal. . . . . 94

4.10 GMM applied on the HR features space for Zeynep.  $\mu, \sigma$  refer respectively to the mean and the standard deviation of the HR signal,  $\mu_v, \mu_a$  refer respectively to the mean of the speed and the absolute acceleration of the HR signal. . . . . 95

4.11 Silhouette scores for each student when using GMM clustering on HR chunks of 2 minutes 30 seconds. . . . . 95

4.12 Intersection over Union of HR (Component 1) and EFE (Component 2) SM for each student. . . . . 96

4.13 Average scores of the three students for the three emotions across all sessions. 100

B.1 Left: Oximeter used to record the heart rate beats of the students. Right: ViHealth application that receive the data from the oximeter. . . . . 114

# LIST OF TABLES

---

2.1	Energy absorbed at different scale for some invariance scale $T$ and $N_s$ the total number of scatter coefficients for TIMIT audios sampled at $8\text{ kHz}$ . . .	27
2.2	Energy absorbed at different scale for some invariance scales $T$ and $N_s$ the total number of scatter coefficients for TIMIT audios sampled at $16\text{ kHz}$ . . .	28
2.3	Classification accuracy per sentence of the best values of invariance scale and depth of WST + 1D-CNN and baselines trained and tested on TIMIT sampled at $8\text{ kHz}$ and $16\text{ kHz}$ ( $\#param$ is the number of learnable weights in each architecture). . . . .	31
2.4	Identification accuracy and the time processing of each method applied on 20 speakers TCD-TIMIT [50] re-sampled at $16\text{ kHz}$ . . . . .	35
2.5	Data characteristics after word segmentation. . . . .	42
2.6	Identification accuracy per each view of depth video. Real depth refers to the depth map obtained directly from a depth camera. In contrast, estimated depth stands to depth map generated from RGB images using the method outlined in paragraph 2.3.2.1. . . . .	43
2.7	Identification accuracy per each modality and their fusion for the three benchmarks. Real depth refers to the depth map obtained directly from a depth camera. In contrast, estimated depth stands to depth map generated from RGB images using the method outlined in paragraph 2.3.2.1. . . . .	44
3.1	Performance of our method in Eq. (3.2) on short utterances and full audio modality in FakeAVCeleb, compared to SOTA methods (the standard deviation values are calculated from a 5-fold cross-validation). . . . .	65
3.2	Performance of our proposed method on the whole audio length modality of FakeAVCeleb compared to the SOTA methods (Real: 500 videos from $R_v R_a$ , Fake: 500 videos from $F_v F_a$ ). Bold entries indicate the best performance. . . . .	65

---

3.3	Performance of the SOTA and our proposed method on whole length visual sequence from low-quality videos of DeepfakeTIMIT. Bold entries indicate the best performance. . . . .	66
3.4	Comparison of SOTA performance with our proposed approach using visual sequences from low-quality DeepfakeTIMIT videos over various time segments. . . . .	66
3.5	Comparison of SOTA performance with our proposed approach [45] on various time segments from FakeAVCeleb on balanced 1,000 samples (500 $F_vF_a$ , 500 $R_vR_a$ ) subset, the train-test split was set to 80% – 20%. . . . .	67
3.6	Detection accuracy of Multi-view CNN on short videos from FakeAVCeleb (train-test split 80% – 20% on 2,000 videos of (500 $F_vF_a$ , 500 $F_vR_a$ , 500 $R_vF_a$ and 500 $R_vR_a$ ), the classes number for classification is set to 4). . . . .	68
3.7	Detection accuracy of Multi-view CNN on short videos from FakeAVCeleb (train and test on 1,000 videos of (500 $F_vF_a$ , 500 $R_vR_a$ ) cutted into 1 s frame length with an overlap 50%, the classes number for classification is set to 2). . . . .	68
4.1	The convergences between: $C_1$ (results obtained by our HR signal-based method) and OBS (observation of the course progression and students ‘activities’); and between $C_2$ (results obtained by our EFE based method) and OBS; the Multimodal Decision Fusion and OBS. . . . .	98

# LIST OF ACRONYMS

---

<b>WST</b>	Wavelet Scattering Transform
<b>LPC</b>	Linear Predictive Coefficients
<b>MFCC</b>	Mel-Frequency Cepstrum Coefficients
<b>CNN</b>	Convolutional Neural Network
<b>LSTM</b>	Long Short-Term Memory
<b>TDNN</b>	Time Delay Neural Network
<b>SI</b>	Speaker Identification
<b>MLP</b>	MultiLayer Perceptron
<b>ECAPA-TDNN</b>	Emphasized Channel Attention, Propagation and Aggregation in TDNN
<b>ResNet</b>	Residual neural Network
<b>GAN</b>	Generative Adversarial Network
<b>SVM</b>	Support Vector Machines
<b>DCT</b>	Discrete Cosine Transform
<b>ECG</b>	Electrocardiogram
<b>EDA</b>	Electrodermal Activity
<b>EMG</b>	Electromyography
<b>HR</b>	Heart Rate
<b>EEG</b>	Electroencephalogram
<b>PPG</b>	Photoplethysmogram
<b>ViT</b>	Vision Transformer
<b>MSA</b>	Multi-Head Self-Attention
<b>GMM</b>	Gaussian Mixture Models
<b>LOF</b>	Local Outlier Factor

# INTRODUCTION

---

## 1.1 Research Context

A modality is a specific format used to encode a particular type of information, such as RGB, depth, thermal image, textual, audio, video, or physiological signals [1, 2]. The integration of these heterogeneous modalities, known as multimodal fusion, has emerged as a critical area of research due to its ability to produce more robust and informative systems compared to those relying solely on unimodal data.

One practical example of the need for multimodality can be observed in video communication systems, where audio signals can be difficult to perceive and understand in the presence of background noise. Integrating visual cues, such as lip movements from RGB videos, with voice signals significantly improves speech recognition, particularly in noisy environments, thereby enhancing the overall communication experience [3]. Similarly, the fusion of RGB and depth video with vehicle state measurements, such as speed, has been instrumental in advancing autonomous vehicle technologies. These multimodal systems are capable of making more accurate real-time decisions, contributing to safer navigation and better adaptability to diverse driving conditions [4]. Another compelling application of multimodality is in the domain of content moderation on social media platforms. By merging textual information with RGB images or videos, machine learning models can effectively identify harmful or inappropriate content, thereby enhancing the safety and integrity of online spaces [5]. Moreover, combining RGB-Depth video with audio data can lead to a richer understanding of emotional stimuli and environmental context. This approach has been employed to accurately gauge student engagement in educational settings, providing educators with deeper insights into classroom dynamics and individual student needs [6, 7, 8, 9, 10, 11, 12]. Biometric systems also benefit from multimodal integration, such as combining facial recognition with voice signals to improve identity verification. In noisy environments, visual data may be more reliable, while voice data becomes more dependable in low-light conditions [13, 14, 15].



These examples highlight the critical role of multimodal integration in enhancing system robustness across various domains. However, integrating multiple modalities—each with distinct structures, constraints, and sampling rates—presents substantial challenges. For instance, while videos are spatiotemporal sequences with frame rates of 24-60 frames per second, audio is a 1D signal typically recorded at 8-48 kHz. These differences contribute to the complexity of multimodal fusion [1].

## 1.2 Multimodal Fusion Architectures

To address these challenges, three main deep multimodal fusion architectures have been proposed in the literature:

- **Early fusion** methods integrate raw or feature-level data from multiple modalities at the initial stages of processing, facilitating early cross-modal interactions. Given  $n$  input modalities  $x_1, x_2, \dots, x_n$  and a model  $I$ , the output  $y$  can be expressed as:

$$y = I(x_1, x_2, \dots, x_n). \quad (1.1)$$

These techniques aim to optimally combine information during feature extraction, enhancing the model’s ability to learn robust features. However, challenges include high-dimensional feature spaces, temporal alignment issues due to varying modality characteristics, and the risk of one modality dominating the others. Also, when modalities have the same size (e.g: RGB-Depth maps), a simple concatenation of features offers limited benefits, as it may not fully exploit interdependencies among modalities [16]. Therefore, more advanced techniques, like attention mechanisms or cross-modal transformers, have been developed to address these limitations and better capture relationships between modalities, yet they remain computationally costly.

- **Late fusion** aims to integrate multimodal feature maps at the decision level. In this approach, multimodal data is processed separately in different branches as unimodal data. During the final stage, feature maps computed by these branches are mapped into a common feature space through fusion operations, such as concatenation, addition, averaging, or weighted voting, followed by a series of fully connected layers or a classifier. Let  $x_1, x_2, \dots, x_n$  be the  $n$  input modalities and

$I_1, I_2, \dots, I_n$  be the models used for each modality, the output  $y$  can be written as:

$$y = P(I_1(x_1), I_2(x_2), \dots, I_n(x_n)), \quad (1.2)$$

where  $P$  is the fusion operation as well as the following fully connected layers or classifier. This fusion model offers greater flexibility and scalability. However, because the model is trained to learn unimodal features separately, there is a lack of cross-modal interaction, which may limit its ability to fully exploit the interdependencies among different modalities [16].

- **Hybrid fusion** methods are sophisticated approaches that combine the advantages of both early and late fusion strategies. They can adaptively create a joint feature representation that aims to project features from different modalities into a common latent space. This dynamic approach allows achieving superior performance, enhancing both accuracy and robustness in multimodal systems [16, 17].

These architectures differ in the stage at which the modalities are combined within the machine learning model as illustrated in figure 1.1.

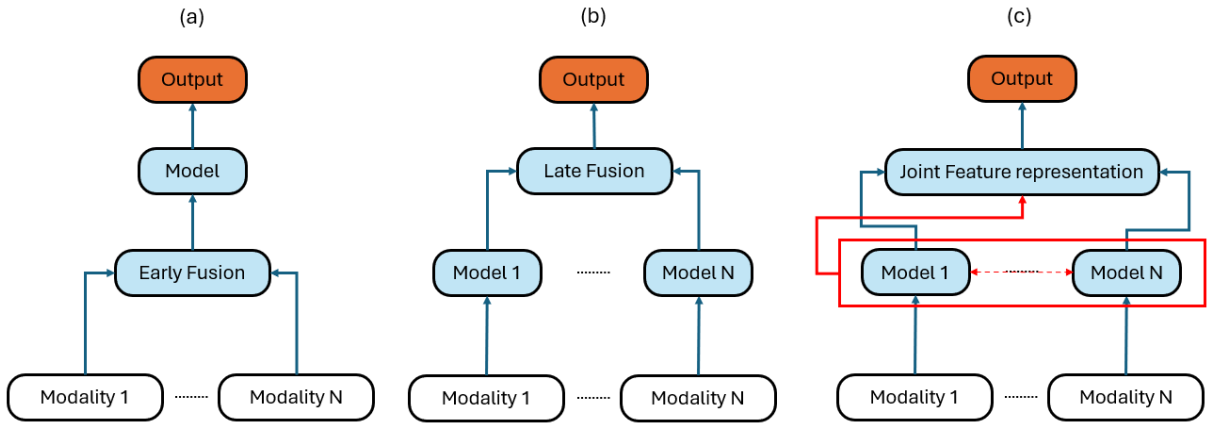


Figure 1.1 – Types of fusion, (a) represents the early fusion, (b) late fusion and (c) hybrid fusion (joint feature representation projects features from different modalities into a common latent space). Red arrows represent the interconnections between models in hybrid fusion.

While these architectures have proven to be effective in various multimodal tasks, their application to short-duration temporal data remains an open research area. This thesis explores how late fusion technique can be adapted to process short-duration signals more efficiently, emphasizing accuracy while minimizing computational overhead.

## 1.3 The Core Research Problem

Despite the advantages of multimodal systems, there remains a critical research gap: the integration of short-duration time series (less than one second) within multimodal frameworks is underexplored. Existing multimodal fusion methods often rely on computationally intensive models, unsuitable for real-time applications, or overly simplify the problem by excluding temporal information, focusing solely on spatial features [17, 18]. These approaches limit the applicability of multimodal systems in real-world scenarios that require efficient processing and rapid decision-making.

This thesis addresses the core research problem: how to effectively and efficiently integrate short-duration temporal modalities in multimodal systems while maintaining high performance, reducing computational complexity, and ensuring scalability. This research problem is particularly relevant for applications such as biometric identification and student engagement detection, where multimodal fusion of short time-series data can provide significant performance improvements but is currently constrained by computational and temporal limitations.

## 1.4 Research Objectives

To address the core research problem, this thesis focuses on the following objectives:

- **Handling short-duration utterances:** Develop methods to effectively process short-duration utterances by multimodal systems, improving model robustness and accuracy across both multimodal and unimodal settings.
- **Efficiency and Practicality:** Design computationally efficient models that maintain high performance while reducing complexity. The goal is to create models that are suitable for real-world applications by utilizing low-cost sensors, such as microphones, RGB-depth cameras, and oximeters.
- **Temporal Information Integration:** Improve the integration of temporal information in multimodal frameworks by leveraging pre-trained unimodal models, such as X-vectors, ResNet, to capture temporal or both spatial and temporal features effectively [19, 20].

## 1.5 Methodological Challenges in Multimodal Fusion

Integrating multimodal data, especially for short-duration time series, presents several key challenges:

- **Temporal and Spatial Complexity:** Short-duration time series require precise, resource-efficient, and time-efficient spatiotemporal feature extraction for modalities such as RGB videos. However, these temporal dependencies are often underutilized in conventional fusion techniques, leading to suboptimal performance.
- **Heterogeneity of Modalities:** Different modalities have varying structures and sampling frequencies, making it difficult to fuse them effectively [1].
- **Computational Efficiency:** Many existing multimodal fusion models, particularly those utilizing deep learning, are computationally intensive. This limits their practicality, especially in real-time applications that require rapid processing of data from multiple modalities [17].
- **Scalability and Interpretability:** Beyond accuracy, models must be interpretable and scalable for real-world datasets, which often contain noisy or missing data. This makes the design of robust, efficient models a significant challenge.

## 1.6 Research Questions

In light of the challenges outlined above, this thesis seeks to address the following key research questions:

- 1. What methods can be developed to enhance the accuracy of multimodal integration while minimizing the number of trainable parameters?**
  - This question will be explored through the development of novel fusion methods aimed at combining information from different short duration temporal modalities without excessively increasing model complexity.
- 2. What strategies are effective in detecting synthetic data manipulations, such as deepfakes, using multimodal data?**
  - This will involve crafting robust detection mechanisms to identify and counteract sophisticated data manipulation techniques, such as deepfakes, by leveraging cross-modal interactions between visual and auditory.
- 3. How to identify student engagement on short utterances using multimodal wearable sensors in educational settings?**

- This question focuses on exploring unsupervised learning methods that can capture the complexities of real-world multimodal data.

These questions guide the exploration of how multimodal integration can be optimized to handle the unique challenges posed by short-duration time series.

## 1.7 Contributions and Organization of the Document

This thesis contributes to the field of multimodal machine learning by addressing the challenge of integrating short-duration temporal data. The dissertation is structured as follows:

- **Chapter 2** investigates how integrating multimodal information enhances biometric recognition accuracy and confidence, particularly in the context of short utterances. First, we focus on unimodal voice-based biometric identification using a well-known signal processing tool called the Wavelet Scattering Transform (WST). We apply this method to improve accuracy for short utterances and to reduce the number of trainable parameters by optimizing its hyperparameters. Finally, we explore the advantages of multimodality by proposing a novel, low-computational-cost multimodal late fusion approach that integrates audio and lip depth videos, demonstrating how this integration improves both robustness and accuracy.
- **Chapter 3** proposes two robust multimodal deepfake detection methods that leverage both RGB visual and audio modalities. These methods are based on an interpretable, novel shallow learning architecture and the deep learning model developed in Chapter 2. The focus is on distinguishing real from fake videos, with particular emphasis on the analysis of RGB lip videos and audio data, all while maintaining a low computational cost.
- **Chapter 4** explores a practical application of multimodality in real-world settings, specifically in detecting student engagement through heart rate signals and facial expression. We introduce a novel dataset that combines these modalities, taking into account the complexities inherent in teaching sessions. Furthermore, we explore an unsupervised subject-based methodology to monitor student engagement, developed in collaboration with a didactic researcher.
- **Chapter 5** summarizes the key contributions, discusses limitations, and outlines potential future research directions, particularly in extending the proposed meth-

ods to new domains and applications.

## 1.8 Publications

### Journal Articles

- **Abderrazzaq Moufidi**, David Rousseau, and Pejman Rasti, "Attention-Based Fusion of Ultrashort Voice Utterances and Depth Videos for Multimodal Person Identification", *Sensors*, 23.13 (2023), p. 5890.
- **Abderrazzaq Moufidi**, David Rousseau, and Pejman Rasti, "Toward Comprehensive Short Utterances Manipulations Detection in Videos", *Multimedia Tools and Applications* (2024): 1-14.
- Delphine Guedat-Bittighoffer, **Abderrazzaq Moufidi**, Jean-Marc Dewaele, David Rousseau, Hugo Voyneau, Pejman Rasti, "Heart rates, facial expressions and self-reports: A multimodal longitudinal approach of learners' emotions in the Foreign Language classroom" *Computers and Education*, (**Submitted**).

### Proceedings

- **Abderrazzaq Moufidi**, David Rousseau, and Pejman Rasti, "Wavelet Scattering Transform Depth Benefit: An Application for Speaker Identification," *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, Springer, 2022, pp. 97–106.
- **Abderrazzaq Moufidi**, David Rousseau, and Pejman Rasti, "Multimodal Deepfake Detection for Short Videos," *IMPROVE*, 2024, pp. 67–73.
- Delphine Guedat-Bittighoffer, **Abderrazzaq Moufidi**, David Rousseau, and Pejman Rasti, "Regards interdisciplinaires croisés sur les émotions éprouvées par trois apprenants de Français langue étrangère en contexte universitaire," *Colloque Langage et éMOTions*, 2024.



# FROM SUPERVISED UNIMODAL TO MULTIMODAL BIOMETRICS IDENTIFICATION

---

As discussed in the introduction, the integration of multimodal information improves accuracy and increases confidence in the output by exploiting the strengths of each modality. In this chapter, we will explore this claim in the context of fully supervised biometric recognition, from the unimodal part to the benefits of multimodality. We begin with an introduction in Section 2.1, which summarises unimodal and multimodal biometric identification. Next, Section 2.2 is dedicated to speaker identification based on Wavelet Scattering Transform Depth Benefit. In the last Section 2.3, we delve into the benefit of multimodality by inserting the lip depth video to the identification system.

## 2.1 Introduction

Biometrics recognition finds widespread applications in various fields such as security systems, access control, and surveillance [21]. This technology capitalizes on the unique traits inherent to each individual, encompassing:

- Physiological: face [21, 22], iris [21, 23], fingerprint [21], ears [24] or lips [25],
- Behavioral: lips movement [26] or voice [27].

Using voice to identify an individual is cheaper than other biometric methods such as face [22] or iris [23] recognition, as the material to record an audio is less expensive than cameras and it does not require a light source to operate as long as the extrinsic variations are not severe (noise, reverberation or low sampling frequency). Speaker identification aims to identify a person based on their voice [27]. Although it has been widely studied [27], the acquisition of long and clear voice samples with a high sampling frequency can be challenging [28, 29, 30].



Deep embedding methods have gained popularity due to their robustness in speaker identification task by feeding raw samples or hand crafted methods to a convolutional neural network (CNN). Among these CNN methods, SincNet [28], x-vectors and its improvements [19, 31]. W. Ghezaiel et al. [29] proposed also an hybrid architecture based on sparse method called Wavelet Scattering Transform (WST) [32, 33] and 2D-CNN network called HWSTCNN, this architecture strengthens the capturing of texture (dominant energy in patterns) for short utterances (i.e: audio signals under 3 seconds) by extracting the WST coefficients up to the second order and fed them to a CNN network, this method has shown its ability to increase identification performance under short utterances circumstances, in addition it reduced the number of parameters compared to other methods [28, 30]. Despite the superior performance of the HWSTCNN among other methods, there has not been enough investigation on the deepness of the WST architecture to determine if it can provide more discriminant features.

In Section 2.2, we explore the importance of the depth and the invariance scale i.e frame length of WST [32, 33] in the speaker identification task on TIMIT dataset [34] under the circumstances where we have short utterances ( $< 3$  seconds) and low sampling frequency  $8 kHz$ .

On the other hand, the rapid advancement and the democratization of technology has lead to the abundance of multimodal sensors. To extract this diversity of information, multimodal deep learning has emerged as a powerful approach for various tasks by combining information from different modalities, exploiting their complementary nature, and enhancing overall performance [35, 26, 36, 37]. In the realm of speaker recognition incorporating multiple features, such as lip movements, depth modality images, and voice, can lead to improved accuracy and robustness in applications like security systems, access control, and surveillance [26, 38, 25, 39, 40, 41]. Utilizing a combination of physiological and behavioral features, including face [22], iris [23], fingerprints [21], voice [27], and lip movements, can benefit person identification [26, 25]. While each feature offers unique advantages and limitations, integrating them using multimodal deep learning methods can lead to more reliable identification systems [26, 36]. Among these features, lip movements serve as a vital behavioral feature for speaker recognition encompassing both physiological (static) and behavioral (temporal) aspects [25].

Traditional methods, such as RGB images and videos, have been commonly used for lip verification [26, 25, 42, 38]. However, they are susceptible to issues such as varying lighting conditions, pose variations, and occlusions. In contrast, depth images offer sev-

eral advantages over traditional RGB images. They are more robust against attempts to deceive recognition systems using photographs and exhibit strong invariance to lighting conditions, which can adversely affect RGB images in challenging environments [43].

One of the challenges in multimodal deep learning is effectively fusing the information from different modalities [36]. Existing fusion techniques, such as early, late, and hybrid fusion, have their inherent limitations, often resulting in suboptimal integration of multimodal information [16, 35]. Our proposed attention-based fusion model effectively addresses these limitations, enabling the model to better adapt to the varying degrees of informativeness in different modalities and achieve improved performance.

Despite the significant strides made in multimodal deep learning, a clear research gap persists when applied to speaker recognition. The existing methods heavily rely on long speech utterances and RGB images for identification. This dependence often leads to impractical and inconsistent outcomes, especially in scenarios where obtaining extended utterances or ideal lighting conditions for RGB images is challenging. Additionally, these methods face difficulties in robustly handling various types of noise, which are typical in real-world situations. Our study directly addresses this gap by developing a novel approach that utilizes ultra short voice utterances and depth videos of lip, ensuring practicality and improved accuracy even in less than ideal conditions.

In Section 2.3, we introduce a fresh perspective on the application of multimodal deep learning for speaker identification in the context of short duration utterances (i.e: at the word level, under 1 second). The crux of our innovation is a distinctive encoding model that processes time-series depth modality images of lip and ultra short voice utterances, demonstrating superior performance on benchmark datasets despite the limited information available. This is paired with an attention-based model that enhances identification accuracy by effectively fusing multimodal data and focusing on the most informative regions. Therefore, the main contributions and novelty of our work are listed as follows:

- We introduce a new encoding model for time series depth images of the lip and ultra short voice utterances (at the word level, i.e: <1 second), which leverages convolutional networks to capture both spatial and temporal features. Despite the challenges posed by the limited information available in short utterances and depth images, our encoding model demonstrates superior performance on benchmark datasets.
- We propose an attention-based model for the fusion of multimodal data, effectively combining information from different modalities such as lip and ultra short voice

utterances, and depth images. The attention mechanism allows the model to concentrate on the most informative regions of the input data, enhancing the accuracy of person identification.

- Our deep learning model is robust to various noises, such as ambient noise and background music, thanks to depth modality, which effectively handles these challenges and makes the model suitable for real-world conditions. In contrast to many state-of-the-art methods, our model effectively handles different types of noise and achieves superior performance.
- We present a comprehensive evaluation of our proposed approach on benchmark datasets, demonstrating its effectiveness in challenging scenarios with ultra short voice utterances (less than 1 s) and depth videos. This showcases the potential of our method in real-world applications, such as security systems and access control.

To summarize, our contributions in unimodal and multimodal biometrics field are:

- In Section 2.2, we explore the importance of the depth and the invariance scale i.e frame length of WST [32] in an application on the speaker identification task on TIMIT dataset [34], for both text dependent and independent tasks under the conditions, shortness of the utterances and the small value of the sampling frequency.
- In Section 2.3, we propose a novel method to encode depth lips times series and a self attention-based model for the fusion of multimodal data, effectively combining information from different modalities such as ultra short voice utterances and depth lips videos.

The work presented in section 2.2 has been published at ANNPR Workshop 2022 [44], while the section 2.3 was published on MDPI Sensors Journal [45].

## 2.2 Speaker Identification based on Wavelet Scattering Transform

### 2.2.1 Materials and methods

#### 2.2.1.1 Datasets

TIMIT [34] corpus is an audio dataset sampled at  $16\text{kHz}$  (recording conditions), with the primary aim of furnishing speech data for studies in acoustic-phonetics and for

Table 2.1 – Energy absorbed at different scale for some invariance scale  $T$  and  $N_s$  the total number of scatter coefficients for TIMIT audios sampled at  $8kHz$ .

$\frac{\ Sx\ ^2}{\ x\ ^2} \%$	Depth				
	$T N_s$	1st	2nd	3rd	4th
16ms 35	90.18	0.21	-	-	-
32ms 74	86.69	0.61	$4.5e-3$	-	-
64ms 152	82.79	1.21	$2.08e-2$	$2.13e-4$	-
128ms 308	71.04	3.39	$6.1e-2$	$1.4e-3$	$1.38e-5$

assessing automatic speech recognition systems. It contains 630 speakers (192 females and 438 males), each speaker reads 10 English sentences, where 2 sentences are common to all speakers. For all our experiments, we consider only 462 speakers in the "TRAIN" folder and we use only 8 sentences, where "SX" sentences (5 for each speaker) are destined to training phase and "SI" (3 per each speaker) are intended for the testing phase. The average duration of the "SX" sentences is about 4 s and the duration of "SI" sentences is about of 2 – 6 s. Which makes a train/test ratio of 5 sentences/3 sentences.

### 2.2.1.2 Experimental setup

While S. Mallat proposed in [32, 33] to limit the depth of the WST (detailed in Annex A) at the second depth for an invariance scale i.e frame length of 32 ms adapted to audio applications, this was based on the fact that the majority energy absorbance is situated at this scale. In this section, we explore the importance of the depth and the invariance scale of WST [32] in an application on the speaker identification task on TIMIT dataset [34], for both text dependent and independent tasks under the conditions, shortness of the utterances and a small value of the sampling frequency (i.e: 8 kHz). We report the WST energy absorbance at different invariance scale until their maximum depth on all audios from TIMIT dataset [34] sampled at the well known frequencies in audio speaking applications  $8kHz$  (telecommunications) in Table 2.1 and  $16kHz$  (recording instruments) in Table 2.2.

To find the optimal values of depth and invariance scale of WST that are suited for Speaker Identification (SI), we perform SI text-independent on 462 speakers from TIMIT dataset and SI text-dependent on two speakers reading the same sentence. The method is then fused with a CNN and compared to the three baselines [29, 30, 28].

#### 2.2.1.2.1 Speaker identification text-independent

Table 2.2 – Energy absorbed at different scale for some invariance scales  $T$  and  $N_s$ , the total number of scatter coefficients for TIMIT audios sampled at  $16\text{ kHz}$ .

$\frac{\ Sx\ ^2}{\ x\ ^2} \%$	Depth					
	$T  N_s$	1st	2nd	3rd	4th	5th
16 ms  74	89.65	0.54	$4.0e - 3$	-	-	-
32 ms  152	86.13	1.01	$2.17e - 2$	$2.13e - 4$	-	-
64 ms  308	82.18	1.6	$5.05e - 2$	$1.4e - 3$	$1.32e - 5$	-
128 ms  620	70.47	3.72	0.11	$4.4e - 3$	$1.11e - 4$	$9.34e - 7$

**2.2.1.2.1.1 Depth and invariance scale** In order to extract the optimum depth and invariance scale of WST for SI task, we work on SI text-independent task done on 462 speakers from TIMIT [34] sampled at  $8\text{ kHz}$  and  $16\text{ kHz}$ . Firstly, we preprocess the audio files by removing the silent frames at the beginning and the end of an utterance (no pre-emphasize was applied). Secondly, we apply the WST by using different invariance scales 16, 32, 64, 128 ms until their maximum depth. We omitted the WST of silent frames for this experiment. The WST coefficients are log-normalized following the normalization equations defined in [32] and Annex A. The resulted frames from 5 utterances beginning with "SX" are used for the training phase and the last 3 utterances starting with "SI" are used for the testing part. A multilayer perceptron (MLP) is used as a classifier with cross entropy as a loss function. The batch size and the maximum epoch were set respectively to 256 and 100. The optimizer used was ADAMAX with a learning rate  $2.10^{-2}$ .

**2.2.1.2.1.2 Comparison to baselines** To enhance the classification accuracy per sentence, feeding WST coefficients to a neural network, inspired from x-vectors architecture [19], is essential to have performance as the SOTA. At the input, the architecture given in Fig. 2.1 receives WST coefficients of  $230\text{ ms}$  frame length and an overlapping of  $58\text{ ms}$ . At the testing phase, we average the probabilities resulted from each frames of a given sentence to give the corresponding speaker. The experiment was conducted for different depths for each invariance scale 16, 32, 64 ms, in order to observe the effect of WST depth under CNN fusion. The architecture was trained using cross-entropy loss, the batch size was set to 256, the maximum epoch was set to 100. The optimizer used was ADAMAX with a learning rate  $5.10^{-3}$ .

**2.2.1.2.2 Speaker identification text-dependent** We visually demonstrate the benefit of depths  $> 2$  in speaker identification text-dependent task. We choose two speak-

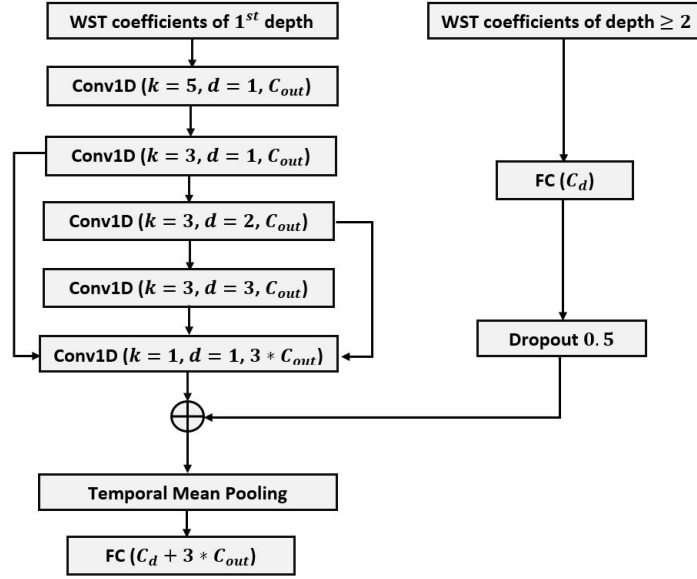


Figure 2.1 – Fusion of 1D-CNN architecture with WST ( $C_{out}$  is the number of output channels from a 1D temporal convolution, here it was set to 128,  $C_d$  is the number of WST coefficients of depth  $\geq 2$ ).

ers from TIMIT dataset reading the sentence *"she had your dark suit in greasy wash water all year"* under a clean speaking record environment, then we apply the WST for the optimal depth and invariance scales values found previously. To observe the impact of depth, we plot the spectrum at each depth for each speaker. One can say; the magnitude of the WST coefficients depends on the intensity of the speaker when he pronounces a word, therefore for visualization reasons, we choose to normalize (max-normalization) our data per scale in order to see which interference value is maximum for a speaker. The color-bar scale was adjusted to clearly visualize the differences. The WST coefficients were not log-normalized as in the previous part.

## 2.2.2 Results and Discussion

### 2.2.2.1 Speaker identification text-independent

**2.2.2.1.1 Optimal invariance scale and depth** To highlight the impact of the WST invariance scale and its depth on TIMIT dataset, we used a different combination of invariance scales (ms) and the number of depths to realize which structure best fits our data i.e compromise between classification accuracy per sentence and the time execution.

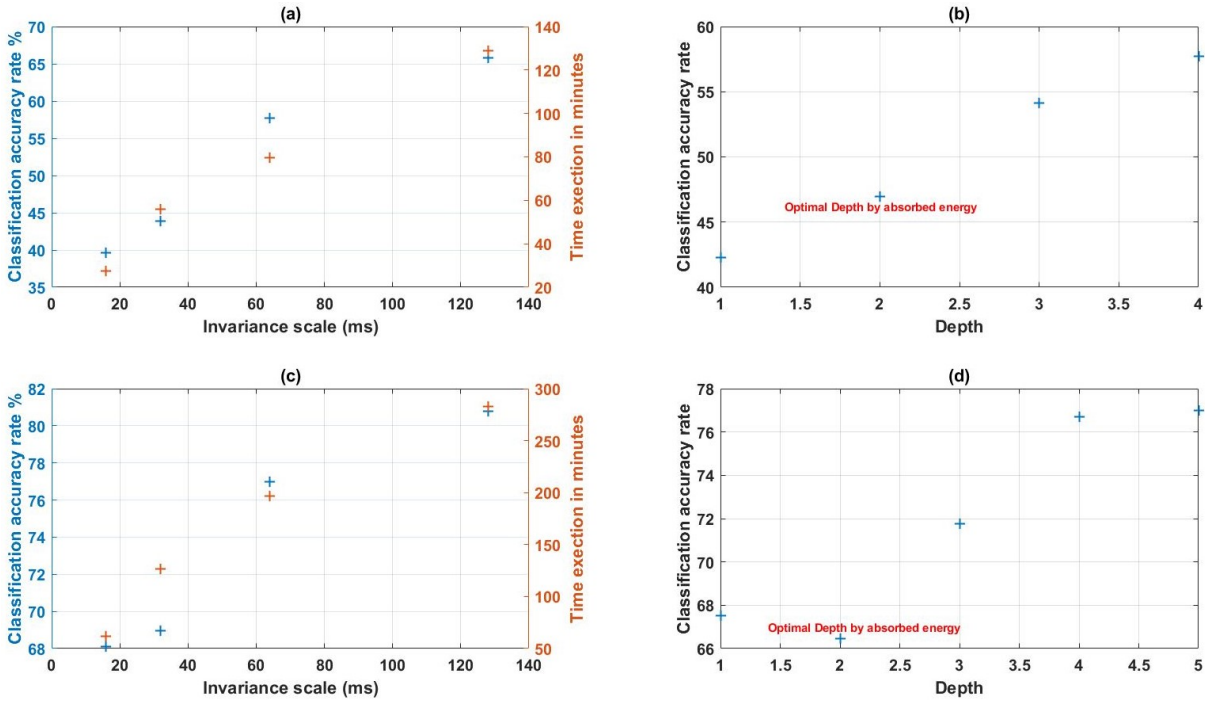


Figure 2.2 – Classification accuracy per sentence for different invariance scales when the audios are sampled at  $16\text{ kHz}$  (c) and at  $8\text{ kHz}$  (a), and for different depths of the best invariance scale  $64\text{ ms}$  sampled at  $16\text{ kHz}$  (d) and at  $8\text{ kHz}$  (b).

The Fig. 2.2 shows the maximum classification accuracy per sentence offered by a given invariance scale at its optimal depth when the sampling frequency is  $8\text{ kHz}$  (a) and  $16\text{ kHz}$  (c), the performance increases linearly for both sampling frequencies from  $16\text{ ms}$  to  $64\text{ ms}$  then we do not observe higher improvement after  $64\text{ ms}$ . Based on our criterion i.e with less time we get more performance, the optimal invariance scale is  $64\text{ ms}$ . From  $16\text{ ms}$  to  $64\text{ ms}$ , when the conditions are more constraint, i.e  $8\text{ kHz}$  sampling frequency, we see a  $18,04\%$  improvement in the classification accuracy per sentence, while for  $16\text{ kHz}$  sampling frequency the improvement was around  $8,87\%$ . At  $64\text{ ms}$ , after each increasing of depth, we get an average improvement of  $3,05\%$  for  $8\text{ kHz}$  (Fig. 2.2 (b)) and an average improvement of  $5,24\%$  for  $16\text{ kHz}$  (Fig. 2.2 (d)) until a stabilisation at the depth 4. These observations can be deduced theoretically from Eq. (38) in [33], where it affirms that increasing invariance scale and depth create more invariant features.

**2.2.2.1.2 WST + 1D-CNN** From the results shown in Fig. 2.2, a fusion of WST with a CNN is crucial to increase the classification accuracy. CNNs captures details from

the first depth of WST and depths  $\geq 2$  gives information on large structures. In Fig. 2.3 we report the results of the WST + 1D-CNN (Fig. 2.1) applied for different depths and invariance scales. The optimal invariance scale for WST + 1D-CNN that increases the classification accuracy per sentence is  $16\text{ms}$  instead of  $64\text{ms}$ , this can be explained by the locality of CNN and the temporal high resolution offered by  $16\text{ms}$ . Depending on the invariance scale used the optimal WST depth that improves the classification accuracy per sentence differs, yet increasing it always leads to an improvement of the classification accuracy per frame. To evaluate our architecture, a performance overview and the number of learnable weights of the baselines and the WST + 1D-CNN are given in the table 2.3. Compared to the baselines methods, the number of learnable weights required by our architecture is less by 94%, yet it outperforms HWSTCNN under  $8\text{kHz}$  sampling frequency condition by an improvement of 7,57%, and makes the same order of performance compared to the baselines under  $16\text{kHz}$  sampling frequency.

Table 2.3 – Classification accuracy per sentence of the best values of invariance scale and depth of WST + 1D-CNN and baselines trained and tested on TIMIT sampled at  $8\text{kHz}$  and  $16\text{kHz}$  ( $\#param$  is the number of learnable weights in each architecture).

<i>Architecture</i>	<b>8 kHz</b>	<b>16 kHz</b>
<b>CNN-raw</b> ( $\#param = 22.8M$ )	<b>97.75%</b>	98.91%
<b>SincNet</b> ( $\#param = 22.7M$ )	97.54%	<b>99.4%</b>
<b>HWSTCNN</b> ( $\#param = 18.1M$ )	85.93%	98.12%
<b>WST + 1D-CNN</b> ( $\#param < 1M$ )	93.5%	98.12%

### 2.2.2.2 Speaker identification text-dependent

To understand visually the behavior of WST across depths, we performed a comparison between the WST coefficients of two speakers reading the same sentence. Fig. 2.4 is the spectrogram of the scatter coefficients of the first and the second orders of the two speakers, the first depth presents two main peaks approximately located at 200 Hz which represents the pitch contour, and the second one at around 425 Hz. At the second depth, the invariant features are more enhanced compared to the first depth. At depths  $> 2$  in Fig. 2.5 for a given word, the distribution along frequencies differs strongly from a speaker to another, therefore going deeper generates more invariant features that their distribution along frequencies depends on speaker’s identity.



(a)		Invariance Scale (ms)		
		16	32	64
WST Depth	1	92.20%	90.25%	87.73%
	2	<b>93.50%</b>	91.99%	88.96%
	3	-	91.55%	90.33%
	4	-	-	89.89%

(b)		Invariance Scale (ms)		
		16	32	64
WST Depth	1	38.76%	40.15%	35.51%
	2	42.67%	42.82%	43.06%
	3	-	43.17%	51.56%
	4	-	-	<b>53.22%</b>

(c)		Invariance Scale (ms)	
		16	32
WST Depth	1	<b>98.12%</b>	96.96%
	2	96.89%	95.81%
	3	96.24%	95.38%
	4	-	95.31%

(d)		Invariance Scale (ms)	
		16	32
WST Depth	1	53.89%	54.89%
	2	59.76%	61.91%
	3	60.25%	65.88%
	4	-	<b>66.61%</b>

Figure 2.3 – The classification accuracy % per sentence (a-c) and the classification accuracy % per frame (b-d) for TIMIT sampled at  $8kHz$  (a-b) and  $16kHz$  (c-d) when using WST + 1D-CNN for different depths of invariance scales  $<$  the optimal invariance scale  $64ms$ .

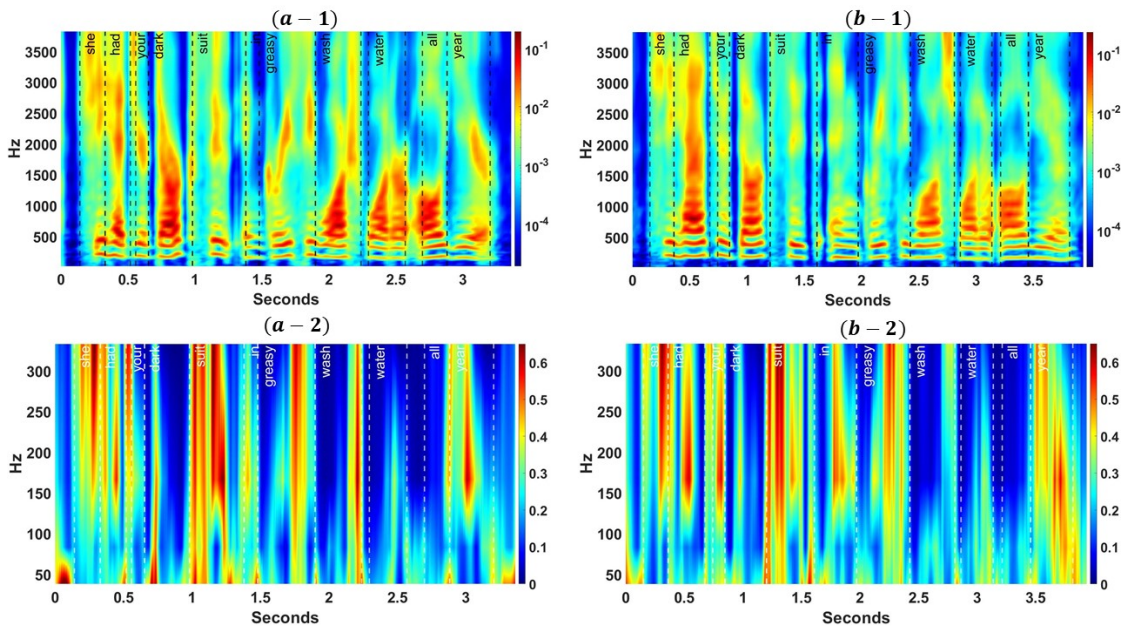


Figure 2.4 – Spectrogram of the WST coefficients at the first depth (first speaker (a-1) and the second speaker (b-1)) and the second depth (first speaker (a-2) and the second speaker (b-2)) (x-axis time and y-axis is frequencies).

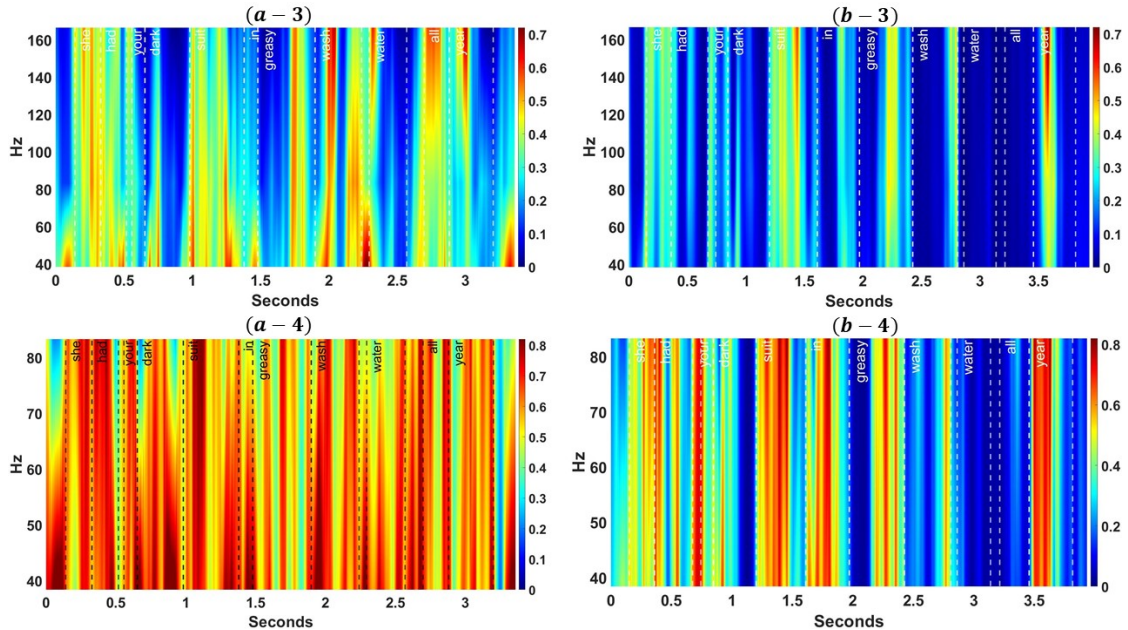


Figure 2.5 – Spectrogram of the WST coefficients at the third depth (first speaker (a-3) and the second speaker (b-3)) and the fourth depth (first speaker (a-4) and the second speaker (b-4)) (x-axis time and y-axis is frequencies).

### 2.2.3 Conclusion and perspectives

In this section, we have shown the importance of WST depth and its invariance scale for speaker identification. We have proved that instead of looking for energy concentration at the first two depths of WST, we should go deeper to generate invariant features. Experimental results on TIMIT have shown that the deeper WST can achieve dominant results with limited data. An optimized method based on a compromise between the classification accuracy per sentence and execution time has been successfully proposed to select a priori the best scatter transform architecture for speaker identification text-independent task. To enhance the classification accuracy per sentence and compete the SOTA, we have proposed a fusion between CNN and WST and we have shown that increasing the WST depth enhances the classification accuracy per frame. The resulted optimal values of WST invariance scale and depth, were used to observe visually the benefits in a speaker identification text-dependent task. These results show significant promise for considerable improvement in speaker identification.

As a possible way to enhance the identification accuracy is the insertion of a new modality to the system to add complementary informations to the audio modality. In

the next Section 2.3, we will present our self attention-based fusion system that balances between the audio and lips depth video modality applied on the person identification under ultra short utterances circumstances at the word level ( $< 1$  second).

## **2.3 Attention-based Fusion of Ultra-short Voice Utterances and Depth Videos for Multimodal Person Identification**

### **2.3.1 Methodology**

Drawing from the literature that explores various sources of information, such as those from diverse modalities or spatiotemporal data, our primary goal is to recognize individuals using depth visuals, dynamic lip movements, and brief audio utterances lasting less than one second. To efficiently extract valuable information from our chosen modalities, namely audio and depth, we have decided to employ a late fusion strategy. This approach entails feeding each modality into a network specifically designed to handle the structure of the respective information. The resulting feature vectors from both systems will then be combined to make a decision regarding the speaker’s identity. In the following sections, we will outline the CNN employed to extract features from each modality, based on existing literature.

Transitioning from one modality to another, we must ensure that the network architecture is designed to optimally process the specific data structure. This allows us to maximize the accuracy and efficiency of the speaker identification process. By adopting a late fusion strategy, we can effectively combine the strengths of each individual modality to improve overall performance.

As we delve further into this topic, we will discuss the details of the CNN used for feature extraction in each modality. By drawing upon the work done in the literature, we aim to create a robust and efficient system that can accurately identify speakers based on their depth visuals, dynamic lip movements, and short audio utterances.

#### **2.3.1.1 Voice Speaker identification**

To identify a speaker from brief speech segments. We used the SOTA speaker verification methods [19, 31, 28] and WST + 1D-CNN from our previous work 2.1. According to

our in Fig. 2.3, the WST depth and invariance scale for the sampling frequency  $16\text{ kHz}$  are 1 and  $16\text{ ms}$  respectively. For the purpose of time processing, we compare this architecture with other pre-trained models from SpeechBrain [46] and Hugging Face [47], which were originally trained on TIMIT [34] and VoxCeleb datasets [48]. We evaluated our architecture shown in Fig. 2.1 in terms of overfitting, performance and time processing (with the same PC capacities) and different models such as SincNet [28], Mel-Frequency Cepstrum Coefficients [49] (MFCC) + x-vectors [19] and MFCC + ECAPA-TDNN [31] to find the most suitable architecture for our KinectsDigits [36] and TCD-TIMIT [50] data. Since we have different and many words in TCD-TIMIT [50], it is wiser to choose this dataset for our preliminary comparison to select the customised audio architecture for our next experiments with all datasets. For time reasons, we eliminate end-to-end networks such as SincNet [28] as it requires high processing time.

Table 2.4 – Identification accuracy and the time processing of each method applied on 20 speakers TCD-TIMIT [50] re-sampled at  $16\text{ kHz}$ .

<i>Architecture</i>	<b>Identification Accuracy (train/test)</b>	<b>Time Processing</b>
<b>WST + 1D-CNN (our architecture Fig. 2.1)</b>	99.21%/96.67%	~ 45 min
<b>MFCC + x-vectors [19]</b>	99.25%/95.21%	~ <b>3 min</b>
<b>MFCC + ECAPA-TDNN [31]</b>	97.71%/95.84%	~ 15min

The results presented in Table 2.4 demonstrate:

- The time limitations of our proposed method, WST + 1D-CNN, under ultra-short utterances (less than 1 second), a result from the fact that the WST coefficients are extracted using a MATLAB implementation, which is the only existing method capable of capturing deeper feature representations of WST. However, this implementation is not optimized for speed, making it slower compared to extracting the hand-crafted features MFCC [49] using Python,
- The ability of MFCC + x-vectors to handle dilemma identification accuracy, overfitting and time processing under ultra-short utterance circumstances. This architecture will be used for our future experiments and we will only use the name x-vectors.

The performance differences between these models presented in Table 2.4 leads to choose x-vectors [19] due to its short time processing. Furthermore, for the VoxCeleb2 dataset [48], we fine-tuned the x-vectors [19] architecture using ultra short utterances ( $< 1\text{ s}$ ), after initially conducting transfer learning with x-vectors [19] pre-trained on both

VoxCeleb1 and VoxCeleb2 [48, 47] datasets.

All audio samples from the three benchmark datasets were either originally sampled or resampled at a  $16\text{ kHz}$  frequency. Before inputting the audio signal into the x-vectors [19] architecture, a MFCC [49] transform with a  $25\text{ ms}$  frame length and a  $10\text{ ms}$  overlap was applied. The cepstrum coefficients were mean normalized across word duration instead of the 3 seconds mentioned in the reference article [19], resulting in a feature vector  $X_a \in \mathbb{R}^{512}$ . For training x-vectors [19] on ultra short utterances from VoxCeleb2 [48], a batch size of 32 was used, along with the ADAMAX optimizer and a learning rate of  $5.10^{-3}$ .

### 2.3.1.2 Lip identification

In the realm of voice production, a robust correlation exists between lip movement and the ensuing audio, as the lip serves to modulate the vibrations of the vocal cords, particularly within the lower frequency portions of the audio spectrum. Literature has demonstrated that these low-frequency components encompass both dynamic and visual biometric elements [25]. Guided by this knowledge, we have devised a novel spatiotemporal architecture, which aims to extract not only spatial features, but also temporal characteristics from the used data.

The 3D convolutional neural network is a popular architecture for feature extraction from videos [51]. However, this architecture demands a significant number of parameters and more time compared to 2D CNN. To address this challenge, we propose projecting the 3D data (2D space + time) onto all possible 2D combinations, namely  $(X, Y)$ ,  $(X, T)$ ,  $(T, Y)$ . Subsequently, each 2D combination undergoes mean normalization as a distance calibration and is fed into a ResNet18 neural network pretrained on ImageNet [20], as delineated in Figure 2.6.

For every view -  $(X, Y)$ ,  $(X, T)$ , or  $(T, Y)$  - the network produces a two-dimensional output matrix,  $Y_{in} \in \mathbb{R}^{512 \times n}$ , where  $n$  represents the quantity of frames along the absent dimension. The first pair of static parameters across this missing third dimension is gleaned through an attentive statistic pooling technique [31].

The input  $Y_{in}$  of this module passes by a tanh activation and 1D-Conv (same input-output channels) layer to calibrate the channel weight (first line in Eq. 2.1), the output score  $Y_{in1} \in \mathbb{R}^{512 \times n}$  is normalized across the missing dimension by using the softmax function, the generated weights  $W_{c,i}$  ( $c = 1, \dots, 512$  is the channel index and  $i = 1, \dots, n$  is the frame index) are then used to compute the weighted mean  $\mu_c$  and the standard

deviation  $\sigma_c$  across the third axis.

This module describes importance to a frame (space or time) contingent on their ability to augment feature invariance, ultimately yielding a vector  $Y_{\text{out}} \in \mathbb{R}^{1024}$  for each view  $\{(X, Y), (X, T), (T, Y)\}$ .

$$\begin{aligned}
 Y_{\text{in}1} &= \text{Conv}(k = 1) (\tanh(Y_{\text{in}})) \in \mathbb{R}^{512 \times n}, \\
 W_{c,i} &= \frac{\exp(Y_{\text{in}1}(c, i))}{\sum_{i=1}^n \exp(Y_{\text{in}1}(c, i))}, \\
 \mu_c &= \sum_{i=1}^n W_{c,i} Y_{\text{in}}(c, i), \\
 \sigma_c &= \sqrt{\sum_{i=1}^n W_{c,i} (Y_{\text{in}}(c, i) - \mu_c)^2}, \\
 Y_{\text{out}} &= (\mu_1, \dots, \mu_{512}, \sigma_1, \dots, \sigma_{512})^T \in \mathbb{R}^{1024}.
 \end{aligned} \tag{2.1}$$

To derive a holistic understanding of the video, the resultant vectors from the attentive static pooling operation  $Y_{\text{out}(X,Y)}$ ,  $Y_{\text{out}(T,Y)}$ , and  $Y_{\text{out}(X,T)}$  are projected onto the space  $\mathbb{R}^{512}$  through a 1D-convolution employing a kernel size of 1, followed by batch normalization and the implementation of a tanh activation function. The projected vectors are concatenated and averaged utilizing the self-attention module.

We introduce a self-attention module, built upon the foundation of the attentive static pooling concept, with the distinct purpose of optimizing both channel and view characteristics. The vector  $Z_{\text{in}} \in \mathbb{R}^{512 \times 3}$ , which arises from the concatenation of the three views, is subjected to a tanh activation and a 1D-Conv layer with the same input-output channels of 512, enabling channel weight calibration. Following this, the transposed output vector  $Z_{\text{in}1}$  is passed through a tanh activation and a 1D-Conv layer with the same input-output channels of 3, as illustrated by the second equation in Eq. 2.2. This final step allows for the calibration of the weights of the views.

The output vector  $Z_{\text{in}2}$  is normalized across the views axis by using the softmax function, the generated scores  $W'_{c,i}$  ( $c = 1, \dots, 512$  is the channel index and  $i = 1, 2, 3$  is the view index) are then used to compute the weighted mean  $\mu'_c$  across the views axis. Therefore, the generated vector  $Z_{\text{out}} \in \mathbb{R}^{512}$  captures the comprehensive spatio-temporal information of the video and shares identical dimension with the audio feature vector

$$\begin{aligned}
 Z_{in1} &= \text{Conv}(k=1) (\tanh(Z_{in})) \in \mathbb{R}^{512 \times 3}, \\
 Z_{in2}^T &= \text{Conv}(k=1) (\tanh(Z_{in1}^T)) \in \mathbb{R}^{3 \times 512}, \\
 W'_{c,i} &= \frac{\exp(Z_{in2}(c,i))}{\sum_{i=1}^3 \exp(Z_{in2}(c,i))}, \\
 \mu'_c &= \sum_{i=1}^3 W'_{c,i} Z_{in}(c,i), \\
 Z_{out} &= (\mu'_1, \dots, \mu'_{512})^T \in \mathbb{R}^{512}.
 \end{aligned} \tag{2.2}$$

To train this depth view fusion system, a batch size of 32 was used, along with the ADAMAX optimizer and a learning rate of  $10^{-2}$ .

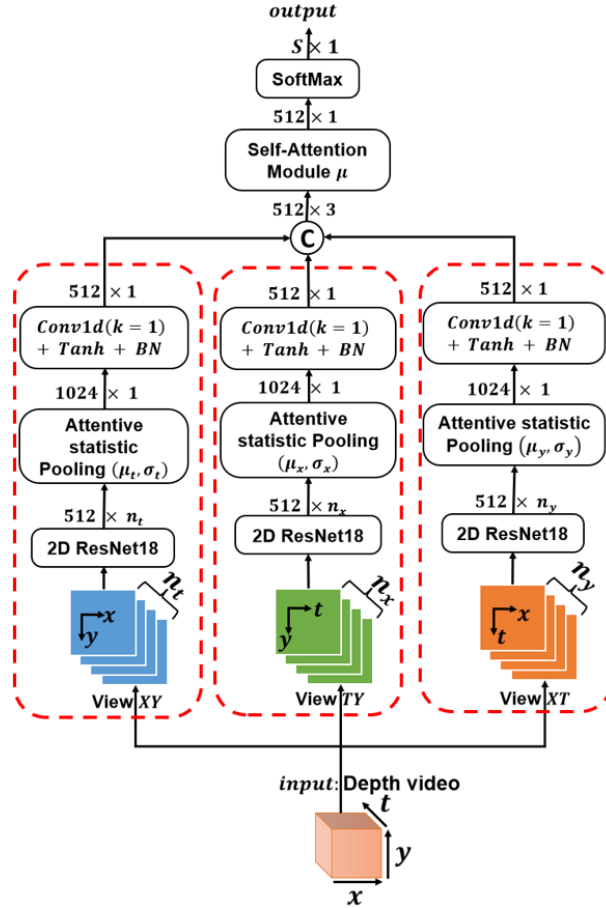


Figure 2.6 – Multi-view Video CNN architecture used on lip depth videos (The red dashed line represent the extraction of the features vector from the view projection of the video).

### 2.3.1.3 Fusion of Depth and Audio modalities

In our proposed method, we employ a late fusion architecture, as illustrated in figure (2.7). This architecture utilizes distinct networks designed to handle the unique information present in each modality, ensuring each modality’s specific characteristics are appropriately processed. For the audio component, we apply a 1D-convolution with a kernel size of 1 to the x-vector outputs that are 1D vector in  $\mathbb{R}^{512}$ . This procedure is further enhanced with batch normalization and a tanh activation function. For the depth video (2D Spatial + Time) modality, we utilize a multi-view CNN (as shown in figure (2.6)) on the depth video to extract a feature vector  $X_d \in \mathbb{R}^{512}$ , encapsulating both visual and dynamic lip movement information.

Following the individual processing of these modalities, the resultant vectors are routed through our self-attention module (illustrated in figure (2.7)). This module calculates the weighted sum of the two vectors as per Equation 2.2. Our approach is informed by the insights of [36], underscoring the advantages of using weighted fusion. By utilizing this strategy, we minimize potential redundancy or ambiguity associated with each modality, reducing contradictions and significantly boosting the overall performance of our model.

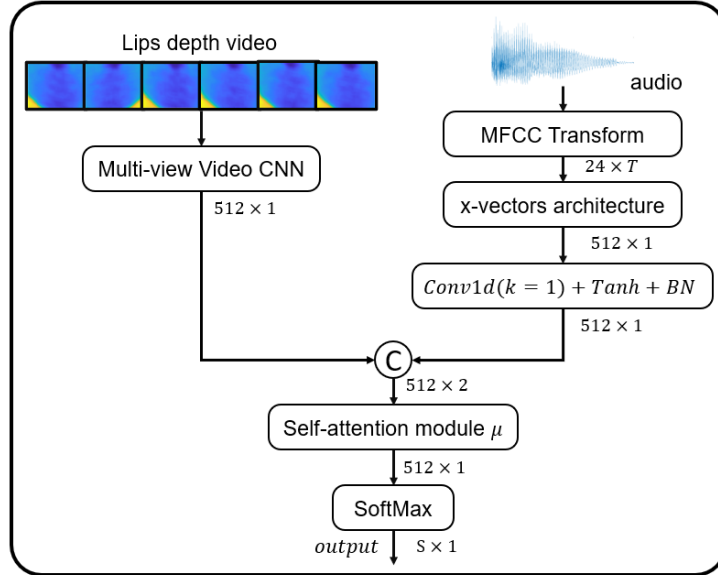


Figure 2.7 – Architecture of late fusion of the two modalities (audio and depth video).



## 2.3.2 Experimental results and Discussion

### 2.3.2.1 Data Collection

To evaluate our proposed methodology, we used three well-known benchmark datasets: KinectsDigits [36], VoxCeleb2 [48], and TCD-TIMIT [50].

KinectsDigits [36] is a valuable resource for multimodal speaker recognition research. The dataset, captured using Microsoft’s Kinect sensor, consists of RGB, Depth, and Audio modalities. It features recordings of individuals lips articulating digits from 0 to 9 under various environmental conditions. We only kept one situation to not have repeated words. The videos have a resolution of  $104 \times 80$  pixels, a frame rate of  $30 \text{ fps}$ , and a voice signal sampling rate of  $16 \text{ kHz}$ .

VoxCeleb2 [48] is an essential resource for speaker recognition research, containing RGB and Audio modalities from a diverse range of speakers. The dataset features over a million utterances from thousands of speakers in various accents, languages, and speech content. Videos in VoxCeleb2 have a  $25 \text{ fps}$  and a low resolution  $224 \times 224$ , while audio signals are sampled at  $16 \text{ kHz}$ . As we have very low resolution, we suggested that the depth estimation network [52] will give less details about the lip region as shown on figure (2.8), therefore we only select randomly 1000 speakers from this dataset.

TCD-TIMIT [50] is a valuable resource for multimodal speech recognition research, comprising continuous speech in both RGB and Audio modalities. The dataset features individuals delivering sentences from the TIMIT corpus. Videos in the TCD-TIMIT database have a resolution of  $1920 \times 1080$  pixels, a frame rate of  $29 \text{ fps}$ , and a voice signal sampling rate of  $48 \text{ kHz}$ . Only audio signals were re-sampled to  $16 \text{ kHz}$ .

While KinectsDigits [36] includes RGB, Depth, and Audio modalities, VoxCeleb2 [48] and TCD-TIMIT [50] only provide RGB and Audio data. Our experimental approach focused on the depth information of the speakers’ lip, present in KinectsDigits [36] but not in the other two datasets. To ensure consistency across all datasets, we applied a lip detection and depth estimation technique to the RGB videos of the three databases, obtaining the same modalities for the three benchmarks. The following subsections outline our lip detection and depth estimation strategies.

#### *Depth estimation*

The primary aim of this research is to demonstrate the synergistic value of depth and audio data for multimodal person identification. However, most publicly accessible datasets

suitable for this purpose only include RGB video and audio modalities. To address this limitation, we utilize a pre-trained face depth estimation network (employing an Encoder-Decoder architecture) [52] to generate pseudo depth map from full RGB facial images. This method, trained on an extensive synthetic face dataset with varying head poses, expressions, backgrounds, and image resolutions, outperforms state-of-the-art models on real face depth datasets such as Pandora [53], Eurecom Kinect Face [54], and Biwi Kinect Head Pose dataset [55]. Consequently, this model can produce depth modality for RGB faces in datasets.

The depth estimation network was independently applied to each facial frame within a given video for Voxceleb2 and TCD-TIMIT. For the KinectsDigits dataset, we retrained the depth estimation network on Pandora [53] lip regions to enable a fair comparison between estimated and actual depth in the person identification task. Although the resulting depth estimation may be considered low-resolution due to some missing depth information, it remains a valuable data source for our analysis, as evidenced in the results section. The figure (2.8) presents a sample of real or estimated depth mouth crop frames for the three benchmark datasets.

#### ***Lip extraction***

In our approach to data preparation, we employed an existing method for lip detection and applied it to two supplementary datasets, namely VoxCeleb2 [48] and TCD-TIMIT [50]. The aim was to accurately identify and extract the lip region of speakers in RGB videos for further analysis. To accomplish this, we utilized Google’s Mediapipe tool. By leveraging the power of Mediapipe, we were able to identify the key landmarks associated with the mouth region. This allowed us to isolate the lower part of the face on the depth video and focus our attention on the lip.

#### ***Word segmentation***

In our investigation of the ultra-short utterance problem, we segmented words from the audio in both the VoxCeleb2 and TCD-TIMIT datasets (while KinectsDigits audio only contained digit words from 0 to 9). We removed duplicate instances of the same word for each speaker to concentrate on the text-independent speaker identification task. Subsequently, we interpolated the resulting timestamps on the video to extract the lip video section corresponding to the spoken word. We utilized models from the Kaldi toolkit for word segmentation, which are available for various languages. It is worth noting that for some words, we obtained a single frame, which the spatiotemporal network processed as 3D data, just like longer videos.

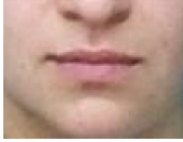
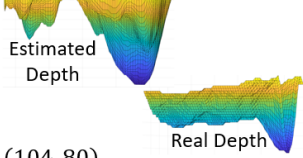
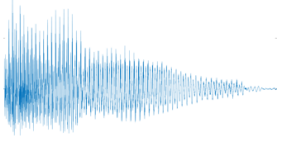
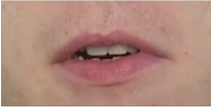
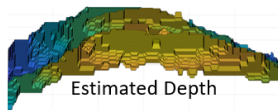

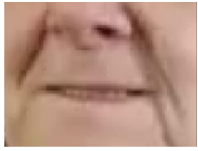
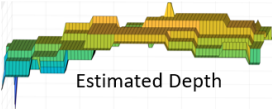
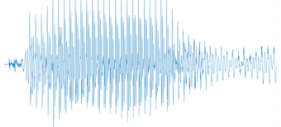
Database	RGB	Real or Estimated Depth	Audio
Kinects Digits	 (104, 80)	 (104, 80)	 $F_s = 16kHz$
TCDTIMIT	 (300, 150)	 (300, 150)	 $F_s = 48kHz$
Voxceleb2	 (60, 60)	 (60, 60)	 $F_s = 16kHz$

Figure 2.8 – General Informations on the three benchmarks used during the study ( $F_s$  represents the frequency sampling before the re-sampling to  $16kHz$ , the audio curves are full utterance or a part of a word spoken by a speaker in each dataset). Real depth refers to the depth map obtained directly from a depth camera. In contrast, estimated depth stands to depth map generated from RGB images using the method outlined in paragraph 2.3.2.1.

Table 2.5 presents pertinent information regarding each dataset and its usage, with an 80%|20% train|test split. Remarkably, up to 95% of the data has a duration of less than 1 second, demonstrating that the task is focused solely on short utterance identification.

Table 2.5 – Data characteristics after word segmentation.

Dataset/ Information	#spk	#avg/spk	(min, max, mean)	Fps	Mouth crop ( $h_x, h_y$ )
Kinects Digits [36]	30	10	(35 ms, 2.02 s, 600 ms)	30	(104, 80)
TCD-TIMIT [50]	59	470	(35 ms, 2.07 s, 610 ms)	29	(300, 150)
VoxCeleb2 [48]	1000	815	(40 ms, 1 s, 520 ms)	25	(60, 60)

### 2.3.2.2 Results

This section addresses the experimental results of the proposed architecture in figure (2.7) and its components applied on the three datasets. We investigate the performance of our spatiotemporal fusion system, the importance of adding depth information to audio,

and the complementary information added by each modality. To understand how the fusion of modalities can improve the performance of the system, we first examine each individual modality and their identification capabilities.

We begin our discussion by exploring the impact of depth video encoding. The fusion of all three views, as depicted in figure (2.6), led to a noticeable improvement of approximately 2% in performance, as detailed in Table 2.7. This improvement validates the significant advantage of integrating both spatial and temporal information into a single modality. Furthermore, it is noteworthy to emphasize the critical role the self-attention module played in these results. By finely balancing the spatial and temporal information, this module effectively amplified the overall performance of the speaker identification process.

To provide additional insights, we dove deeper into the contribution of each view towards person identification based on the depth video of their lip. To achieve this, we dissected the identification performance for each 2D view of the video using the architecture illustrated in figure (2.6). The ensuing findings, presented in Table 2.6, allow for a thorough analysis of the various views.

The analysis unveils that the  $(XY)$  view, which encapsulates the spatial information of the video, has a substantial edge over the other views. Its performance improvement is quite significant: 5% for the KinectsDigits dataset trained on real or estimated depth, 9.27% for TCD-TIMIT, and 8.08% for VoxCeleb2. Following the  $(XY)$  view, the horizontal view  $(XT)$ , which covers the horizontal movements of the lip, shows remarkable results on the real depth KinectsDigits dataset [36]. In fact, it surpasses the  $(TY)$  view by 8.33%, indicating the importance of temporal dynamics. Furthermore, the performance improvement from depth estimated video reveals an interesting prospect of considering the valuable information provided by estimated depth.

Table 2.6 – Identification accuracy per each view of depth video. Real depth refers to the depth map obtained directly from a depth camera. In contrast, estimated depth stands to depth map generated from RGB images using the method outlined in paragraph 2.3.2.1.

<i>View/ Dataset</i>	<i>XY</i>	<i>XT</i>	<i>TY</i>	<i>Spatio-temporal fusion</i>
<b>KinectsDigits (Real Depth) [36]</b>	96.66%	91.66%	83.33%	<b>98.33%</b>
<b>KinectsDigits (Estimated Depth) [36]</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>
<b>TCD-TIMIT (Estimated Depth) [50]</b>	96.20%	83.33%	86.93%	<b>98.58%</b>
<b>VoxCeleb2 (Estimated Depth) [48]</b>	11.82%	8.62%	8.58%	<b>17.49%</b>

Diving into another critical aspect of our analysis, the audio component of videos, we discover its pronounced significance in the sphere of speaker identification tasks. We used

the x-vectors architecture, a powerful method for dealing with audio data. However, our results, as depicted in Table 2.7, highlight a considerable drop in performance, primarily attributed to the short duration of the voice samples. This brevity tends to trigger substantial overfitting, depending on the specificities of the dataset.

We found that the audio-based identification performs rather well in controlled recording environments, such as those present in the KinectsDigits [36] and TCD-TIMIT [50] datasets, where the identification accuracy surges beyond 75%. Nonetheless, in scenarios more reflective of real-world conditions, as is the case with the VoxCeleb2 [48] dataset, the accuracy nosedives to nearly half of its initial value. Such a drastic dip in accuracy raises concerns about the reliability of the system for accurate speaker recognition in practical applications.

The combination of different modalities can, however, offer a solution to this problem. Integrating various modalities not only harnesses their individual strengths but also supplements each other’s limitations. Particularly, an effective fusion of audio and depth or estimated depth data can create a more robust and efficient system. Our results support this claim, as shown in Table 2.7, where we observed a marked enhancement in performance when both modalities were utilized in unison.

Upon fusing the modalities, the performance of audio identification shot up by 20%, while the real or estimated depth’s spatiotemporal encoding registered an average boost of 1.5%. These findings validate the significant potential of integrating the strengths of each modality. The result is a more potent speaker identification system capable of adjusting to diverse conditions and tackling the inherent challenges associated with short utterances and noisy recording environments effectively.

Table 2.7 – Identification accuracy per each modality and their fusion for the three benchmarks. Real depth refers to the depth map obtained directly from a depth camera. In contrast, estimated depth stands to depth map generated from RGB images using the method outlined in paragraph 2.3.2.1.

<i>Modality/ Dataset</i>	<i>Audio</i>	<i>Depth</i>	<i>Multimodal Fusion</i>
<b>KinectsDigits (Real Depth) [36]</b>	75%	98.33%	<b>100%</b>
<b>KinectsDigits (Estimated Depth) [36]</b>	75%	<b>100%</b>	<b>100%</b>
<b>TCD-TIMIT (Estimated Depth) [50]</b>	89.25%	98.58%	<b>99.76%</b>
<b>VoxCeleb2 (Estimated Depth) [48]</b>	56.03%	17.49%	<b>64.11%</b>

### 2.3.2.3 Discussion

These results not only emphasize the significant contribution of depth information to the overall performance, but they also highlight the valuable supplementary information provided by the audio component. By effectively combining these two modalities, it is possible to develop more advanced and reliable systems that capitalize on the strengths and complementary nature of both audio and depth information. Building on this observation, addressing the issue of speaker identification reliability requires considering both audio and visual modalities. Fusing the information from these modalities results in a more robust and effective system that can overcome limitations posed by short utterances and variable recording environments. This multimodal approach significantly enhances the performance of speaker identification systems, ensuring their effectiveness in various real-world applications.

In speaker identification tasks, our spatiotemporal fusion method (illustrated in figure (2.6)) is an efficient way to process depth videos, while deliberately avoiding the use of resource-intensive 3D CNNs. The logic behind steering clear of 3D CNNs is based on their inherent drawbacks, such as the high computational demand and memory requirements which can impede their use in real-time applications or on devices with limited resources. Furthermore, the greater number of parameters in 3D CNNs has potential implications for longer training times and increased susceptibility to overfitting, especially when working with datasets of limited size. In contrast, our chosen approach, which relies on a more efficient spatiotemporal fusion method, has the advantage of pulling out distinctive features from real or estimated depth videos, thus ensuring comparably strong performance in speaker identification tasks without facing the challenges tied to the use of 3D CNNs, such as high computational complexity and prolonged training periods.

This spatiotemporal fusion architecture consists of multiple 2D views, and an investigation into the contribution of each view is required. For the real depth data, specifically the KinectsDigits [36] dataset, the horizontal view of the mouth ( $XT$ ) outperforms the vertical view ( $TY$ ). This could be attributed to the mouth’s greater horizontal length, which enhances discriminant dynamic features. However, the ( $TY$ ) view exhibited more pronounced and rapid overfitting, suggesting its susceptibility to overfitting in this context. It is crucial to select an appropriate mouth crop resolution, as the speaker identification performance can be significantly impacted by the resolution or depth estimation, as seen in the VoxCeleb [48] dataset.

Considering the spatial view ( $XY$ ), it outperforms other fusions for both estimated

and real depth. This could be due to visual features conveying the 3D shape of the human face, while dynamic features only represent low frequencies of voice and behavior. The limited context for temporal features in short utterances might also contribute to this disparity. As depth resolution increases, the gap between the  $(XY)$  view and other views grows larger, suggesting the increasing importance of spatial information for distinguishing speakers with improved resolution, leading to enhanced discriminative power for the  $(XY)$  view.

Another key element of our discussion is the analysis of errors that occur during the speaker identification process. Several noteworthy patterns emerge after examining the errors made by our proposed model. The primary source of error, especially for the audio modality, appears to be ultra short utterances. Shorter utterances provide limited temporal context, and the brevity of these utterances can lead to the omission of certain speaker-specific cues, thus causing the model to underperform in these cases. This issue is compounded in the VoxCeleb2 dataset, where the utterances are notably brief and the recording conditions are significantly more varied than in the other datasets, thus challenging the model’s performance.

Additionally, another pattern we have noticed is the decrease in identification accuracy as the environmental noise increases, particularly with the VoxCeleb2 dataset. This dataset features recordings in diverse and often challenging conditions, such as outdoor recordings or instances with significant background noise. These factors can mask or distort speaker-specific cues, making it harder for the model to correctly identify the speaker.

In our examination of the depth video modality, we found that errors frequently stemmed from the low resolution of the RGB videos, particularly noticeable in the VoxCeleb2 dataset. This resulted in less accurate depth estimation, which in turn affected the precision of the spatial representation of lip movements, subsequently lowering identification performance. It’s worth noting that the depth estimation algorithm is specifically trained on facial features and thus performs optimally with high-resolution RGB videos, as is the case with the KinectsDigits and TCD-TIMIT datasets. Therefore, when applied to lower resolution videos like those in the VoxCeleb2 dataset, the depth estimation’s accuracy diminishes. The lower resolution videos also struggle to capture the finer details of lip movements, limiting the depth modality’s discriminatory ability and potentially leading to identification errors.

### **2.3.3 Conclusion**

This study demonstrates the effectiveness of a multimodal approach for speaker identification, incorporating both audio and depth information to achieve more accurate and reliable results. Through the examination of three benchmark datasets – KinectsDigits, VoxCeleb2, and TCD-TIMIT – we have shown that the fusion of these modalities leads to significant improvements in identification performance. The proposed spatiotemporal architecture effectively extracts spatial and temporal features from depth information, while the x-vectors architecture processes the audio modality.

Our findings highlight the importance of integrating multiple modalities to overcome the limitations posed by short utterances and variable recording environments. By fusing audio and depth data, we have achieved enhanced performance in a range of scenarios, including clean recording environments and more realistic, constrained situations. The results indicate that depth information plays a crucial role in performance enhancement, with the addition of audio providing complementary benefits.

This research contributes to the advancement of speaker identification systems by proposing a robust multimodal approach. In our future work, we aim to propose an architecture based on an early fusion, taking into account the high correlation between the lip movement and the voice generated. This would potentially lead to further improvements in performance and system reliability.

Future work can also extend this study by exploring the integration of other modalities or refining the fusion techniques to further improve performance. Additionally, the application of the proposed methodology to other related tasks, such as speaker verification or emotion recognition, could provide valuable insights and contribute to the development of more advanced and reliable systems in these domains.





# TOWARD COMPREHENSIVE SHORT UTTERANCES MANIPULATIONS DETECTION IN AUDIO-VISUAL VIDEOS

---

In the previous chapter, we discussed the advantages of multimodality in the biometrics fields. In this chapter, we will discuss the threat that has been evolving in the recent years, the deepfake, in other words, the multimedia manipulations. This chapter explores the need to integrate counterfeit detection mechanisms for robust security measures. Section 3.1 starts with an introduction about multimodal deepfake generation and detection. Section 3.2 reviews existing literature in audio and visual deepfake detection, emphasizing the limitations and computational challenges of current approaches. Section 3.3 elaborates on our hand-crafted methods for audio and visual deepfake detection, as well as our multimodal deep learning model. Section 3.4 details the experimental design and SOTA datasets used to evaluate our methods. Section 3.5 lists the results of all the experiments that we conducted in this chapter to evaluate our methods. Section 3.6 presents a comprehensive analysis of our results, followed by Section 3.7 that concludes the chapter and discusses future research directions.

## 3.1 Introduction

In an era defined by the rapid advancements in artificial intelligence and Generative Adversarial Networks (GANs), the manipulation of multimedia content has become both more sophisticated and more accessible [56, 57]. This growing accessibility not only unveils a new frontier of creative and practical possibilities but also precipitates a significant challenge in the form of multimedia manipulations that are increasingly harder to detect. Such manipulations have serious implications as they can mislead both human judgment and automated systems, jeopardizing the integrity of information channels [58]. Exist-

ing countermeasures for manipulations detection often hinge on resource-intensive deep learning models, thus limiting their applicability in real-world scenarios with constrained computational resources [18, 59, 60, 61].

The latest developments in deepfake technology have greatly influenced multimedia manipulation, leading to the creation of various methods capable of generating highly realistic synthetic media [62]. Notably, lip-syncing algorithms, which alter lip movements in videos to align with specific audio tracks, have gained significant attention [63, 62]. These techniques create a convincing illusion that the person in the video is speaking the provided audio. Previous research has indicated that detecting manipulations in the lower part of the face, particularly the lips and mouth, is more challenging compared to other facial features like the eyes, especially when using deep neural networks such as XceptionNet [64, 65]. This is a critical aspect of our study, as many deepfake techniques target the lower facial region to create deceptive effects [56]. The significance of the lip area is further amplified in scenarios where the upper face is obscured, making other facial features less reliable for verifying authenticity. By focusing on lip area manipulation detection, our research tackles a vital component of deepfake techniques and seeks to offer a robust solution for ensuring the authenticity of digital media.

Voice conversion represents a recent advancement in deepfake technology, enabling the transformation of one person’s voice to closely mimic the vocal characteristics of another [57, 66]. This technology can be combined with visual manipulations to create even more convincing synthetic media [67, 57]. Additionally, text-to-speech synthesis has advanced to the point where synthetic voices are becoming nearly indistinguishable from human voices, facilitating the creation of highly realistic audio content from text inputs [68, 57, 66]. While these developments are impressive, they highlight the critical need for effective detection mechanisms to ensure the authenticity of multimedia content in this new era of synthetic media.

One critical and under-addressed area within this context is the detection of manipulated short utterances in both visual video and audio content. The paucity of data in short sequences exacerbates the challenge of reliable detection. Algorithms find it increasingly difficult to discern patterns from limited information, creating an acute need for efficient and effective models.

Against this backdrop, our work in this field introduces several key contributions:

- Novel Detection Method: Introduced a specialized shallow learning technique for detecting deepfake content by analyzing the visual and auditory components of

- multimedia, specifically targeting the lower region of the face in videos.
- **Focus on Ultra-short Segments:** Optimized the detection method for ultra-short video segments ranging from 200 ms to 600 ms, addressing a significant challenge in existing methods.
  - **Multi-scale Audio Analysis:** Employed a multi-scale analysis for audio features using the wavelet scattering transform (WST), which effectively captures essential frequency characteristics of audio signals.
  - **High-frequency Video Analysis:** Developed a video feature extraction method based on high-frequency spatial analysis using discrete cosine transforms (DCT), focusing on the lip region for enhanced detection accuracy.
  - **Versatility:** Designed the method to be versatile, applicable in both unimodal and multimodal settings, leveraging visual and auditory cues for a comprehensive evaluation.
  - **Performance and Efficiency:** Demonstrated that the proposed method not only excels in accuracy for ultra-short segments but also scales efficiently to longer video lengths, making it suitable for real-world applications with constrained computational resources.
  - **Lastly,** for environments endowed with substantial computational resources, we introduce a state-of-the-art multimodal deep learning model [45], synthesizing both audio and visual data for enhanced detection accuracy.

Moreover, the complexity of detection becomes increasingly arduous with shorter iteration cycles in the generation of deepfakes. We, therefore, propose a testing strategy that seeks to assess the robustness of our models, accounting for various types of manipulations and different levels of compression in tampering methods.

The contribution of this work lies not only in the introduction of these novel methods but also in their potential to fill the gap between high computational requirements and real-world applicability. Our hand-crafted methods, in particular, offer a viable alternative for real-world scenarios where the deployment of resource-intensive deep learning algorithms is not feasible.

To rigorously validate our contributions, we perform extensive evaluations using multiple benchmark datasets, thereby confirming the efficacy and generalization ability of our methods. Our results indicate that when computational resources are not a limitation, our multimodal deep learning model performs exceptionally well, while the hand-crafted methods demonstrate unparalleled performance in resource-constrained environments.

The work presented was split into hand crafted and deep learning method, they were respectively published in the journal *Multimedia Tools and Applications* [69] and in the *IMPROVE* conference 2024 [70].

## 3.2 Related Works

Previous studies in deepfake detection have charted a range of technical methodologies for analyzing audio and visual components, employing strategies that extend from detailed low-level hand-crafted features to advanced high-level neural network architectures [60, 57].

In the realm of combating audio deepfakes, the use of Mel-Frequency Cepstral Coefficients (MFCC) is widespread, analyzed through 2D neural network architectures such as VGG16 and EfficientNet [59, 71, 18]. These methods are promising, yet they demand significant computational resources and have not been examined when dealing with ultra-short audio samples. In this chapter, we propose to explore shallow learning and deep learning methods which require less resources.

A. Pianese et al. [72] employed a distinct approach for audio deepfake detection by harnessing the Person of Interest (POI) concept, echoing the core ideas of speaker verification systems [31]. Their strategy focuses on assessing the similarity between the voice under scrutiny and a pre-existing reference collection of the claimed identity, employing two unique non-supervised methods: centroid-based and maximum-similarity testing [72]. The primary challenge of this method is its reliance on a comprehensive reference set for each identity analyzed. Our methodology aims to address this limitation by proposing an alternative approach that minimizes the need for such extensive reference collections, thereby enhancing the practicality and scalability of audio deepfake detection.

Visual-based deepfake detection methods have seen a diverse range of strategies. Some leverage 3D networks for in-depth sequence analysis [60, 17]. Zhou et al. [17] proposed a system that exploits the intrinsic synchronization between audio and visual elements, particularly focusing on the lips' movement and corresponding audio at the word level. Employing a multimodal neural network, they experimented with three types of fusion mechanisms based on attention mechanisms [17]. However, it's noteworthy that [17] employed 3D networks for video feature extraction and attention mechanisms, processes that are known for their high resource demands.

On the other side, alternative approaches to deepfake video detection primarily em-

ploy image-based methods, placing a significant focus on facial features as discussed in references [60, 18]. In other words, these methods involve the independent analysis of each frame within the video input by the network, culminating in a conclusive decision through hard or soft voting [60, 18]. While these techniques demand fewer resources compared to their 3D counterparts, they lack the incorporation of temporal information essential for thoroughly examining videos.

Building on these contributions, our work aims to address some of the identified limitations. Firstly, our methodologies designed for audio-visual detection rely on shallow learning and only the lips part of the face, demanding fewer resources and demonstrating a comprehensible and interpretable nature. Our visual deepfake detection is based on sequence level, in other words, we exploit the temporal anomalies with less computational cost. Secondly, based on our previous [45], we tune our architecture that dissects videos into three distinct views: one spatial and two temporal. This approach is designed to mitigate the computational costs associated with some of the existing methods, while still maintaining high detection accuracy.

### 3.3 Methodology

In the following subsections, we elaborate on our approach, rooted in both novel hand-crafted techniques and adapted deep learning strategies. The former serve as the primary contributions of this study, tailored specifically for video manipulation detection. The latter, initially developed in our previous work [45], tackle challenges unique to this research domain. Our methods operate effectively in both uni-modal and multi-modal settings.

#### 3.3.1 Pre-processing: lips part selection

The primary objective of this research endeavor is to advance the field of manipulation detection in videos, with a focus on several modalities including audio, visual, and audio-lips correlation. Previous studies have indicated that the lower part of the face—specifically the lip and mouth area—demonstrates reduced performance in manipulation detection when compared to other facial features like the eyes. This is particularly obvious when using deep neural networks such as XceptionNet for the classification of manipulated content [64, 65].

To address this challenge, our methodology involves the precise isolation of the lip

region from the talking face within the video frame. We employ an existing algorithm designed for lip detection and tailor it to the datasets being studied. For this purpose, we utilize Google’s Mediapipe tool [73], a comprehensive framework known for its proficiency and speed in identifying facial landmarks. By leveraging the capabilities of Mediapipe, we isolate key landmarks corresponding to the mouth and lips for further detailed analysis.

The ultimate goal of focusing on this specific region is to capture subtleties often missed by manipulation-generating algorithms, such as the intricacies of teeth [74]. By concentrating on these nuanced areas, our research aims to both improve existing methods of manipulation detection in videos and provide insights into the limitations and potential enhancements in the modeling of complex facial regions.

### 3.3.2 Rationale for Hand-crafted Techniques

#### 3.3.2.1 Proposed Pipeline

The global view of the proposed method is depicted in Figure 3.1 and will be detailed in this paragraph. After some standard pre-processing steps, we apply a dual-phase method for deepfake detection. Initially, we extract features separately from the video’s facial imagery and the corresponding audio. We then proceed to independently classify the authenticity of both video and audio. Finally, we bring together these independent classifications at a decision level, harnessing the temporal information inherent in both modalities.

#### 3.3.2.2 Hand-crafted Methodology

We detail our method, which is based on new hand-crafted techniques for extracting features from both audio and video. This is shown in the proposed method module of our pipeline in Figure 3.1.

**Audio Feature Extraction** Our approach aims to refrain from making assumptions regarding the specific frequency domain impacted by deepfake alterations in the audio signal. As a result, we have chosen to employ a multi-scale analysis for the development of our audio features. The literature presents various multi-scale decomposition methodologies, such as Mel Frequency Cepstrum Coefficients (MFCC) and Mel-Spectrograms, along with the Wavelet Scattering Transform (WST). For the purposes of this study, we have

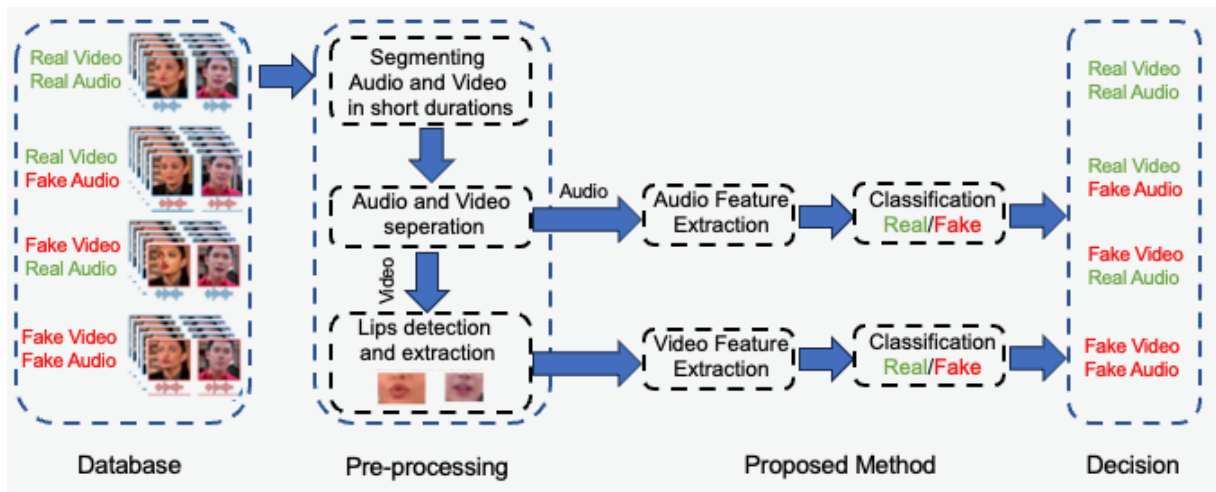


Figure 3.1 – Proposed pipeline for short utterances manipulations detection in videos.

selectively employed the WST. The selection was based on WST’s proven efficacy in capturing essential frequency characteristics of audio signals, which are crucial for identifying the subtle alterations introduced by deepfake techniques.

Consider an audio signal  $x \in \mathbb{R}^{t_x}$  sampled at  $16kHz$  and maximally normalized, where  $t_x \in \mathbb{N}^*$ . We partition this signal into its positive  $x_p \in \mathbb{R}^{t_x}$  and negative  $x_n \in \mathbb{R}^{t_x}$  components as follows:

$$\begin{cases} x_p = ReLU(x) \\ x_n = ReLU(-x) \\ x = x_p - x_n \end{cases} . \quad (3.1)$$

Let us introduce  $\Psi$ , denoting the Wavelet Scattering Transform (WST) designed for one-dimensional signals as presented in existing works [32, 33]. The fundamental idea behind WST is the iterative application of the wavelet transform [75] coupled with a modulus operation serving as a non-linear function and subsequently averaging the result through a Gaussian filter. This transformation technique is subject to multiple hyper-parameters, including the invariance scale (window length), the transform depth, and the quality factors which determine the number of wavelets per octave. This WST is suited for our purpose as it gives a frequency characterization of an audio.



A preliminary assessment of disparities between an authentic audio and its cloned using the audio deepfake generation method described in [66]. The two scattergrams (c-d) in the figure 3.2 illustrate variations in the WST between the upper and lower samples of the two audio sources, while the scattergrams (a-b) of the two signals without decomposition have small dissimilarities. The two plots in (c-d) shows that the fake audio is characterized by predominantly negative values with a positive value at  $0Hz$ , whereas the authentic one primarily consists of positive values and exhibits null values at  $0Hz$ .

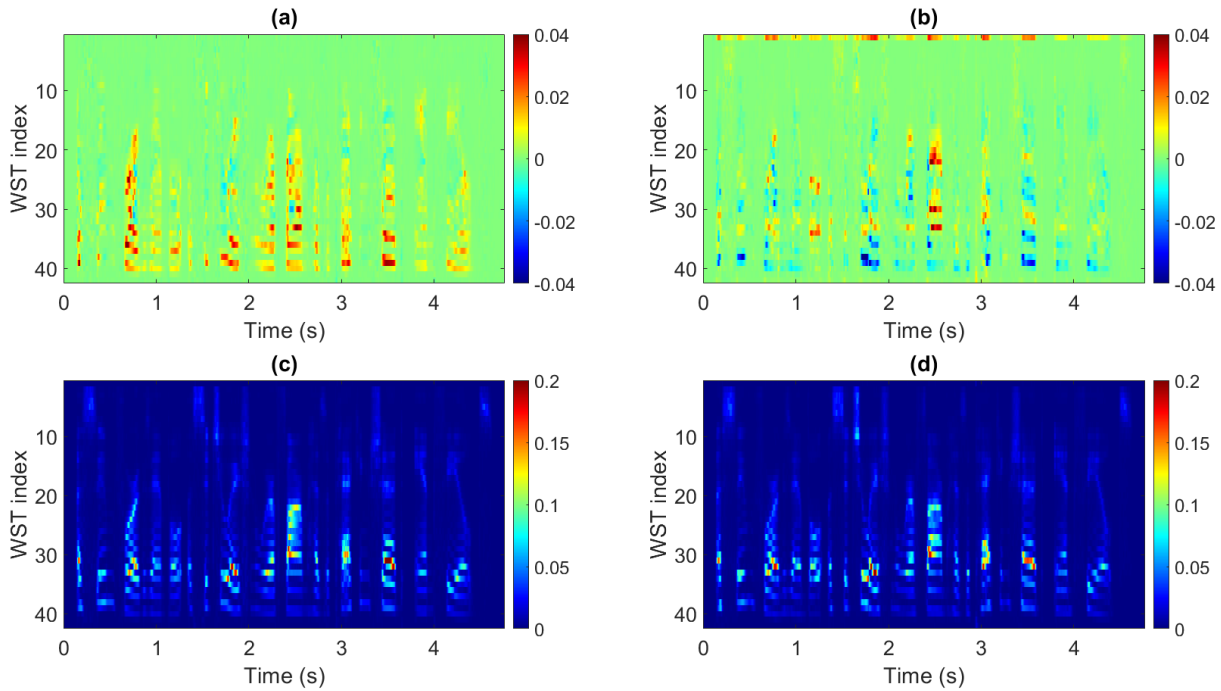


Figure 3.2 – (a) The difference between the WST of positive and negative sample of Real audio. (b) The difference between the WST of positive and negative sample of the fake audio represented in (a). (c) The WST of the same real signal. (d) The WST of the fake signal. Only the zero-th and the first order WST are presented and the fake signal was generated using the method described in [66] (i.e: the WST index refers to the mel-scaled frequencies between 0 Hz and 8 kHz).

In line with the existing research on audio processing [32], we have set the window length at  $64ms$  and chosen four layers, which are adequate for better capturing invariant features [44]. With this setup and a sampling frequency of  $16kHz$ , we derive 153 WST coefficients. Subsequently, we compute the WST  $\Psi$  of our signal  $x$  and both negative  $x_n$  and positive  $x_p$  components. This results in the output matrices  $\mathbf{X}_n, \mathbf{X}_p, \mathbf{X} \in \mathbb{R}^{C \times T_x}$ ,

where ( $C \in \mathbb{N}^*$ ) denotes the number of WST coefficients, ranging from zero order up to the fourth order, here  $C = 153$ .

The core concept of our method hinges on using transformed matrices to calculate the lower bounds of the Pearson correlation coefficient, focusing on the interactions between  $\mathbf{X}_p$ ,  $\mathbf{X}_n$ ,  $\mathbf{X}$ , and  $\mathbf{X}_p - \mathbf{X}_n$ . This approach is integral to our analysis, as it leverages only the inherent elements of the audio signal, thus eliminating the need for external benchmarks. Mathematically, the framework of our proposed method can be described as follows:

$$\mathbf{S}(c) = \min \left( \begin{aligned} &\rho(\mathbf{X}_n(c), \mathbf{X}_p(c) - \mathbf{X}_n(c)), \\ &\rho(\mathbf{X}(c), \mathbf{X}_p(c) - \mathbf{X}_n(c)), \\ &\rho(\mathbf{X}_p(c), \mathbf{X}_p(c) - \mathbf{X}_n(c)) \end{aligned} \right), \quad (3.2)$$

where  $c = 1, \dots, C$  is the channel index of the WST coefficients,  $\mathbf{S} = (\mathbf{S}(1), \dots, \mathbf{S}(C)) \in \mathbb{R}^C$  is the features vector to detect a fake audio and  $\rho$  is the Pearson correlation coefficient across temporal axis.

We employ the Pearson Correlation Coefficient to elucidate the interrelationships among the distinct components of the audio signal. Analyzing the correlation patterns between the positive and negative aspects of the WST coefficients allows us to detect inconsistencies characteristic of deepfake manipulations. Such irregularities are indicative of alterations, given that genuine audio signals typically demonstrate a stable and consistent correlation pattern, which is often disrupted in manipulated audio.

**Video Feature Extraction** Following the exposition of our audio deepfake detection methodology, we now introduce the algorithm we designed to discern the authenticity of visual sequences. This method hinges on both the spatial and temporal attributes of a video, focusing primarily on detecting any anomalies or inconsistencies in the motion or appearance of a speaking subject in a video with a zero head pose. Given that deepfake generation algorithms exhibit difficulties in accurately replicating the high-frequency characteristics inside the mouth area such as the teethes [76, 77], our method emphasizes the high-frequency components of the video signal.

Let  $V \in \mathbb{R}^{t_v \times 3 \times N_x \times N_y}$  be a video sequence depicting the lip movements of a subject speaking without head pose. The video frames are converted to gray-scale for analysis. For each temporal instance  $t = 1, \dots, t_v$ , we take the fourth-order spatial derivative of the frame

$V(t) \in \mathbb{R}^{N_x \times N_y}$  with respect to both  $x$  and  $y$  axes to yield  $\frac{\partial^4 V}{\partial x^2 \partial y^2} \in \mathbb{R}^{t_v \times 3 \times N_x \times N_y}$ . This operation effectively shifts the energy of the signal toward the high-frequency domain.

Subsequently, we apply the two-dimensional discrete cosine transform (DCT2), denoted as  $\Phi$ , to  $\frac{\partial^4 V}{\partial x^2 \partial y^2}$ , resulting in the frequency representation of the rate of intensity change across frames. Drawing inspiration from existing work [78], which leverages temporal variations in Pearson correlation coefficient to identify talking regions, we compute this coefficient for successive frames. Mathematically, the process can be described as follows

$$\left\{ \begin{array}{l} \Phi\left(\frac{\partial^4 V}{\partial x^2 \partial y^2}\right)(t) = (\nu_{i,j}(t))_{1 \leq i \leq N_x, 1 \leq j \leq N_y} \in \mathbb{R}^{N_x \times N_y} \\ \nu_{i,j}(t) = \sum_{p=0}^{N_x-1} \sum_{q=0}^{N_y-1} \frac{\partial^4 V}{\partial x^2 \partial y^2}(p, q, t) \cdot \cos\left(\frac{\pi}{N_x} \left(p + \frac{1}{2}\right) i\right) \cdot \cos\left(\frac{\pi}{N_y} \left(q + \frac{1}{2}\right) j\right), \quad (3.3) \\ \rho(t) = \text{pearson}\left(\Phi\left(\frac{\partial^4 V}{\partial x^2 \partial y^2}\right)(t), \Phi\left(\frac{\partial^4 V}{\partial x^2 \partial y^2}\right)(t-1)\right) \in \mathbb{R}, t = 2, \dots, t_v \end{array} \right.$$

where  $\Phi\left(\frac{\partial^4 V}{\partial x^2 \partial y^2}\right)(t)$  represents the high-frequency components obtained from the DCT2 transform  $\Phi$  of the frame  $V(t)$ , and  $\rho(t)$  denotes the Pearson correlation coefficient between two successive frames  $V(t)$  and  $V(t+1)$ . To perform an analysis over a specific duration or the entire length of the video, we construct a scatter plot using the temporal mean  $\mu$  and the standard deviation  $\sigma$  of the  $\rho$  values.

### 3.3.3 Deep learning Methodology

In scenarios with abundant data, deep learning, especially CNNs, proves invaluable due to its capability to autonomously discern and extract pivotal features, a feat often surpassing the performance of hand-crafted methods. Building on [45] developed deep learning model for biometric identification tasks, we have incorporated specific enhancements to address the unique challenges presented by deepfake detection in audio-visual data. These enhancements, detailed in the subsequent sections, refine the model's architecture and functionality, where we have changed the number of classes to 2 (Real Audio - Real Video and Fake Audio - Fake Video) or 4 (Real Audio - Real Video, Fake Audio - Fake Video, Real Audio - Fake Video and Fake Audio - Real Video), ensuring optimized performance for this application.

In our approach to deepfake detection, we opt for a late fusion technique inspired

by the prior research in biometric tasks [45]. This architecture is presented in Fig. 3.4. The decision to employ late fusion is motivated by several factors. Firstly, it offers a more straightforward implementation compared to alternatives like early or hybrid fusion, effectively balancing information derived from both audio and visual modalities.

Secondly, the networks used for feature extraction in each modality are already pre-trained on expansive datasets. For audio, we use x-vectors pre-trained on the VoxCeleb dataset [48], and for the visual aspect, we utilize ResNet18 pre-trained on ImageNet [79]. This allows us to concentrate solely on fine-tuning the fusion and classification layers, streamlining the overall training process.

Additionally, by dividing the video into multiple views,  $XY$  representing the spatial view,  $TY$  and  $XT$  incorporating the spatio-temporal respectively views across the y-axis and x- axis, we enhance our detection capabilities. This multi-view framework permits the identification of specific features such as jaw location thanks to the  $XY$  view, motion jitters (a demonstration of this effect can be observed in Fig. 3.3)  $TY$  view, and common artifacts that deepfake generators often struggle to simulate convincingly.

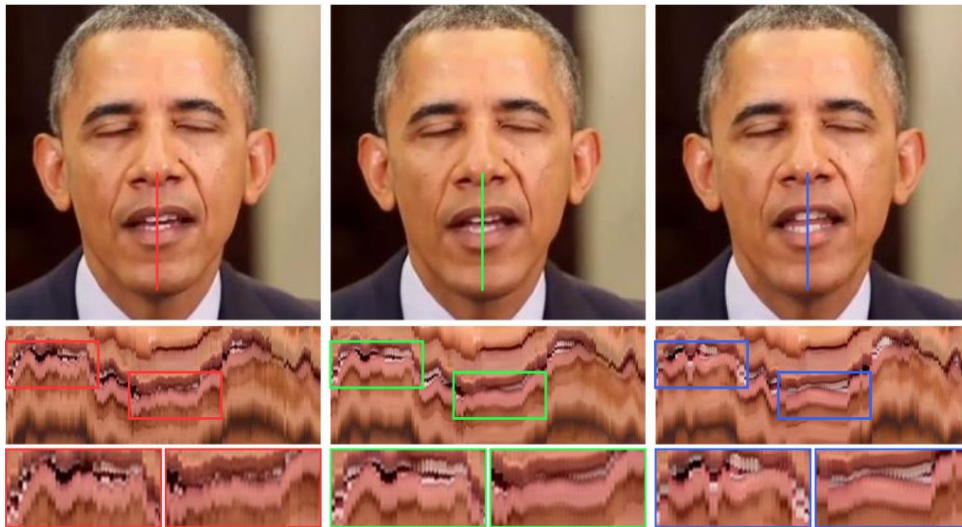


Figure 3.3 – Motion jitters problem illustration from the article [62]. Left and middle are two deepfake generation methods, and the right image is taken from a real video. For each video, the vertical cuts (the vertical red/green/blue line) are made in each frame along time, and then show their concatenation at the bottom.

Lastly, the computational efficiency of a multi-view architecture makes it a more practical choice for feature extraction than using 3D CNNs, particularly in real-world applications.

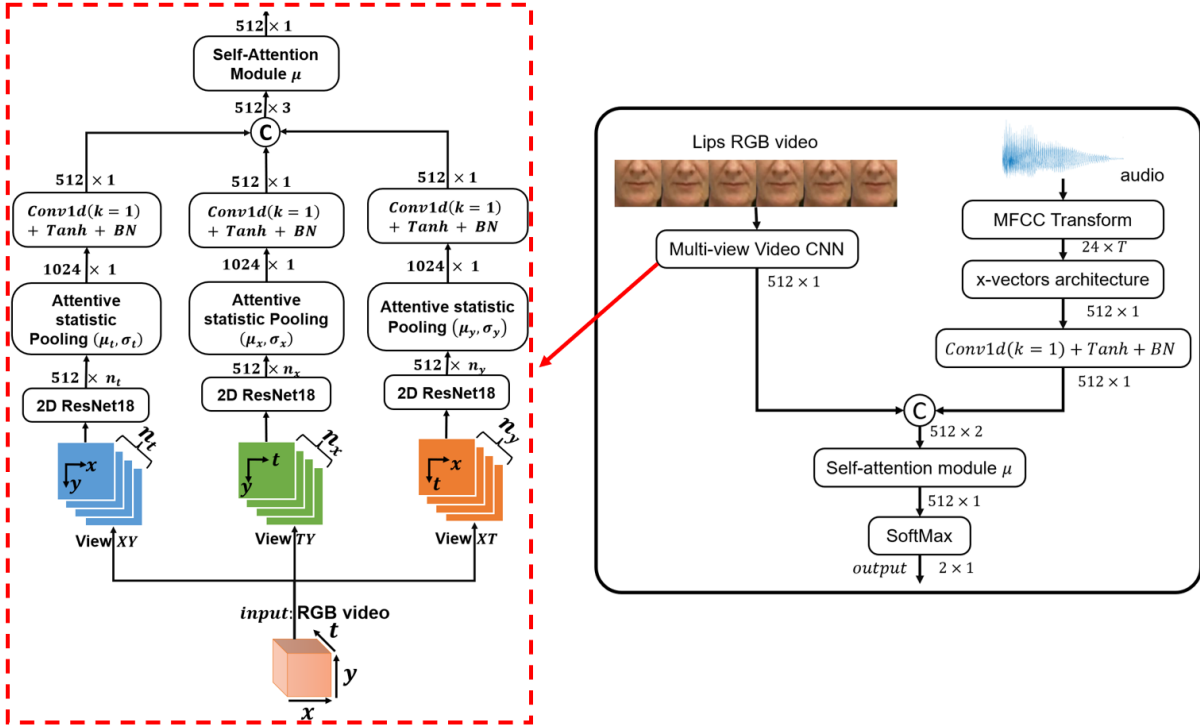


Figure 3.4 – Multi-view CNN late fusion architecture for audio-lips correlation [45]. For multimodal deepfake detection, the number of classes are set to 2 or 4, in addition, we only tune the fusion layers (all layers except 2D ResNet-18 and x-vectors).

### 3.4 Experimental Strategy and Materials

In this section, we detail the experimental procedures used to validate our methods against existing state-of-the-art (SOTA) solutions [45, 72, 59, 60] for both audio and visual deepfake detection. Our validation process includes experiments with two reputable datasets which would be explained in the following.

#### 3.4.1 Datasets

To evaluate our proposed methods, we introduce benchmark and reference datasets that are widely accepted and commonly employed in the field of deepfake detection to assess the effectiveness of our proposed methods. Note that this may not be an exhaustive list.

### 3.4.1.1 FakeAVCeleb

The FakeAVCeleb dataset serves as a comprehensive and developed resource for audio-visual deepfake detection [80]. Originating from the frequently cited VoxCeleb2 multi-modal corpus [48], FakeAVCeleb stands out for its frame rate of 25 fps and an average video duration of approximately 7.8 seconds. Employing an array of state-of-the-art deepfake generation methods such as lip-syncing and face-swapping technologies [81, 82, 83, 84], the authors have fabricated videos that pose a realistic and formidable challenge in discerning their authenticity.

To bolster its applicability, the dataset has been curated to offer ethnic and gender diversity, thus paving the way for equitable and representative evaluation. Structurally, it comprises four distinct categories of audio-visual data:

- A collection of 500 videos with real-audios and real-videos ( $R_v R_a$ ),
- Another set of 500 videos where the audio has been cloned or generated from a speech text while the visual content remains authentic ( $R_v F_a$ ),
- A larger set of 9,700 videos featuring real audio coupled with manipulated visual content ( $F_v R_a$ ),
- Finally, a comprehensive set of 10,800 videos in which both the visual and audio components are synthetic ( $F_v F_a$ ).

The FakeAVCeleb dataset introduces three distinct classes of manipulated content, alongside a dedicated category for wholly authentic videos. For testing our method, videos from this dataset have been chunked into segments ranging from 200 ms, 600 ms to 1 s in duration. Samples from this dataset are depicted in the Fig. 3.5 from the original article [80].

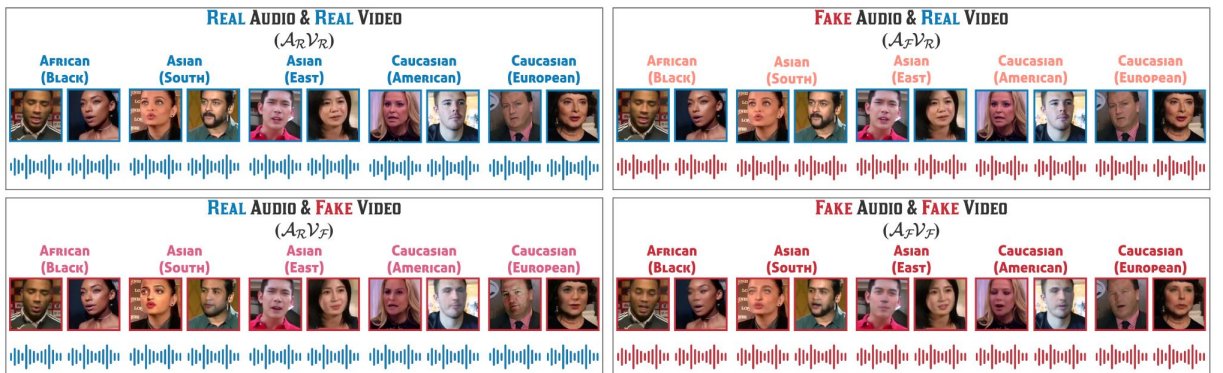


Figure 3.5 – Samples from the 4 classes of FakeAVCeleb dataset [80].

### 3.4.1.2 DeepfakeTIMIT

The DeepfakeTIMIT dataset emerges as an essential benchmark in the realm of deepfake detection, featuring an average video length of approximately 4.25 seconds. Originating from the VidTIMIT database [85], it employs GAN-based face-swapping techniques to generate manipulated content [86, 87]. This corpus is bifurcated into two main categories: the first contains 320 clips with manipulated visuals yet authentic audio ('Fake Video - Real Audio'), while the second consists of 430 clips preserving both the original video and audio ('Real Video - Real Audio'). Two samples representing fake and real frames are depicted in Fig. 3.6.

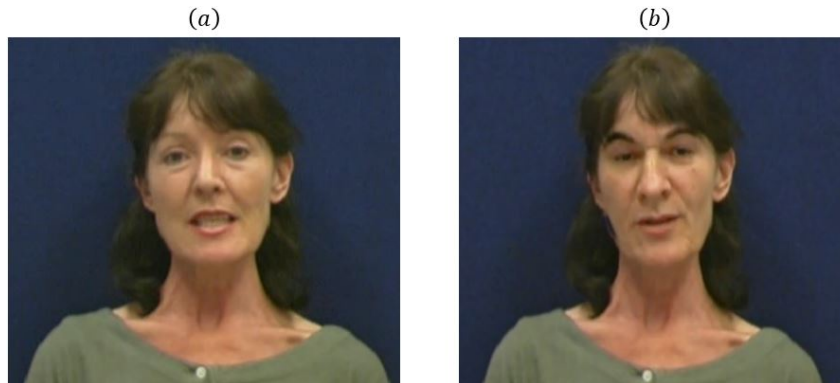


Figure 3.6 – Two frames samples from DeepfakeTIMIT dataset ((a) Real, (b) Fake).

Our investigation is primarily aimed at the low-quality segment of the DeepfakeTIMIT collection. This specific tier provides invaluable insights into the robustness of deepfake detection methods when operating under suboptimal conditions. It thereby furnishes a more nuanced understanding of algorithmic performance constraints, especially considering that diminished visual and auditory quality exacerbate the challenges of distinguishing authentic content from fabricated instances. For the purposes of our study, we implemented a preprocessing step that involved the removal of silent segments from the video clips in the dataset. This ensured that the subsequent segmentation into durations ranging from 200 ms and 600 ms to 1 s resulted in utterances containing at least a word, thereby guaranteeing meaningful audio-visual data for analysis.



## 3.4.2 Experimental Strategy

### 3.4.2.1 Hand-crafted proposed methods

**The first experiment (hand-crafted audio deepfake detection)** assesses the capability of our model in detecting deepfakes within short utterances. For this purpose, we segmented the datasets into frames of 200 to 600 milliseconds, with a 50% overlap between consecutive frames. This approach ensures that the dataset is both robust and continuous, enabling a thorough analysis of the model’s performance.

The data was split into training, validation, and test sets using a 60-20-20% ratio, which is a standard practice to ensure a balanced evaluation of the model. We utilized a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel to classify the features extracted as described in Equation (3.2). The RBF kernel was chosen for its effectiveness in handling non-linear patterns within the data. During the training and validation phases, the regularization coefficient was carefully tuned to 0.01, optimizing the model’s performance by controlling the balance between model complexity and classification accuracy.

**The second experiment (hand-crafted visual deepfake detection)** evaluates the effectiveness of our model detailed in Eq. 3.3, in detecting deepfakes within the visual channel of videos, both in short utterances and across entire video frames. Similar to the approach used for audio signals, we segmented the video datasets into short frames ranging from 200 to 600 milliseconds. This segmentation allows us to focus on brief visual segments, which are particularly challenging in deepfake detection.

We applied the same cross-validation strategy as in the first experiment to ensure consistent and rigorous evaluation of the model’s performance in the visual domain. By maintaining this uniform approach, we were able to accurately assess the model’s ability to detect visual manipulations in both short and extended video sequences.

### 3.4.2.2 Deep Learning proposed method

**The third experiment (Comparison to the SOTA):** we evaluate [45] network (cf. Fig. 3.4), we exclusively consider audio RGB lips videos from the classes  $R_v R_a$  (Real Video - Real Audio) and  $F_v F_a$  (Fake Video - Fake Audio) from FakeAvCeleb dataset [80], labeled respectively as real and fake audiovisual. These sequences are split to frame lengths ranging from 200 ms, 600 ms and 1 s with an overlap of 50%. We have ensured that there was an equal partition between fake and real labels.



**The fourth experiment (Multi class detection)** our emphasis is on evaluating our network’s performance on short segments containing four distinct classes. The goal is to determine whether the network’s performance remains consistent when at least one real modality is present, in other words, there are four main labels:

- Fake Video - Fake Audio ( $F_vF_a$ ),
- Fake Video - Real Audio ( $F_vR_a$ ),
- Real Video - Fake Audio ( $R_vF_a$ ),
- Real Video - Real Audio ( $R_vR_a$ ).

For this purpose, we selected a dataset comprising equally distribution among four subsets:  $F_vF_a$ ,  $R_vF_a$ ,  $F_vR_a$  and  $R_vR_a$ , token from FakeAVCeleb [80]. We then segmented the lip portions of videos from these categories into frame lengths of 0.2 s, 0.6 s, and 1 s, with a 50% overlap between consecutive frames. The dataset was further divided into training and testing sets following an 80 – 20% split, and the classification layer was set to 4 classes.

**The fifth experiment (Investigation on the role of each view and modality)** to delve deeper into the impact of each view—namely,  $XY$ ,  $XT$ ,  $TY$ —and their fusion (cf. Fig. 3.4), we conducted a dedicated experiment. For this evaluation, we considered a balanced dataset consisting of 500 real videos ( $R_vR_a$ ) and 500 fake videos ( $F_vF_a$ ), segmented into 1-second frame lengths with overlapping intervals. The data was partitioned into training and testing sets at an 80 – 20% ratio, and the experiment was designed with two distinct classes.

## 3.5 Experimental Results

### 3.5.1 Hand-crafted methods

**Experiment 1:** Our experimental framework, as detailed in Equation (3.2), is designed to evaluate the effectiveness of our hand-crafted method for detecting deepfake audio. We rigorously tested our approach using the FakeAVCeleb dataset [80], which predominantly consists of synthesized audio samples that pose significant challenges for detection algorithms. The results, shown in Table 3.1, indicate that our method not only surpasses SOTA deep learning techniques in accurately identifying deepfakes within short audio utterances but also maintains a high level of performance when applied to the full duration of the audio as indicated in Table 3.2.

Model	Accuracy (%)		
	Whole audio	600 ms	200 ms
X-vectors + SVM [72, 61, 19]	<b>99.75±0.04</b>	<b>96.13±0.12</b>	85.38±0.18
ECAPA-TDNN + SVM [72]	99.65±0.05	88.81±0.07	75.01±0.05
Our Method + SVM	99.12±0.74	96.08±0.92	<b>85.71±1.25</b>

Table 3.1 – Performance of our method in Eq. (3.2) on short utterances and full audio modality in FakeAVCeleb, compared to SOTA methods (the standard deviation values are calculated from a 5-fold cross-validation).

Model	Accuracy
MFCC + XceptionNet [59]	76.6%
Mel-Spectrograms + DST-Net [18]	97.51%
MFCC + DST-Net [18]	88.5%
X-vectors + SVM [72, 61, 19]	<b>99.98%</b>
ECAPA-TDNN + SVM [72]	99.97%
Our hand-crafted method	99.83%

Table 3.2 – Performance of our proposed method on the whole audio length modality of FakeAVCeleb compared to the SOTA methods (Real: 500 videos from  $R_vR_a$ , Fake: 500 videos from  $F_vF_a$ ). Bold entries indicate the best performance.

**Experiment 2:** We present the empirical evaluation of our visual deepfake detection approach, formulated as per Eq. (3.3). The focus of this investigation is restricted to the lip region of the subject, contingent upon the maintenance of a zero-degree head pose throughout the recording session. Consequently, this evaluation is exclusively conducted on the DeepfakeTIMIT dataset.

Our aim is to track the temporal behavior of the Pearson coefficient defined in Eq. (3.3), then we decide the authenticity of the visual sequence. We considered 20 fake videos from DeepfakeTIMIT and 20 real videos from VidTIMIT. We then split every video into chunks of  $600ms$  and an overlap of 50%, we obtain 279 fake and 279 real videos. To imprint the temporal dynamics of our parameter in Eq. (3.3), we choose the mean and the standard deviation.

Figure 3.7 provides a visual demonstration of our method’s effectiveness. The figure presents a scatter plot that includes various measurements, such as the average and variation (standard deviation) of the Pearson coefficient. This visual arrangement effectively separates real videos from fake ones. The top-left part of the plot illustrates the application of the DCT coefficient to unaltered video frames. In contrast, the top-right part

shows the results of applying the first spatial derivative to the video frames. Notably, the bottom-left section of the plot underscores our method’s strong capability in distinguishing real from fake videos. Based on this last observation, the authenticity of the video can be done by setting a threshold on the inverse coefficient of variation  $C_v^{-1}$  plotted at the bottom right of the Figure 3.7. For various values of the threshold  $C_v^{-1}$  ranging from  $-1$  to  $10$ , we plot the receiver operating characteristic (ROC) curve (not shown) and the optimal threshold to distinguish genuine and fake videos giving a maximum detection accuracy  $98.39\%$  was found at  $1$ .

Setting the threshold at  $1$  allowed us to achieve a remarkable detection accuracy of  $99.87\%$  across the entire VidTIMIT and DeepfakeTIMIT video datasets. This result strongly validates the effectiveness of our method in reliably distinguishing between authentic and manipulated content. To further highlight the capability of our approach, we conducted a performance comparison with SOTA methods, as detailed in Table 3.3 and Table 3.4.

Model	Accuracy
XceptionNet (Image level detection) [80]	65.98%
I3D (Sequence level detection) [60]	96.38%
Our hand-crafted method	<b>99.87%</b>

Table 3.3 – Performance of the SOTA and our proposed method on whole length visual sequence from low-quality videos of DeepfakeTIMIT. Bold entries indicate the best performance.

Type	Model	Accuracy (%)		
		Whole video	600 ms	200 ms
Deep Learning	XceptionNet [59]	65.98	68.19	61.98
	Multi-view video CNN [45]	98.43	<b>99.39</b>	<b>99.88</b>
Shallow Learning	Our hand-crafted method	<b>99.87</b>	99.12	97.36

Table 3.4 – Comparison of SOTA performance with our proposed approach using visual sequences from low-quality DeepfakeTIMIT videos over various time segments.

### 3.5.2 Audio - Visual late fusion deep learning method

**Experiment 3:** The system’s visual input consists of the lower facial region, selected for its computational efficiency. The 2D decomposition allows for an economical extraction of spatiotemporal information from the video footage. Despite this optimization, our

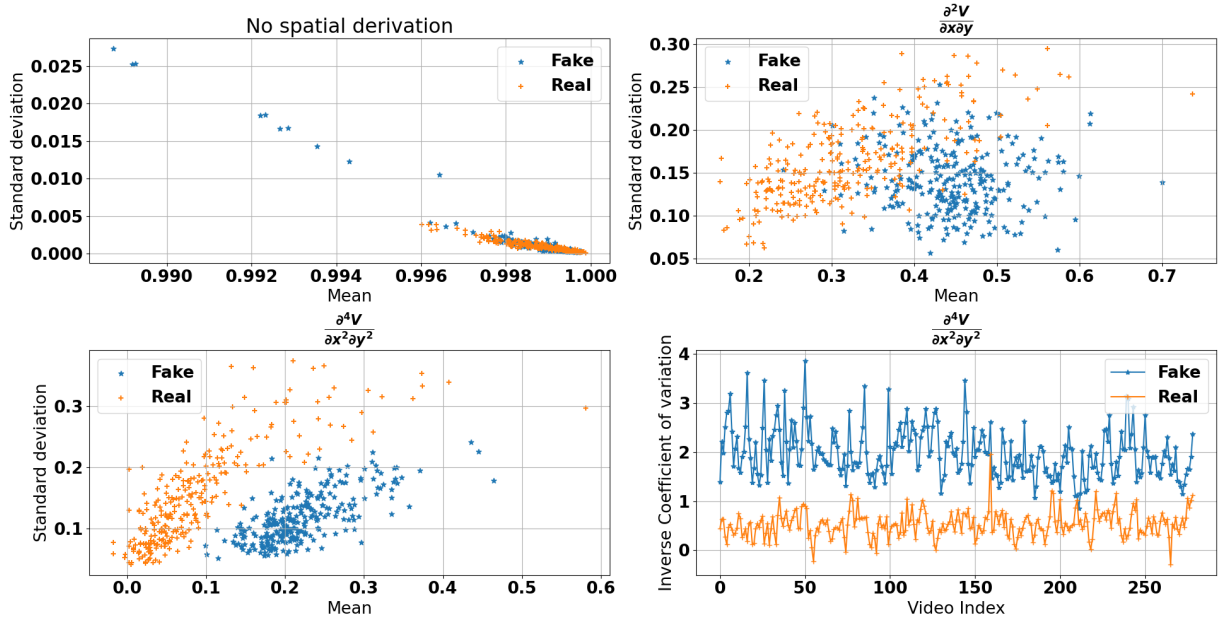


Figure 3.7 – Analysis of low-quality 20 fake and genuine videos split into 600ms with an overlap of 50% using our method. Top-left: Mean and standard deviation without spatial derivation. Top-right: After second-level spatial derivation. Bottom-left: After fourth-level spatial derivation. Bottom-right: Temporal inverse coefficient of variation of videos from DeepfakeTIMIT and VidTIMIT.

approach yields superior detection accuracy relative to current state-of-the-art models, as indicated in Table 3.5.

Model	Accuracy (%)		
	200 ms	600 ms	1 s
XceptionNet (Soft-Voting) [60]	77.33	77.69	73.34
Multi-View CNN (Ours) [45]	93.19	97.68	98.55

Table 3.5 – Comparison of SOTA performance with our proposed approach [45] on various time segments from FakeAVCeleb on balanced 1,000 samples (500  $F_v F_a$ , 500  $R_v R_a$ ) subset, the train-test split was set to 80% – 20%.

**Experiment 4:** The results displayed in Table 3.6 confirm a huge decrease in detection accuracy compared to the scenario with only two classes, as outlined in Table 3.5. These observations indicate that the inclusion of a real modality tends to strongly influence our network’s determination of a video’s authenticity. To deeply understand the dominance of one modality over another, we carry out an experiment in the upcoming

subsection to pinpoint the modality where our network demonstrates lower detection accuracy. Additionally, we explore the contribution of each view in improving the overall performance.

Table 3.6 – Detection accuracy of Multi-view CNN on short videos from FakeAVCeleb (train-test split 80% – 20% on 2,000 videos of (500  $F_vF_a$ , 500  $F_vR_a$ , 500  $R_vF_a$  and 500  $R_vR_a$ ), the classes number for classification is set to 4).

Window length	Accuracy (train/test)
200 ms	85.95%/74.04%
600 ms	82.89%/82.00%
1 s	88.68%/85.96%

**Experiment 5:** The findings, presented in Table 3.7, illuminate the crucial roles played by individual views and modalities in the system’s performance. Notably, the temporal view ( $TY$ ), capable of accounting for issues like motion jitters, outshines the spatial ( $XY$ ) and the spatiotemporal ( $XT$ ) views. The relatively poorer performance of the spatial view equipped with a self-attention model can be explained by the focus on the lower part of the face—a known challenging aspect for many state-of-the-art deepfake detection methods, as cited in [65, 64].

When fusing all visual views, we observe a marked increase in detection accuracy, thereby showcasing the model’s prowess in resolving the ambiguity or confusion that could arise from individual views. Moreover, our results indicate that the audio modality holds a distinct edge in detection accuracy, contributing to an overall performance lift of 7.86% when integrated with the visual modality.

Table 3.7 – Detection accuracy of Multi-view CNN on short videos from FakeAVCeleb (train and test on 1,000 videos of (500  $F_vF_a$ , 500  $R_vR_a$ ) cutted into 1 s frame length with an overlap 50%, the classes number for classification is set to 2).

Modality	View	Accuracy (train/test)	Precision (train/test)	Recall (train/test)
Visual	$XY$	85.76%/80.22%	88.12%/82.17%	89.64%/79.96%
Visual	$XT$	85.47%/83.04%	88.45%/87.68%	87.22%/84.93%
Visual	$TY$	88.15%/88.21%	91.31%/88.93%	88.72%/92.76%
Visual	$XYT$	94.88%/90.05%	96.21%/93.52%	95.23%/90.41%
Audio	–	100%/100%	100%/100%	100%/100%
Audio + Visual	$XYT$ + Audio	98.77%/97.91%	99.06%/97.96%	98.90%/98.72%

## 3.6 Discussion

### 3.6.1 Hand-crafted methods

The methodology we have formulated for the detection of fake audio signals exhibits numerous distinct advantages, with its performance and generalizability being particularly noteworthy.

Unlike some methods that depend on a predefined set of reference speakers, as discussed in [72], our approach distinguishes itself by removing the need for these comparisons and trainable parameters. This distinction is clearly demonstrated in Figure 3.8, which shows the Pearson coefficients across WST channels for both authentic and fake audios produced using Text-To-Speech (TTS) techniques [66].

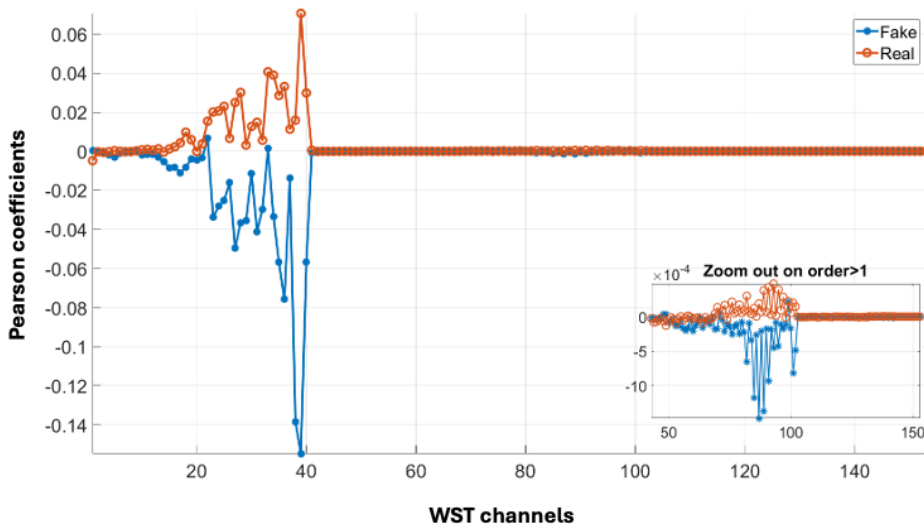


Figure 3.8 – Pearson coefficient value (y-axis) described in Eq. 3.2 per each channel (x-axis) (orange circles refers to the real audio and the blue star stands to the fake audio).

Moreover, our technique demonstrates a robust capacity to handle short audio utterances, maintaining satisfactory performance which incrementally improves with the length of the audio sample. This trend is particularly advantageous for scenarios commonly faced in the real world, where audio clips are often brief.

On the visual modality, the robustness of our visual-level deepfake detection method is underscored by its efficacy under a range of challenging conditions. Specifically, the technique exhibits high performance even when subjected to low-quality videos. What sets our method apart is its unique focus on the lip region for detection—a region notoriously difficult for traditional deepfake detection methods to analyze. This focus doesn't merely serve to fill a gap in existing methodologies; it provides our system with a marked advantage over SOTA deep learning-based approaches. Moreover, this is achieved with very limited number of hyperparameters, which significantly reduces the computational overhead and simplifies the implementation.

The crux of our method is its strategic utilization of high-frequency spatial energy patterns. By pushing spatial energy towards these higher frequencies, our system is able to significantly amplify the contrast between real and fake visual sequences, a fact that is empirically supported by the results illustrated in Figure 3.7. This approach not only serves to enhance detection capabilities but also fortifies our system's adaptability. This adaptability is further substantiated by the system's performance metrics under scenarios involving short utterances, ranging in length from  $200ms$  to  $600ms$ . Even under these non-ideal conditions, the system was able to maintain a reasonable performance level.

However, it's important to note that the scope of the current study did not extend to evaluating the technique's robustness against videos with varying luminosity or noise levels. Given the demonstrated performance of the method, this remains a crucial avenue for future research, to comprehensively assess the system's applicability under a wider array of real-world conditions. As long as the luminosity does not change temporally or spatially in temporal or spatial frequency ranges which are discriminant between real and fake this should not have impact on the performance of our method.

### 3.6.2 Audio-Visual late fusion deep learning method

Navigating the intricate landscape of fake video detection necessitates a methodological framework that is both efficient and nuanced. In tuning our pre-existing model, cited as [45], we have not only successfully adapted it for fake video detection but also advanced our understanding of how different modalities contribute to the detection process. One of the primary strengths of our architecture lies in its utilization of pre-trained networks: ResNet-18 for ImageNet and x-vectors for VoxCeleb2. These well-established, data-rich training sources confer upon our model a robust initial feature set. Moreover, we adopt a computationally economical approach by using a  $2D$  scheme for visual sequences and a  $1D$

scheme for audio, thereby circumventing the need for more resource-intensive networks.

Our research goes beyond mere detection to dissect the relative contributions of each sequence view:  $XY$ ,  $XT$ , and  $TY$ . The temporal motion jitters, belonging to  $TY$  view, emerges as the most accurate in detecting fake videos. This likely capitalizes on the inherent difficulties that deepfake algorithms have in accurately reproducing the temporal dynamics of human behavior. Conversely, the spatial view  $XY$  underperforms, which is consistent with existing literature [65, 64] indicating that the lower facial region presents substantial challenges for deepfake detection systems.

The fusion of these three views adds an additional layer of complexity, further refining our model’s detection capabilities. Such a fusion approach effectively exploits both spatial and temporal information, without the need for resource-intensive  $3D$  models. Importantly, the incorporation of audio via x-vectors lends a significant boost to the model’s performance. This may be attributed to the transfer learning advantages offered by Vox-Celeb2, or it could point to a more fundamental characteristic of deepfake generation algorithms—that they are currently more proficient in visual manipulation than in audio.

Despite these promising outcomes, the architecture’s performance is not without limitations. Most notably, its efficacy diminishes when applied to short utterances. This finding is significant and indicates a key area for future research: optimizing the model to maintain high detection rates irrespective of video length.

## 3.7 Conclusion and perspectives

In this study, we have introduced a full-pipeline approach based on hand-craft method and tuned our previous architecture [45] to detect short audiovisual deepfake content.

On the shallow learning side, we introduced a full-pipeline approach to detect fake audio and video content, leveraging hand-crafted features for audio and distinctive visual cues, particularly in the lip region. Our method demonstrates high interpretability and computational efficiency, achieving robust performance on the FakeAVCeleb and DeepfakeTIMIT datasets. This unified strategy underscores the synergy between auditory and visual elements, reflecting a comprehensive stance against the rising tide of deepfake technologies. The robustness of our approach is evident as it maintains performance metrics across various conditions without requiring a reference dataset or a complex train-test split, marking a significant advancement over existing deep learning methods.

On the deep learning side, we expanded upon the prior work in late fusion biometrics



identification [45] to address the detection of deepfake videos using two distinct modalities. The model has demonstrated superior performance compared to the SOTA on the FakeAVCeleb dataset. Additionally, we have delved into the influence of three views in video decomposition and the role of modalities in augmenting detection accuracy. Notably, our findings highlight a substantial contribution from the audio modality in comparison to its visual counterpart.

Our future perspective aims to enhance this integration by considering an hybrid fusion that could further integrate the audiovisual features at multiple levels of extraction. Such a multimodal system could benefit from the inherent strengths of each modality, potentially leading to a more resilient detection mechanism against sophisticated deepfake manipulations. These efforts will contribute to the overarching goal of ensuring the authenticity and trustworthiness of digital media.

# UNSUPERVISED MULTIMODAL STUDENT'S ENGAGEMENT DETECTION

---

Previously, we discussed the advantages of multimodality in deepfake detection on ultra short utterances. In this chapter, we shift our focus to education, specifically investigating the merits of utilizing multimodal approaches to analyze student engagement with the aim of gauging their engagement levels during classroom sessions in foreign language classes, thereby assessing the effectiveness of teaching methods. Unlike the other chapters, here our focus will be on the whole face and the heart rate signal of the student. Section 4.1 starts with an introduction about multimodal student engagement recognition. Section 4.2 reviews existing literature in this domain, emphasizing the limitations of current approaches and datasets. Section 4.3 elaborates on our new dataset designed for didactic purposes, the experimental design, and unsupervised hand-crafted methods to detect significant moments during a class session. Section 4.5 presents the results of our preliminary explored pipeline to detect the significant moment at the local time level cross referenced with the observation notes of the didactic researcher. Section 4.5 presents a comprehensive analysis of our results, followed by Section 4.6 that concludes the chapter and discusses future research directions.

## 4.1 Introduction

The need for interdisciplinary research has never been greater in the field of foreign language education and applied linguistics [88]. While researchers broadly agree that interdisciplinary perspectives allowed the creation of “new knowledge frameworks” [89], there is no denying that “Even today, it is not so simple to transcend these disciplinary boundaries to build an interdisciplinary, collaborative, and relevant scientific approach” (p. 5). While a fair amount of interdisciplinary research has already been carried out in applied linguistics and psychology, there still is much to be done. For example, most

research on learners’ emotions in Foreign Language (FL) classes is based on self-reported data collected from learners after class through interviews, questionnaires and diaries. Gregersen et al. [90] pointed out that these approaches cannot capture the dynamic and fluctuating nature of emotions in the classroom. They argue that the field needs a new type of study based on multimodal data collection, combining physiological data with ratings of emotions while performing a task, and answering questions about the spikes and dips in a subsequent interview [91]. Recent advances in technology have made it possible to collect and analyze multiple data streams in real-time, enabling a more sophisticated triangulation of learners’ dynamic emotions, experiences of flow, neural activities and facial expressions [92, 93, 94]. While there is nothing wrong with self-reports, there is an inherent limitation because not everybody is equally capable of verbalising what they feel [95]. Individuals with low levels of emotional intelligence may give very broad indications on the valence of their emotions (good/bad) while those with higher levels of emotional intelligence detect their own emotions in more detail and nuance, and therefore provide a much more accurate picture of their various interacting emotions. One way to mitigate this source of variation, which could lower the quality of the self-reported data, is to complement it with various physiological measures. We fully agree with the view that multimodal approaches “will make it possible to model emotion in higher dimensionality, and answer fundamental questions about how biological, mental, and contextual features are related over time” [92].

The present study adopts such a cutting-edge multimodal approach combining Heart Rate (HR) monitoring, Emotional Facial Expressions (EFE) analysis, classroom observations and self-report questionnaires on levels of enjoyment, anxiety and boredom of three students over a period of several weeks. We argue that the use of sophisticated data analysis techniques and advanced statistical methods can lead to a better understanding of the dynamics of learner emotions while they perform tasks together in the classroom. This multi-pronged approach could ultimately lead to the development a robust tool that can provide real-time feedback on student emotions.

## 4.2 Related Works

The theoretical basis of the present study is Gobin et al. [96]. The authors argue that an emotion arises in response to a particular emotional situation [97] that is more or less significant. Individual responses to this event consist of three components:

1. A component of physiological arousal (motivational dimension) that refers to physiological responses, meaning all the body's internal reactions;
2. A component that corresponds to motor expression and refers to the visible verbal or non-verbal manifestations of emotion, with the most common expression being "Emotional Facial Expressions" (EFE);
3. A final component called subjective feeling that reflects subjective awareness and englobes the cognitive-experiential responses to emotion. It consists of all cognitive processes related to the perception of the emotional situation, which the individual can verbalize and explain.

### 4.2.1 Multidimensionality of emotions in FL courses

Research into the use of multimodal data for analysing learner's emotions, particularly in educational contexts, has evolved significantly since the 2000s. The idea of integrating physiological and behavioral data to understand dynamic interactions between motivation, emotions, and Willingness To Communicate (WTC) across educational environments is not new [98] but technological developments have allowed researchers to use sophisticated tools that are increasingly affordable to applied linguists. Here, we review some key studies.

D'Mello et al. [99] explored the use of multimodal data to monitor engagement and learning in real-time. This research utilized a combination of EFE, body posture, and interaction logs to understand how students engage during learning activities in various environments, including classrooms and online platforms. The researchers used advanced and expensive equipment which would be unaffordable to most researchers.

Another study, conducted by Tonguç and Ozkara [94], employed a cheaper facial recognition tool with 67 students during basic information technology courses, using a camera placed in each student's computer. The materials used during the lecture were reflected on students' computer screens. EFE were analyzed and digitized to identify seven emotions/feelings: disgust, sadness, happiness, fear, contempt, anger, and surprise based on Ekman's [100] theory of universal emotions. This approach has been strongly criticized for being too static and essentialist. Feldman Barrett [95] and Gendron and Feldman Barrett [101] have proposed an alternative approach, the theory of constructed emotions which posits that emotions are shaped by the broader cultural context and the more specific social context in which the emotion unfolds. They also reject the idea that emotions have universal unique "fingerprints", in other words, a smile does not automatically reflect

happiness and a scowl does not always signify anger. The dynamic theory of constructed emotions is particularly appropriate when studying multilingual and multicultural individuals performing tasks in a classroom.

Adopting a dynamic approach, Gregersen et al. [90] conducted a multimodal study on three high-anxiety and three low-anxiety learners of Spanish FL (based on scores on Horwitz’s et al.’s FL Classroom Anxiety Scale [102]). The authors combined physiological data (heart rates), idiodynamic data (anxiety ratings), and interviews about the fluctuations. Participants did a 3-minute oral presentation in Spanish. They wore heart monitors and, immediately after the task, they provided 42 anxiety ratings on a scale from +5 to -5 while viewing their presentation on a computer. They then explained to the researchers why the spikes and dips in anxiety had occurred. Increased heart rates were positively correlated with anxiety ratings. The high anxiety participants reported difficulties in vocabulary retrieval as the main cause for their anxiety. Low anxiety participants used strategies to mitigate this. They had practised the presentation in the preparation stage rather than attempting to memorize it. The study used various sources of data to highlight the dynamic nature of a single emotion but ignored the fact that participants may have experienced other emotions, which could also have affected heart rates.

The pioneering neurological study by Nozawa et al. [93] was the first to peer into learners “black box”, namely their brain waves, as they were performing tasks in class. The researchers focused on two intact English FL classes in Japan with two groups of four learners each (totaling 16 students). They adopted a multimodal approach, examining the interbrain synchronization among learners working in pairs and the similarity of the flow state<sup>1</sup> dynamics during collaborative learning. Prefrontal neural activities were measured using a wireless functional near-infrared spectroscopy device placed on the students’ heads. Additionally, the researchers asked the learners to watch recorded videos of the classes and to evaluate their own flow levels on a scale from 1 to 7 every two minutes. The study required advanced technology and know-how to process the neurological data in a single cortical area. The authors admit that the correlation between self-reports and interbrain synchronization does not imply causation as there may be “hidden variables” (p. 10).

These studies highlight the diversity of methodologies and technologies used in the multimodal analysis of learners’ emotions in the FL classroom. All relied on supervised

---

1. Csikszentmihályi (1990) defined flow as “an optimal psychological and physiological state characterized by intense concentration, a sense of harmony, a clear goal, a loss of the sense of time, a balance between skill and challenge, a total absence of boredom and anxiety, and a profound enjoyment that contributes to a more general sense of well-being”.

learning approaches that require extensive annotation and labeled data, which can be resource-intensive, time-consuming and expensive.

### 4.2.2 Anxiety, boredom, and enjoyment in FL classes

The introduction of positive psychology in the field of FL acquisition with the publication of MacIntyre and Gregersen [103] made researchers aware that there had been a long-term exclusive focus on negative emotions in FL classrooms [104] and anxiety in particular. Its popularity among researchers had been boosted by Horwitz et al. [102], whose 33-item Foreign Language Classroom Anxiety Scale (FLCAS) covered physical symptoms of anxiety, nervousness, and lack of confidence in the FL class. They defined FLCA as “a distinct complex of self-perceptions, beliefs, feelings, and behaviours related to classroom language learning, arising from a uniqueness in the language learning process” (p. 128). The main cause of FLCA is the inability to project an accurate image of themselves in the FL and the fear of coming across as clumsy and inauthentic [105]. FLCA grows gradually through repeated anxious experiences in the FL classroom. As such, FLCA starts as being a situation-specific state and gradually becomes more stable and trait-like [105].

A meta-analysis by Botes et al. [106] of 67 studies based on the FLCAS has shown that FLCA is moderately negatively linked to FL performance and progress. High FLCA was linked to lower general academic achievement and lower speaking, listening, reading, and writing performance in the FL. Students suffering from high FLCA have a lower degree of Willingness to Communicate (WTC) and may even prefer to hide and remain silent in the classroom [107]. This withdrawal from classroom interactions slows their progress in the FL.

Another negative emotion frequently present in FL classrooms is boredom [108]. Li et al. [109] described it as being characterized “by negative valence, low arousal and being achievement-related activity-focused” (p. 244). Boredom emerges when a classroom activity or task is perceived as irrelevant and when learners feel helpless and fatigued because the activity is either too easy or too difficult [110, 109]. Bored learners lose their confidence and suffer from a perceived lack of control. This lowers their WTC and undermines both their short-term and longer-term motivation, as well as their overall engagement in the FL activities. Learners’ boredom can also originate in the teacher’s inability to hide their own boredom [108]. Li et al. [109] developed a 32-item FLLB scale consisting of 7 factors. The first factor was named Foreign Language Classroom Boredom (FLCB) and consisted of 8 items, which has since been used independently in

later research. Dewaele and Li [111] showed that teacher enthusiasm can counter learners’ FLLB, increase their enjoyment, and stimulate their engagement. Li [112] found enjoyment and boredom to be strongly negatively correlated. Unsurprisingly, a negative relationship exists between FLLB and FL achievement [113].

Researchers increasingly agree that positive emotions such as Foreign Language Enjoyment (FLE) should be part of a more holistic picture and they reject the deficit view of FL learners [114]. The authors referred to Csikszentmihalyi [115] who noted that enjoyment (and sometimes flow) emerges when a person manages to complete a challenging task, reach a state of full concentration, perform a task with clear goals, and receiving immediate feedback on the performance. Dewaele and MacIntyre [114] developed the 21-item FLE scale, which was followed by a shorter 9-item psychometrically validated scale [116]. While a majority of the longitudinal studies on FL emotions focused on change over a period of weeks and months, a smaller number of studies have focused on fluctuations over shorter time spans. Boudreau et al. [117], for example, used the idiodynamic method to investigate second per second fluctuations in FLE and FLCA. Anglo-Canadian participants completed a one-minute speaking task in French FL and then watched the recording and reported their levels of both emotions for every second. Values were found to vary considerably during that minute and were later commented on by participants who pointed to linguistic difficulties or to fleeting moments of anxiety, enjoyment or boredom. Elahi Shirvan and Talebzadeh [118] also used the idiodynamic approach to investigate the fluctuations in FLE of 7 Iranian EFL university learners participating in conversations on simple and more difficult topics. The results showed strong intra- and inter-individual variation linked to the conversational topics. Adopting a multi-case study design, Elahi Shirvan et al. [119] investigated fluctuations in FLE over different time spans, ranging from seconds with the idiodynamic method, to minutes, weeks and months. The researchers used low tech “Enjoymeters” (pieces of paper with thermometer-shaped figures ranging from 0 to 10 indicating the level of FLE) to capture variation in FLE for periods of 5 minutes. Participants were two Iranian EFL students in the classroom. FLE among these two students was found to fluctuate differently over different timespans. The variation was found to be linked to unique social and personal factors, such as the ability to be creative, the appropriate challenge, the opportunity for authentic communication in English with peers, the teacher’s ability to be supportive, humorous and establishing a positive classroom climate. Being laughed at by peers for making a mistake could cause a sudden drop in FLE and a spike in anxiety. Bielak and Mystkowska-Wiertelak [120] also used the idio-

dynamic methodology to investigate fluctuations in FLCA, FLE in 10 Polish EFL learners working in pairs and group. Their interactions were video recorded and then viewed and rated second per second for FLE and FLCA. In later stimulated-recall interviews, they discussed the causes of the fluctuations and the emotion regulation strategies they deployed to control them. The two emotions showed brief periods of stability followed by highly idiosyncratic levels of fluctuation. Levels of FLCA were found to fluctuate more than FLE but the triggers for the fluctuation in both emotions overlapped substantially. Causes for FLCA included the awareness of having made specific errors, frustration at the lack of linguistic sophistication, and ignoring task instructions. FLE was found to be linked to the quality of the peer's performance and a productive collaboration. Taking the floor and deploying new knowledge caused peaks in both FLE and FLCA while yielding the floor cause a dip in both emotions.

Dewaele and Pavelescu [121] used a multiple case study approach to investigate the relationship between FLE, FLCA and WTC in two Romanian secondary school EFL learners. Qualitative data including classroom observations and semi-structured interviews on the emotional sources of fluctuation in participants' WTC in the English classroom. FLE and FLCA were found be influenced by a range of contextual factors including seating arrangements, course material and conversation topic which shaped their WTC in dynamic and unique ways.

The meta-analysis by Botes, Dewaele et al. [122] showed that FLE is strongly positively correlated with WTC. Moderate positive relationships emerged between FLE and FL academic performance.

A crucial awareness that emerged from previous research is that learner emotions do not exist in isolation. Studies on large samples reveal positive correlations between positive emotions, and between negative emotions, combined with negative correlations between positive and negative emotions. This suggests that there is a strong probability that students who are enjoying themselves are less likely to suffer from anxiety and boredom. As the idiodynamic studies show, the emotions are constantly connected with each other, with motivation and linked to the immediate classroom environment and the wider social context [123, 124, 113]. The teacher is central in this context and his/her perceived enthusiasm or happiness can cause a wave of positive emotional contagion [125, 126, 127]. A teacher who cannot hide his/her boredom will spread this negative emotion to students [108]. The relationship with peers and with a specific partner in pair-work will also shape individual learner emotions. Collaborating with someone who is very anxious or bored



will drag down the enthusiasm of learners who actually enjoy the activity. On the other hand, working with an empathic, friendly, funny partner might boost learners’ positive emotions and lower their negative emotions. The task and activity itself will also shape learners’ emotions, as they may -or may not- enjoy it and grow bored with it if it lasts for too long [128].

To sum up, this short overview of the existing literature shows a field in rapid transition because of the emergence of a holistic understanding of learner emotions and of their dynamic nature, combined with technological innovations. Studies using the idiodynamic method used material collected laboriously over a period of no more than a few minutes. They also focused on no more than two emotions in order not to overwhelm participants.

Very few of these studies included a physiological measure. We thus argue that there is an urgent need for multimodal studies focusing on a larger number of emotions tracked for longer periods in real-world FL classroom environments using economical and efficient techniques. The combination of non-intrusive, low-cost instruments with robust data processing techniques, could lead to the development of a scalable and replicable system.

The current study thus proposes to analyze emotions of a small number of participants over a period of several weeks. We adopted a design similar to that of Nozawa et al. [93] with external technical or digital measurement instruments (pulse oximeter and camera placed in the classroom) and self-perception data collected from students (enjoyment, boredom, and anxiety questionnaires completed at the end of the data collection period).

Our two following research questions (RQ) are:

- *RQ<sub>1</sub>*: Do multimodal methods of data collection, i.e. measuring HR, EFEs and class observations allow researchers to gain an overall view of the emotions experienced by students in language classes?
- *RQ<sub>2</sub>*: Do the scores obtained through the FLE, FLB, and FLCA scales, along with the learners’ responses to the open-ended questions, help to better understand and interpret their physiological results (HR variation) and their EFE linked to the class observations ?

## **4.3 Materials and methods**

### **4.3.1 Study design and participants**

The project took place at the Center X (CeLFE) at the University of X. Participants were preparing for the University Diploma in French Studies (DUEF) at the beginner level, A2 according to the Common European Framework of Reference for Languages (CEFR). The courses were held over four days in the second semester of the academic year, from February to April 2023. Our study was conducted over 16 sessions of 2 hours and 40 minutes each ( $2 \times 1$  hour and 20 minutes). Three students, over 7 volunteers, participated in all sessions: Mitch, a 21-year-old American; Zeynep, a 23-year-old Turkish student; and Oksana, a 23-year-old Ukrainian student. Students voluntarily signed a consent form authorizing the recording of their interactions and the use of their data for research purposes, while excluding the direct public release of their data. Of the eleven students in the class, seven provided consents; however, only those consistently present across the majority of sessions (three students) were included in the recordings. A research team member took comprehensive notes for all 16 sessions. During each session, both their HR signals and EFE data were collected. Additionally, at the end of each session, participants completed a questionnaire.

### **4.3.2 Instruments**

Three instruments were used to tap into learners' physiology, EFE and self-report of FLE, FLB and FLCA.

#### **4.3.2.1 Physiological reaction: measurement of HR variation**

An electrocardiogram (ECG) measures ECG signals, which can be used to predict numerous features such as heart rate (HR), interbeat interval, and HR variability [129]. For our task, we selected an affordable wearable device called the "Fingertip Pulse Oximeter," with characteristics detailed in Annex B and Fig. 4.1. A pulse oximeter measures heart rate by detecting pulsatile changes in blood volume using the photoplethysmogram (PPG) signal. The device emits red and infrared light through the finger, and a photodetector on the opposite side measures the transmitted light. The pulsatile component of blood flow generates a PPG waveform, representing variations in blood volume with each heartbeat. The heart rate is calculated by measuring the time interval between successive peaks in

the PPG waveform, determining the number of heart beats per minute [130]. This method is reliable and accurate for HR up to 155 bpm, suitable for non-strenuous activities [130, 131], Therefore, this instrument meets our needs as it is cost-effective and user-friendly. Students can pair the oximeter with the ViHealth application via Bluetooth, and the recordings are stored on the student’s phone in PDF format.



Figure 4.1 – Left: Oximeter used to record the heart rate beats of the students. Right: ViHealth application that receive the data from the oximeter.

#### 4.3.2.2 Expressive behavioural responses: EFE

To collect visual data for recognizing student EFE, we selected an affordable camera, the Razer Kiyo Pro (C100), as illustrated in Fig. 4.2. At the start of each session, the camera was positioned to ensure comprehensive coverage of all students’ faces in the classroom. This device was configured to record RGB video at 30 fps with a resolution of  $1280 \times 720$  pixels, providing a clear visualization of each student’s facial expressions when they faced the camera. However, this modality faced several challenges, including varying lighting conditions, obstruction of student faces by the teacher passing in front, and instances where students hid their faces or did not look directly at the camera. A simple view of the teaching classroom can be seen in Fig. 4.2.

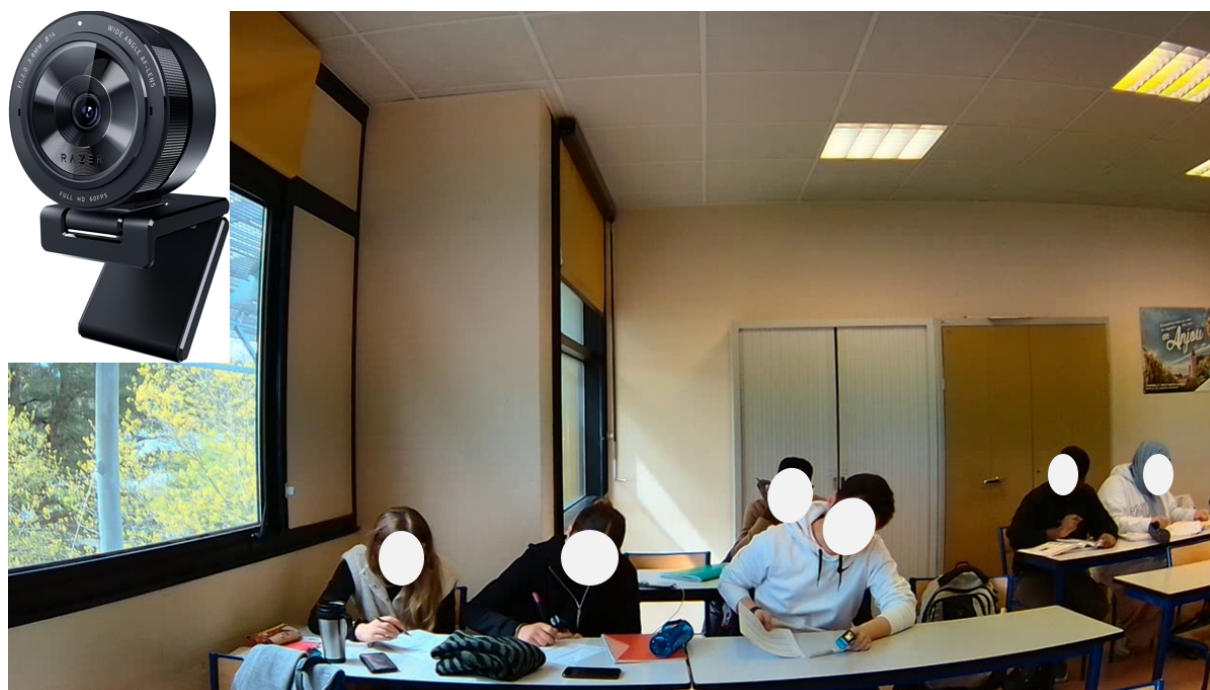


Figure 4.2 – Camera used to record RGB videos (top left) and a sample of a teaching session from NeuroCam dataset.

#### 4.3.2.3 Cognitive-experiential responses: the enjoyment, boredom, and anxiety questionnaires

At the end of each of the 16 sessions, we administered the short version of the enjoyment questionnaire [116] to the three volunteers retained for the study. This questionnaire consists of nine items that assess enjoyment in the French FL class across three dimensions: teacher enjoyment (e.g., "The French teacher is kind"), personal enjoyment (e.g., "I am proud of my progress in French"), and social enjoyment (e.g., "We support each other in the French class"). Participants also filled out the 8-item subdimension Foreign Language Classroom Boredom (FLCB) [109]. These items address lack of concentration, fatigue, and restlessness, such as "My mind begins to wander in the French class." Finally, participants filled out the 8-item short form of Foreign Language Classroom Anxiety scale (S-FLCAS), employed by Dewaele and MacIntyre [114] and validated by Botes et al. [132]. These items refer to physical symptoms of anxiety, nervousness, and lack of self-confidence. Two items refer to low anxiety, such as "I am not afraid of making mistakes in the French class," and six items indicate high anxiety, such as "I become nervous and confused when

I speak in my French class." Items were accompanied by a 5-point Likert scale (1. strongly disagree, 2. disagree, 3. neither agree nor disagree, 4. agree, 5. strongly agree).

The closed items were complemented by two open-ended questions allowing students to freely express their feelings and emotions in their own words. These two questions were concrete, asking students to describe a specific situation in class where they felt really good, a moment when they felt bad, and what they precisely felt at that moment. The qualitative material gathered in this way allowed us “to hear the voices of participants, free from the shackles of the Likert scale items.” [133]. Dörnyei [134] encouraged researchers to include open questions in questionnaires because they “can provide a far greater richness than fully quantitative data.” The two open-ended questions could be answered in French or English. Thus, Oksana and Mitch responded in English, while Zeynep responded in French. The quantitative part of the questionnaire was used for purely illustrative purposes as no inferential statistics could be calculated. Combined with the answers to the open-ended questions, they forced participants to think about the FLE, FLCA and FLCB in the classroom and provided a basis for the interviews.

### 4.3.3 Preprocessing Methods

After detailing the recording instruments per each modality, we can notice that the raw outputs from the instruments of HR and EFE modalities are not suitable for direct data processing. In the following subsections, we will detail the preprocessing methods applied on these outputs to extract the HR signal from the oximeter’s image output and to track each student face in the RGB video.

#### 4.3.3.1 HR signal extraction

The output of the oximeter devices in the **ViHealth** application is an image containing two signals: the oxygen level percentage ( $O_2$ ) in the blood and the pulses representing the heart rate (bpm), sampled at 30 Hz. An illustrative example of this output image is depicted in Fig. 4.3. Our goal is to acquire the digital values from this image output to extract patterns from the HR variation signal.

To extract the heart rate (HR) digital curve from the image section denoted as (g) in Fig. 4.3, we employed a series of image processing techniques. First, we manually cropped the section of the graph labeled "Fréquence du pouls (/min)" from the output image (Fig. 4.3). Next, we used LeNet-5 [135] and image filters, such as Sobel, to identify and

recognize all digits present in the cropped image, determining the center pixel positions of these digits to understand the resolutions along both the x-axis and y-axis. Finally, we applied threshold segmentation to isolate the pixels corresponding to the green-colored areas representing the heart rate pulses. Since multiple y-values could correspond to a single time instant, we used simple averaging across the y-axis to streamline the data.

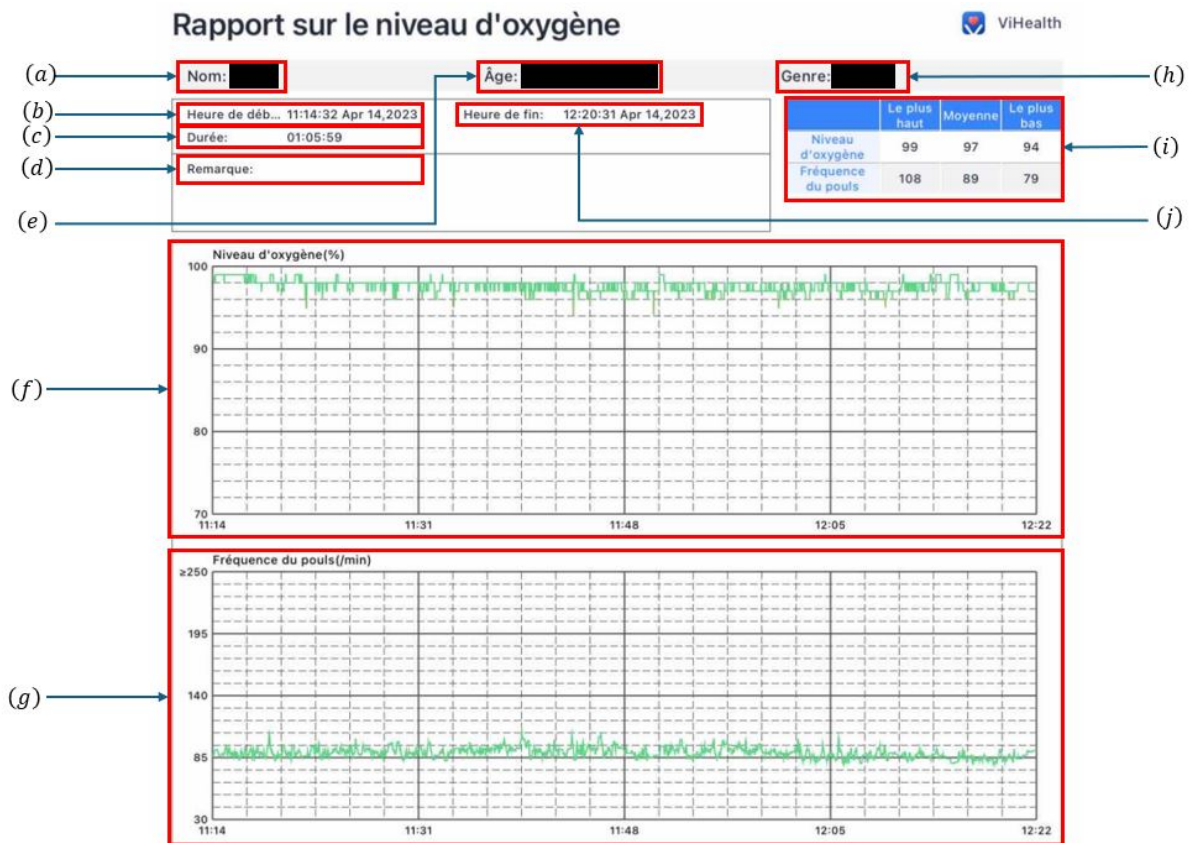


Figure 4.3 – Example of the output image from our oximeter. (a) is the name of the participant, (b) the date and the hour of the beginning of the measurement, (c) test duration, (d) remarks, (e) the birthday date and the age, (f) the signal curve of the oxygen level during the session (y-axis: oxygen level % and x-axis: time instant of the session), (g) the heart rate variation of the student during the teaching session (y-axis: heart rate *bpm* and x-axis: time instant of the session), (h) gender, (i) table representing the statistics (max, mean, min) of the oxygen level % and the heart rate *bpm*, (j) the end date and hour of the recording.

We should note that the sampling frequency rate depends on the recording time and the image resolution. Therefore, the extracted signals had many sampling frequency rates.

The approximation done by our preprocessing method can be found in Fig. 4.4.

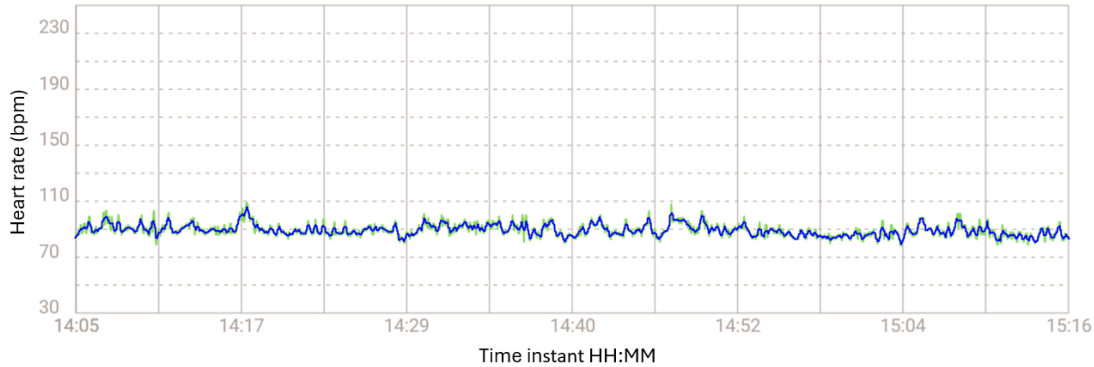


Figure 4.4 – Heart Rate signal of a student during a session. The green curve corresponds to the real and the blue one is its approximation using our method described in 4.3.3.1 (x-axis: time instant of the session, y-axis: the heart rate of the student (bpm)).

#### 4.3.3.2 Face recognition and tracking

Our camera records an RGB video of the whole class at  $1280 \times 720$  pixels. During the recording sessions, we faced several constraints, such as varying distances between each student and the camera depending on the session and the classroom, differing luminosity conditions, students occasionally having extreme head poses and the teacher sometimes passing in front of the camera. Therefore, we need a robust tool to track each student face and store it into a video. For this purpose, we performed a downsampling on our video sessions to 1 fps and stored all frames in a folder. For each frame, we extracted and sorted all corresponding faces using the GhostFaceNet network [136] from a lightweight face recognition Python module called DeepFace [137, 138, 139]. The indexing of frames facilitated the tracking of individual faces, we specify that we did not use face verification methods due to their poor performance under these constraints. After these steps and to verify the absence of outliers i.e overlapping the faces, we manually verified the face tracking results. We only retained faces with head poses that allowed us to predict the student’s emotions.

### 4.3.4 Data analysis

#### 4.3.4.1 Proposed Pipeline for the HR and the EFE

To monitor student engagement, we combined physiological signals, specifically heart rate (HR) variations, with behavioral signals, such as emotional facial expressions (EFE). While previous studies [7, 6, 140, 141] have highlighted HR data as a reliable measure of physiological states, relying solely on one modality may overlook important behavioral cues that provide additional insight into engagement. EFE can capture cognitive and emotional states that may not be fully reflected in HR data, offering complementary information. By integrating both HR and EFE, we can leverage the strengths of each modality, as they together offer a more holistic view of student engagement. To effectively combine these two sources of information, we adopted a decision-level fusion approach, where HR and EFE are processed independently to detect anomalies. The final decision on engagement incorporates both modalities when available, but the method remains robust even in the absence of one modality. This strategy ensures flexibility and improves the accuracy of engagement detection by compensating for limitations in any single data source, offering a more comprehensive and reliable system. The pipeline consists of two separate sub-pipelines, each dedicated to processing a different modality independently.

Given the temporal nature of HR modality, the preprocessed HR data from each session is segmented into equal frames of duration  $T$  with a 50% overlap. To capture HR signal features for emotion recognition purposes, researchers have proposed various methods [7] to reduce the amount of the data by projecting the signal onto a latent space, primarily using statistical features. Let  $X_n$   $n \in [1, N]$ , the extracted HR features vectors are measured as the following:

- the mean of the raw signal  $\mu_X = \frac{1}{N} \sum_{n=1}^N X_n$ ,
- the standard deviation  $\sigma_X = \sqrt{\frac{1}{N} \sum_{n=1}^N (X_n - \mu_X)^2}$ ,
- the mean of the absolute value of the first difference  $\delta_X = \frac{1}{N-1} \sum_{n=1}^{N-1} |X_{n+1} - X_n|$ , representing the average speed of the heart during the sequence,
- the mean of the absolute value of the first difference of the normalized signal  $\Delta_X = \delta_X / \sigma_X$ ,
- the mean of the absolute value of the second difference of the signal  $\gamma_X = \frac{1}{N-2} \sum_{n=1}^{N-2} |X_{n+2} - X_n|$ , representing the average absolute acceleration of the heart during the sequence,
- the mean of the absolute value of the second difference of the normalized signal



$$\Gamma_X = \frac{\gamma_X}{\sigma_X}.$$

Therefore, for each chunk of fixed duration  $T$  (with varying sampling frequency), we obtain six feature vectors that define the student’s state based on their physiological signal. For each student, the feature vectors from all sessions are clustered in an unsupervised manner into two groups: normal moments, representing most of the time, and Significant Moments (SM), indicating deviations from the norm. We called these significant moments because they were identified as instances that did not align with the standard baseline of physiological responses observed in the three students. We therefore wanted to understand precisely what was happening for them at those specific moments by correlating them with classroom observations and triangulating with the questionnaires. This subject-dependent method introduces a personalized approach that accounts for individual variations in health status and cultural background. Since resting heart rate (HR) can be influenced by various factors, this method provides a more accurate and context-sensitive assessment of each student’s physiological data [7].

Simultaneously, the video output from our RGB camera, initially recorded at 25 frames per second (fps), was downsampled to 1 fps and preprocessed to track only the individual corresponding to the HR measurements. The decision to downsample to 1 fps was driven by the need to balance computational efficiency and the temporal resolution of EFE. Since significant changes in facial expressions typically occur over a span of seconds rather than milliseconds, capturing frames at 1 fps is sufficient to detect these variations without unnecessary redundancy. This approach reduces the computational load while still providing the necessary granularity to accurately track and analyze facial expressions in sync with the HR data. For the EFE modality, we applied a Vision Transformer (ViT) designed for image emotion recognition [142] on each frame of the student face across the video. This transformer based architecture, depicted in Fig. 4.5 and adapted from [142], was pre-trained on the FER-2013 dataset [143] and available on Hugging Face platform [144], which recognizes seven emotions (angry, disgust, fear, happy, sad, surprise, and neutral) with an accuracy of 90.92%. At the input, the architecture processes the student’s face image, which has been resized to  $224 \times 224$  pixels. This image gets projected into a 1D high dimensional space thanks to a 2D convolutional layer with a kernel size of  $16 \times 16$  and a stride of  $16 \times 16$ , producing  $768 = 16 \times 16 \times 3$  output channels. Consequently, this operation divides the image into  $\frac{224 \times 224}{16 \times 16} = 196$  patches, along with one position embedding vector. The Transformer Encoder of  $L = 12$  layers is composed with Multi-Head Self-Attention (MSA) and Multilayer Perceptron (MLP) blocks as depicted in Fig. 4.5, it

receives a patch embedding  $X \in \mathbb{R}^{197 \times 768}$  to capture the global relationships within the image. At the last layer, depending on the features vectors, an emotion of 7 is attributed to the image. The number of parameters is  $86M$ .

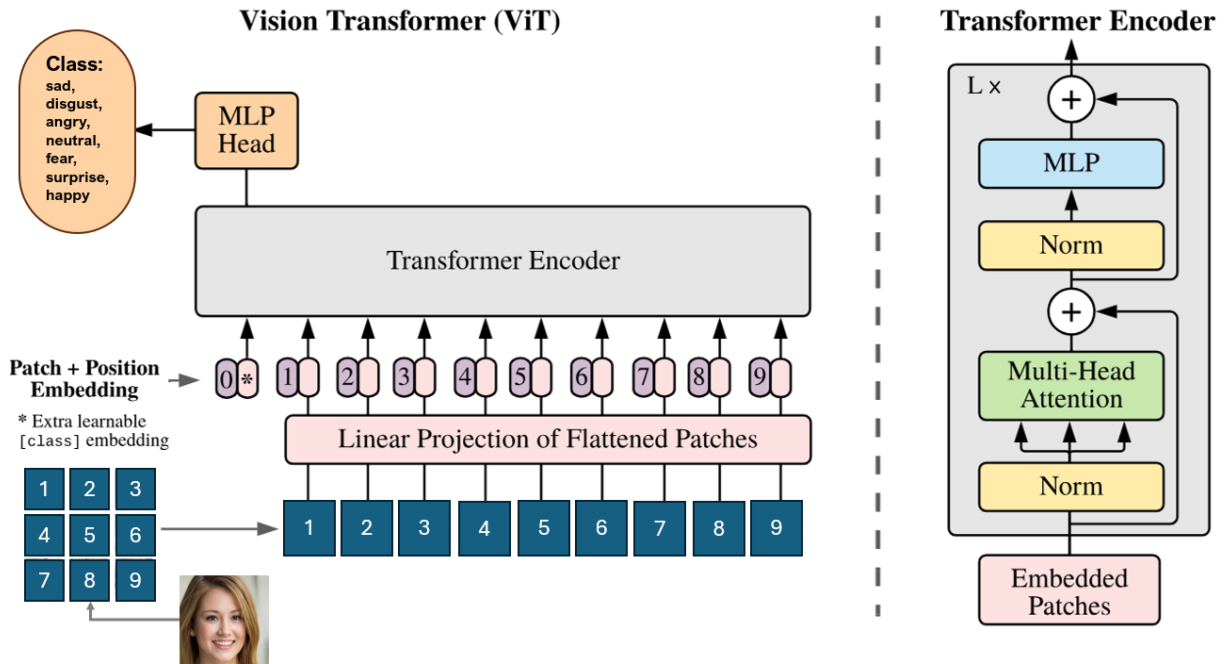


Figure 4.5 – Architecture used for emotional facial expressions recognition taken from [142].  $L$  is set to 12 in the pre-trained model.

The resulting feature vectors are concatenated across all sessions and clustered similarly to the HR signal data to identify SM based on facial expressions. But as we explained earlier, the identification of these 7 emotions mainly allowed us to measure a deviation from the average for each learner, since identifying emotions on learners' faces doesn't really make sense according to Feldman Barrett's theory of constructed emotions. What matters is identifying the most frequent EFE for the learner and seeing when there is a variation, called Significant Moments (SM), compared to this standard emotion for them.

Our entire pipeline for these two modalities is illustrated in Figure 4.6, where the "decision level" fusion means that priority is given to HR data, as referenced in the works [140, 145, 6]. When HR data is available, it takes precedence, otherwise, significant moments are detected using EFE.

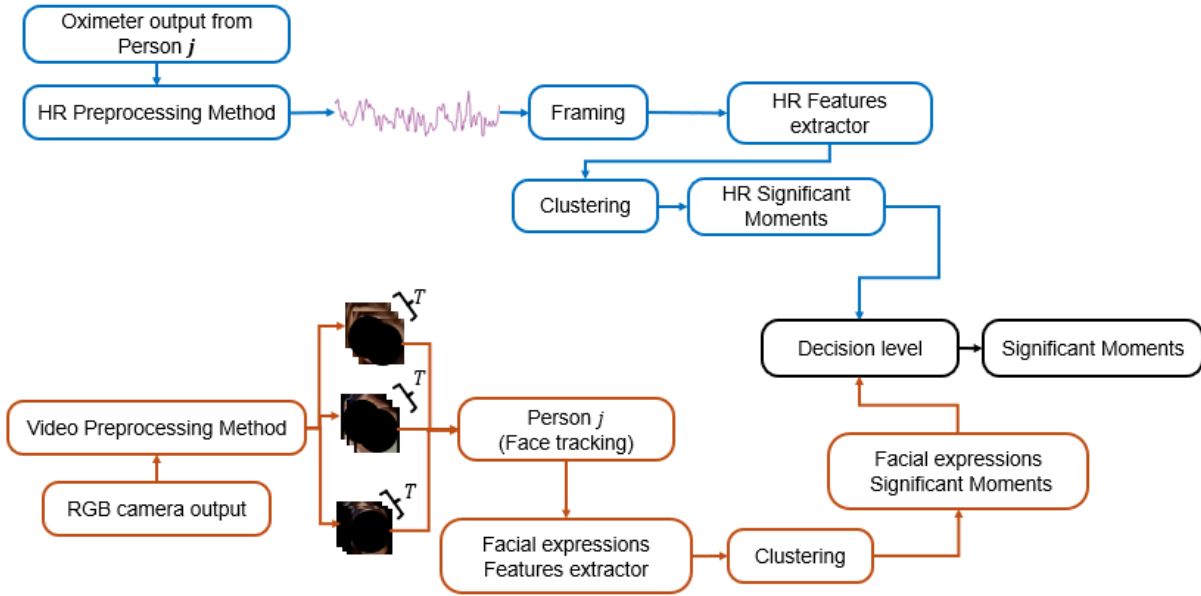


Figure 4.6 – Our proposed pipeline method to detect significant moments during a teaching session.

#### 4.3.4.2 Experimental setup for the HR and EFE

In the following paragraph, we introduce the experimental setups on subject-dependent to monitor students’ emotions using two modalities: HR signals and facial expressions.

For the HR signal, due to its temporal nature, we should establish a specific time window  $T$  on which we can apply the HR statistical features ensuring a  $T$  value that does not affect the emotional decision. According to the literature [6, 146] time interval between an emotional stimulus and the subsequent physiological response varies due to factors such as individual differences and signalling modality. This variability complicates the task of defining a suitable window size for emotion recognition systems. Kreibig found that the most common average time intervals for physiological responses were 60 s and 30 s in a survey of 134 publications [6, 146]. Other common average intervals were 0.5 s, 10 s, 120 s, 180 s and 300 s [6]. Therefore, for the HR signal processing, we choose to work on window of duration  $T = 60$  s, 90 s, 120 s, 150 s, 180 s and 210 s with a 50% overlapping (a hyperparameter that can variate from a physiological signal to another [6]). We segment the heart rate signal of each student in each session, into  $T$  with an overlap of 50%. HR features vectors corresponding to each session are clustered, using Gaussian Mixture Models (GMM) to identify the outliers which correspond to the significant moments. Fig.

4.7 states the pipeline used to track a given student on all teaching sessions  $S_1, \dots, S_M$ .

The final SM are those resulting of the union of all time duration windows.

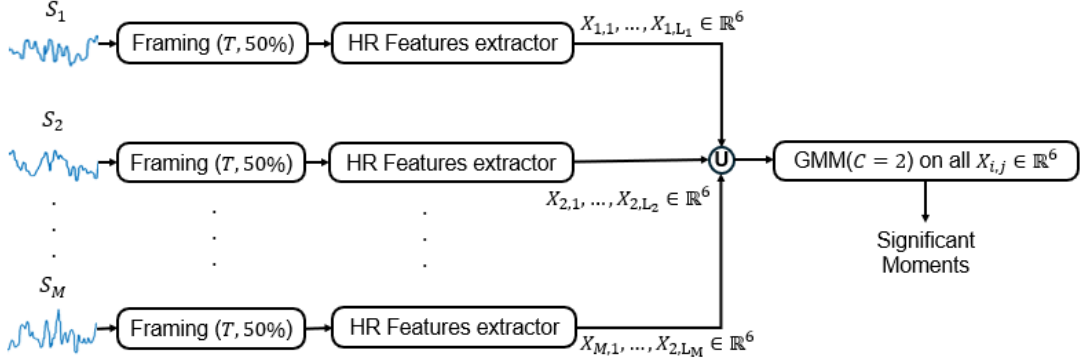


Figure 4.7 – The pipeline designed for tracking the student engagement based on the HR signal.  $T$  is the duration of the framing.

For the EFE analysis, in contrast to HR signal processing, where variable  $T$  chunk durations are employed based on flexible guidelines [146], we focused on frame-level recognition. This approach was selected due to the demonstrated reliability of using individual facial images for emotion recognition tasks in deep learning models [147]. Each processed facial image is passed through a Vision Transformer (ViT) architecture pre-trained for emotion recognition [142]. The extracted feature vectors are then clustered using a Gaussian Mixture Model (GMM) to identify the corresponding SM. The frame-level analysis allows for precise detection of dynamic and subtle emotional variations, ensuring robust identification of significant behavioral patterns while maintaining computational efficiency. This subject-dependent method offers a personalized approach that takes into account individual differences in emotional state and cultural background. Given that emotional facial expressions can be influenced by a range of factors, this approach enables a more precise and context-aware evaluation of each student’s behavioral data.

#### 4.3.4.3 The processing of data from the survey

The questionnaire yielded two types of data: both the scores obtained on items related to enjoyment, boredom, and anxiety for each session, even though the number of sessions for the three learners varies due to occasional absences. These scores are already a good indicator of the students’ emotional state. These scores are supplemented by the students’ responses to open-ended questions, which provide an even more precise insight into the

emotions experienced during the various classes. Sometimes, the self-reported data do not quite match the scores for enjoyment, boredom, and anxiety. For example, during session 9, Mitch has a high enjoyment score (3.77/5) and a lower boredom score (2.875/5). Based on these two scores, we can say that the dominant emotion for Mitch during this session is enjoyment. However, in his responses to the two open-ended questions, he states: “I was happy to finally start learning *passé composé*, but that’s about it otherwise. I don’t feel much normally in class. Just bored. I didn’t feel anxious in class today. It was fine, but I was a little bored. I prefer getting called on to stay engaged or my mind wanders and I stop paying attention.” The dominant emotion in his comments is clearly boredom: thus, the self-reported data diverge in this case. In other cases, on the contrary, the self-reported data are completely convergent. The scores and the responses to the two open-ended questions align completely. For instance, during session 6, Oksana has an enjoyment score of 3.88, a boredom score of 2, and an anxiety score of 2.125, and she states: “Today was a good day. I was happy to speak and happy to prepare the text for this class. I wasn’t stressed today.”

While we sought convergences between the self-reported data, we also aimed to evaluate the convergences between the three components, namely between the HR’s SM (Component 1:  $C_1$ ), the EFE’s SM (Component 2:  $C_2$ ) and finally with the classroom observations (OBS). These observations are primarily descriptive; they were used to precisely describe the course’s progression and to note the students’ activities during the different phases of the class.

## 4.4 Results

### 4.4.1 Necessary clarifications about the experimental results

We first discuss the experimental results of the proposed pipelines presented in Fig. 4.7, applied to the dataset, in relation to the performance of the HR features in detecting anomalous moments. Figures 4.8, 4.9 and 4.10 illustrate the HR feature space following an unsupervised clustering via GMM applied to HR features extracted from 150-second segments. In these figures, we can clearly observe the formation of two clusters for students 1 (Oksana) and 3 (Zeynep). However, student 2 (Mitch) presents a different pattern with a small number of outliers appearing far from the dense area (norm).

To evaluate the cohesion and separation of the clusters resulting from the unsupervised

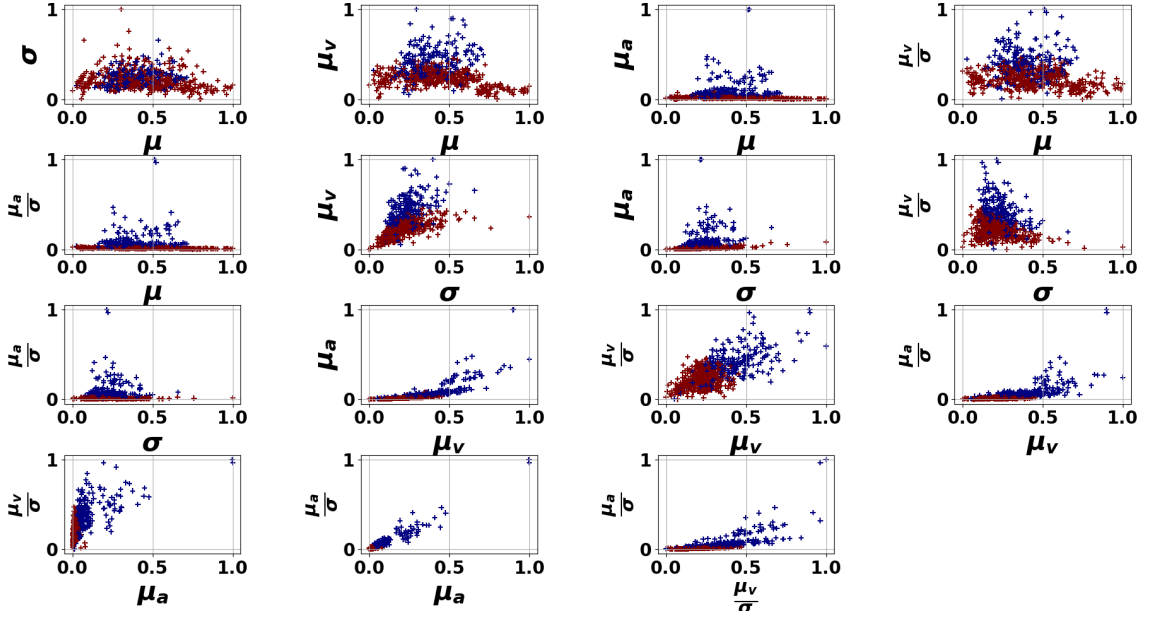


Figure 4.8 – GMM applied on the HR features space for Oksana.  $\mu, \sigma$  refer respectively to the mean and the standard deviation of the HR signal,  $\mu_v, \mu_a$  refer respectively to the mean of the speed and the absolute acceleration of the HR signal.

GMM, we use the mean of silhouette score metric over all samples [148]. Specifically, for a sample  $i$  from the data, the silhouette score  $s(i)$  is calculated as follows:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (4.1)$$

where  $a(i)$  represents the average distance between  $i$  and all other points within the same cluster, capturing the intra-cluster cohesion, and  $b(i)$  denotes the minimum average distance between  $i$  and all points in any other cluster, representing the closest inter-cluster separation. The silhouette score  $s(i)$  ranges between -1 and 1, with values close to 1 indicating well-clustered samples, values near 0 suggesting boundary points between clusters, and values below 0 indicating possible misclassification of  $i$  to its assigned cluster. Fig. 4.11 presents the silhouette scores for each GMM clustering applied to the three students. Notably, it illustrates the dominance of HR acceleration  $\mu_a$  and its normalized value  $\frac{\mu_a}{\sigma}$  in the clustering process, followed by  $\mu_v$  and  $\frac{\mu_v}{\sigma}$ , suggesting that these four features play a significant role in differentiating between the significant moment of the students and their normal behavior. By focusing on these specific physiological parameters, it might be possible to improve the accuracy of clustering or classification models that rely on HR sig-

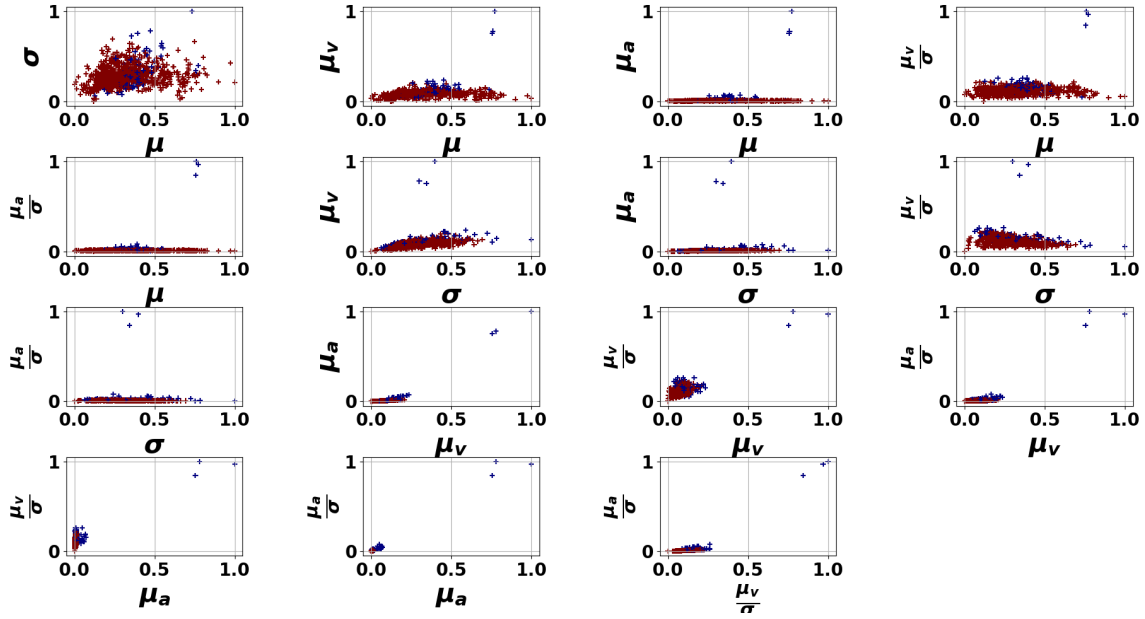


Figure 4.9 – GMM applied on the HR features space for Mitch.  $\mu, \sigma$  refer respectively to the mean and the standard deviation of the HR signal,  $\mu_v, \mu_a$  refer respectively to the mean of the speed and the absolute acceleration of the HR signal.

nals to infer the significant moments of the student. This observation is further supported by findings in emotion classification using physiological signals [149], which demonstrated the superior contribution of these four features over the mean and the standard deviation of HR signal.

#### 4.4.2 Convergences between the cross-referenced results of Components 1 and 2

Integrating HR signals (component 1) and EFE (component 2) could improve the detection of students’ activity and emotional states. Figure 4.12 illustrates the percentage of SM detected through the Intersection over Union (IoU) of HR signals and EFE for the three students. This metric quantifies the overlap between the significant moments identified by HR signals and those identified by EFE. It is important to note that only the anomalies in EFE present during the periods of HR signal recording were considered. For Mitch and Oksana, the IoU between the SM of HR signals and EFE is 21.05% and 20.24%, respectively. These moderate overlaps indicate some consistency between the HR

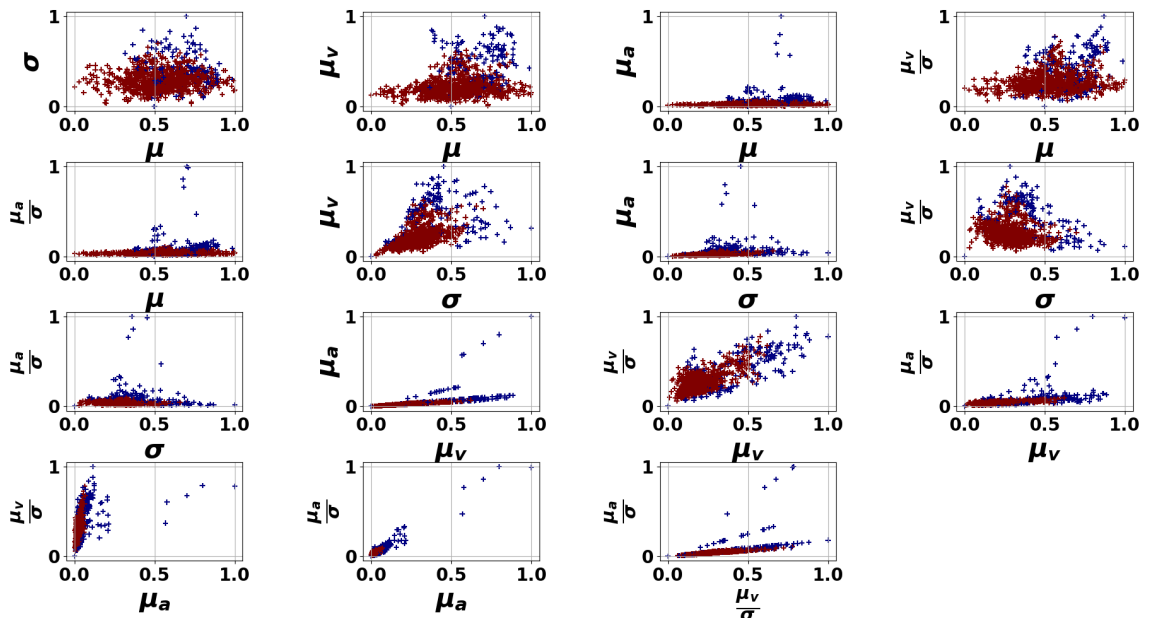


Figure 4.10 – GMM applied on the HR features space for Zeynep.  $\mu, \sigma$  refer respectively to the mean and the standard deviation of the HR signal,  $\mu_v, \mu_a$  refer respectively to the mean of the speed and the absolute acceleration of the HR signal.

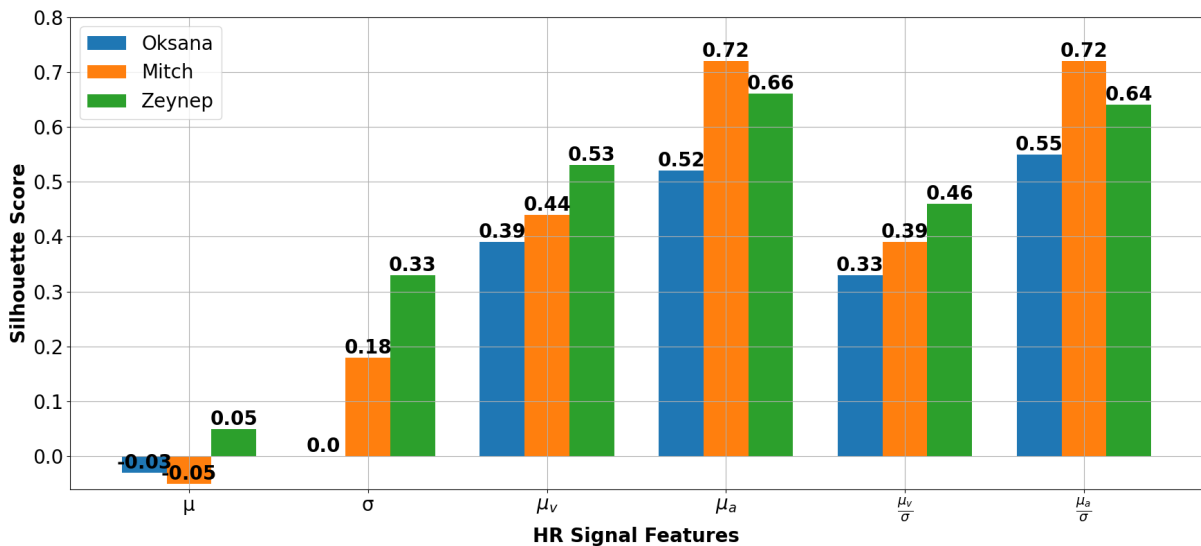


Figure 4.11 – Silhouette scores for each student when using GMM clustering on HR chunks of 2 minutes 30 seconds.



data and EFE, although there may be variability in capturing the students’ moments of activity. In contrast, Zeynep shows an IoU of 27.40%, the highest among the three students, suggesting a stronger convergence between HR signals and EFE for identifying moments of activity for this student. This comparison highlights that although there is some level of agreement between the two modalities for all students, the extent of this alignment varies, with Zeynep presenting the strongest correlation.

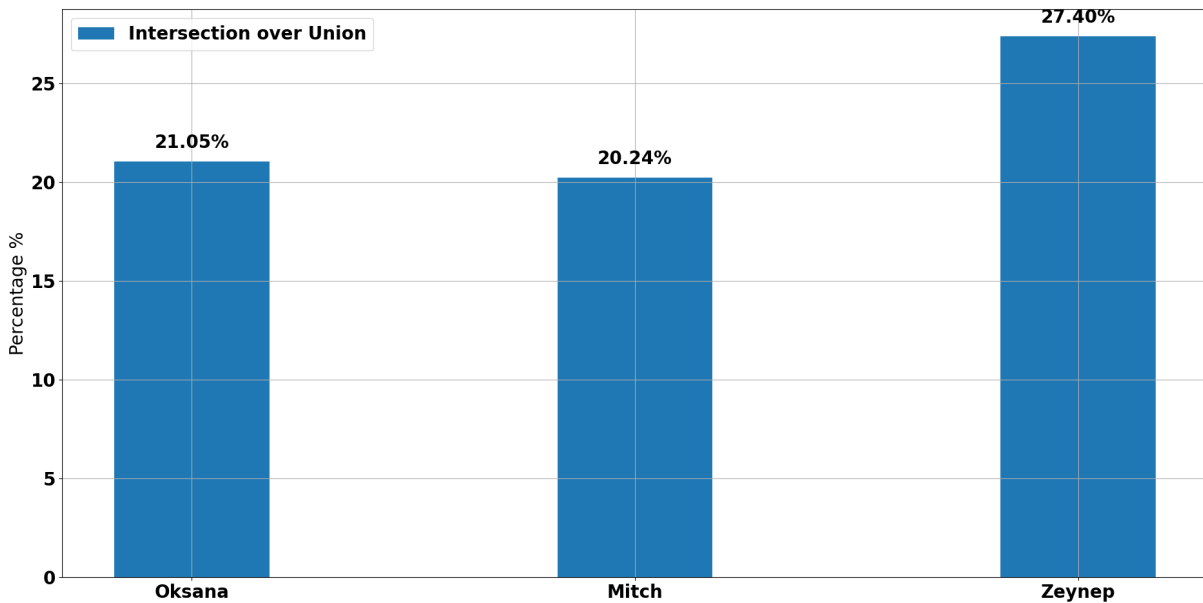


Figure 4.12 – Intersection over Union of HR (Component 1) and EFE (Component 2) SM for each student.

#### 4.4.3 Convergences between the cross-referenced results of Components 1 ( $C_1$ ), 2 ( $C_2$ ) and their fusion and the observations of the courses (OBS)

The results presented in Table 4.1 illustrate the effectiveness of the HR signal-based method for detecting students’ activity during teaching sessions. The table provides a comparative analysis of the percentage of convergence between the results by our HR signal-based method ( $C_1$ ) and the results obtained by our EFE based method ( $C_2$ ) and finally by the students’ observations made during the 16 sessions by an expert in pedagogical methods (OBS).

In addition, fusing both modalities (HR signal-based and EFE-based) through decision fusion achieves the highest convergence rates with expert observations across all students. This outcome highlights that integrating both physiological and behavioral features enhances the accuracy of detecting students' activities compared to using unimodal approach.

The results from the questionnaires and the responses to the open-ended questions ( $C_3$ ) provided a general understanding of the emotions felt by the students during the classes, but they did not give precise indications about specific moments in the class. This is why we were unable to statistically cross-reference the data between  $C_1$  and  $C_3$  or between  $C_2$  and  $C_3$ , as  $C_1$  and  $C_2$  allowed us to identify specific SMs locally during the class. The questionnaires and the responses to the open-ended questions helped us understand the global emotional states of the students throughout the different classes, which allowed us to better interpret the statistical results from  $C_1$  and  $C_2$ , as we will see in the part focusing on each student.

On the other hand, the OBS are very detailed and indicate the course progression with precise time and duration indicators, using the same model as  $C_1$  and  $C_2$  with the SMs. We were therefore able to cross-reference the data between  $C_1$  and OBS, between  $C_2$  and OBS, and also between the multimodal decision fusion of  $C_1$  and  $C_2$ , and OBS, and these results are presented in Table 4.1 below.

In addition to measuring convergence, two additional metrics were introduced: semi-convergence and divergence, to better evaluate the relationships between the different data sets. Convergence measurement reflects full alignment between data sets, while semi-convergence accounts for partial alignment, and divergence indicates complete misalignment. These metrics were applied to assess the correspondence between the self-reported data ( $C_3$ ), as well as  $C_1$ ,  $C_2$ , and observational data (OBS), providing a more detailed and nuanced understanding of the relationships among these data sources.

SMs were identified within the data from  $C_1$  and  $C_2$ , with the aim to determine whether these SMs corresponded to specific events documented during the course. In the case of several students, certain SMs detected by  $C_1$  and/or  $C_2$  were found to coincide with specific tasks the students were engaged in at that moment, which was classified as a moment of convergence between the data sets. In other instances, SMs from  $C_1$  and/or  $C_2$  partially aligned with the student's actions, but the time intervals did not fully overlap; these cases were categorized as semi-convergence. Lastly, in situations where SMs detected by  $C_1$  and/or  $C_2$  did not correspond to any observable activity or task in the class,

indicating that the student was not engaged in a specific action, the instance was classified as divergence. To answer the  $RQ_1$ , for Oksana (Student 1), the agreement between the

<b>Student</b>	$C_1 \cap OBS$	$C_2 \cap OBS$	Decision fusion $\cap$ OBS
Oksana	70%	55%	<b>80%</b>
Mitch	72.22%	62.5%	<b>90%</b>
Zeynep	54.54%	16.67%	<b>63.62%</b>

Table 4.1 – The convergences between:  $C_1$  (results obtained by our HR signal-based method) and OBS (observation of the course progression and students ‘activities’); and between  $C_2$  (results obtained by our EFE based method) and OBS; the Multimodal Decision Fusion and OBS.

classroom observations (OBS) and the HR signal-based method ( $C_1$ ) is significant at 70%, indicating that the HR method is highly effective in objectively detecting when the student is active in class. This convergence shows that the HR method provides a reliable and objective measure of student activity compared to observational methods. In contrast, the convergence between OBS and the EFE-based method ( $C_2$ ) is lower at 55%, indicating that while the EFE method can reflect classroom activity to some extent, it lacks the precision of the HR-based method. The fusion of these two modalities enhances the accordance to 80%, an average improvement of 17.5% compared to unimodal methods, highlighting the need for multimodal benefits in better detecting significant moments at the local time level for this student.

Mitch (Student 2) exhibits a strong agreement of 72.22% between OBS and  $C_1$ , reinforcing that the HR method provides a more objective and reliable measure of classroom activity compared to the subjective or observation-based methods. The convergence between  $C_2$  and classroom observations is also relatively strong at 62.5%, but it is still lower than the HR-based measure, highlighting the HR method’s superiority in offering a more accurate, objective assessment. The fusion of these two modalities improves accuracy by 90%, representing an average increase of 22.64% over unimodal methods. This emphasizes the importance of multimodal approaches in more effectively detecting significant moments at the local time level for this student.

For Zeynep (student 3), the HR signal-based method ( $C_1$ ) shows a moderate agreement with classroom observations at 54.54%, suggesting that while the HR method objectively captures Zeynep’s classroom activity, the complexity of her physiological responses may not always align perfectly with direct observations. Nonetheless, the HR method remains more objective than the EFE-based method ( $C_2$ ), which shows a much lower convergence

of 16.67% with classroom observations. While the unimodal performance has decreased for this student, the fusion of these two modalities still enhances accuracy by 63.62%, representing an average increase of 28.02% over unimodal methods. This underscores the importance of multimodal approaches in more effectively leveraging the complementary information from the EFE and HR modalities, as well as in detecting significant moments at the local time level for this student.

The data in Table 4.1 highlight the ability of our HR signal-based method to effectively detect 3 students' activity compared to the EFE. However, the fusion of these modalities has led to better detection of significant moments at the local time level within the session, shedding light on the need for multimodal benefits in this context. The substantial agreement with students' subjective experiences and the precise, direct classroom observations validates the robustness of our approach. This cross-verification reinforces the credibility of our HR signal-based method, demonstrating its applicability and potential to provide objective and insightful data on student activity in educational contexts.

#### 4.4.4 Detailed analysis of triangulated results with component 3: focus on each student

To answer the first part of  $RQ_1$  and  $RQ_2$ , we will now focus on the three participants during one particular session. As we explained previously, within component 3, the data were not always convergent. As shown by the figure 4.13 of the average scores of the three emotions for the three students, enjoyment is highest. However, the open-ended questions reveal that the dominant emotion for the 14 sessions where Mitch was present is boredom. For Oksana (present at 13 sessions) and Zeynep (present at 14 sessions), enjoyment dominates in the responses to the questions, even though they sometimes experience boredom and anxiety in the French FL class.

We will now examine in more detail one session for each student where the results between the three components and the classroom observations were convergent or semi-convergent. We selected session 14 for Mitch and Oksana where the emotion of enjoyment was predominant for both, and they had moments of shared enjoyment. Session 15 was selected for Zeynep who also reported very high enjoyment.

**4.4.4.0.1 Emotional contagion between Mitch and Oksana** The results concerning Mitch during session 14 are convergent and semi-convergent across all compo-

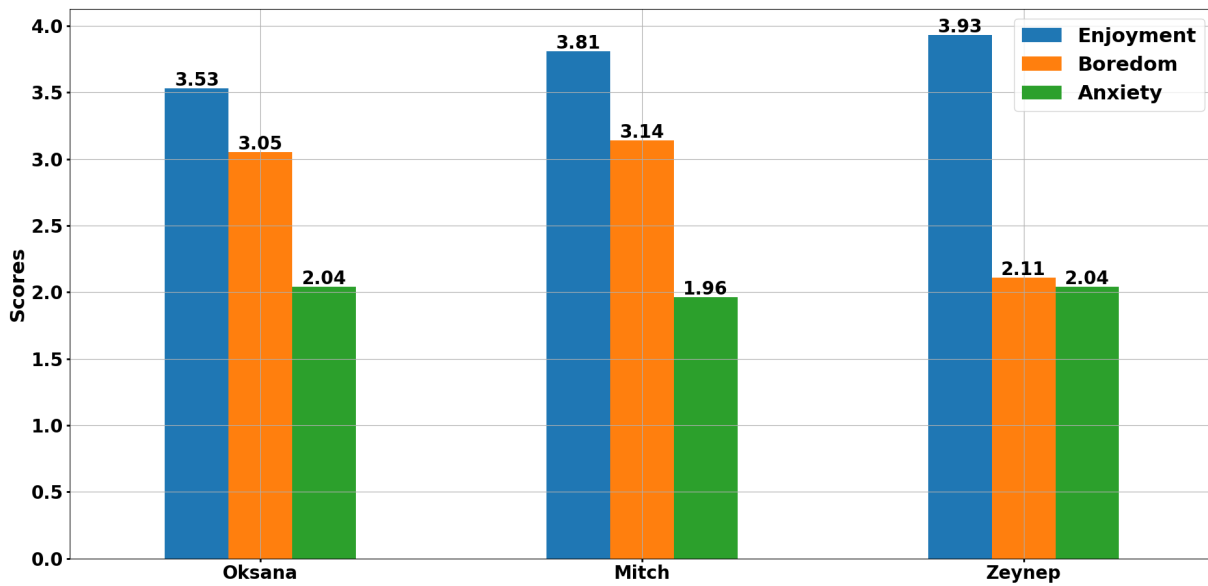


Figure 4.13 – Average scores of the three students for the three emotions across all sessions.

nents and classroom observations. During this session, the dominant emotion for him was enjoyment (which is rare for him): the enjoyment score is higher (3.77) than the boredom score (2.25) and the anxiety score (2). The boredom score during this session is much lower than in the other thirteen sessions, where it hovers around 3 or 3.5. The responses to the two open-ended questions are consistent with these scores, as Mitch wrote: "I felt good throughout the whole class. They were engaging, and their teaching style is better, and I like being able to read and write what I learn." He did not report any negative emotions.

The convergence between the SMs of Component 1 is almost total because 3 out of the 5 SMs correspond to Mitch's very active participation in the class, as shown by the observation of the class progression. He is very engaged in the different tasks and highly motivated. The SMs correspond to the moment when he stood up to read a very personal text about what the meaning of life is for him; during another SM, he worked with the teacher. For the third SM, Mitch did a group activity with Oksana.

The convergence between the observation of the class proceedings and Component 2 (EFE) is not complete because, out of 18 SMs in the EFE, only 7 correspond to a significant element concerning Mitch during the class. The convergence between the SMs of component 1 and those of component 2 is also semi-convergent, as 4 SMs are shared by both components, and only 2 out of the 4 SMs correspond to a significant element of

what Mitch did during the session.

One of the factors that might explain the intense enjoyment Mitch felt during this session is likely related to the fact that the class was led by two student interns from the Master's program in "Language Didactics" and that they did not use the Neurolinguistic Approach (NLA) method which Guillaume, the French FL teacher for the course, relies on. They conducted a class with the theme "Shitty Life." Mitch expressed the boredom he often felt due to the repetitive structure of NLA with its different phases. However, this boredom is mainly linked to the heterogeneity of the group of learners. Mitch, Oksana, and Zeynep were bored during most of the other sessions because they found the tasks too easy, too simple for them.

Oksana experienced very strong enjoyment during this same session 14. Her enjoyment score is 4, while the boredom (2.25) and anxiety (2) scores are lower. She responded to the open-ended questions by saying, "Today was an interesting class," and she did not report any negative emotions. Regarding component 1 (HR), the SMs converge with the observations of the class proceedings. Oksana was very active during these SMs and did most of the tasks with Mitch. Regarding component 2 (EFE), only 5 out of the 12 SMs correspond to significant elements of what she did in class. Six SMs overlap between components 1 and 2, indicating that these results are semi-convergent. Among these 6 SMs, only 4 reflect Oksana's active participation in the class. This semi-convergence means that Oksana's active participation in class is not always aligned with the SMs. It is clear that her internal emotional reactions are not always visible through external observations and that she is not always aware of them either. Therefore, there may be SMs that do not match either her actions in class or what she has reported in the responses to open-ended questions.

**4.4.4.0.2 Zeynep, a strong emotional engagement** The dominant emotion for Zeynep during the 14 sessions she attended in the FL French course was clearly enjoyment. Therefore, we selected session 15, where the results across different components were convergent. Her enjoyment score for session 15 (3.88) is thus higher than those for boredom (2.125) and anxiety (2.125). She responded very positively in French to the open-ended questions in the questionnaire: "I am happy to learn new things. We talked a lot today. " She did not report any negative emotions experienced.

Out of the 17 SMs from component 1 (HR), three correspond to significant elements noted in the observations of the course's progression concerning Zeynep. Indeed, at these

moments, she was highly engaged in oral interactions during the oral phase of the NLA. She was modeling a sentence related to the most important person to her when she was a child. She spoke about her grandfather, with whom she grew up. This was, therefore, a very emotionally intense moment for her. During the other two SMs, she was interacting with other learners, particularly with Oksana. The SMs from components 1 and 2 almost entirely converge, as 16 SMs from component 1 out of 17 are found in six SMs from component 2 (EFE). However, only 4 SMs among these 16 correspond to a significant activity by Zeynep during the class. This is why we can say that the results between these two components are semi-convergent.

We speculate that the strong enjoyment Zeynep experienced is likely due to the emotional contagion that emerged during her interactions with Oksana in the oral phase of the NLA, as well as the fact that she was talking about emotionally significant topics for her, which boosted her motivation and engagement in oral tasks.

## 4.5 Discussion

This study aimed to identify the emotions that students experience during a French FL course over the progression of an entire semester in a university setting using a multimodal approach.

Addressing  $RQ_1$ , which concerns the relationship between scores obtained from classroom observations and measures of physiological reactions (HR variation) and EFE, results varied significantly for each of the three students, confirming the findings in previous research [120, 118, 119, 90].

It is  $RQ_2$  that helps refine the initial results obtained from  $RQ_1$ . Indeed, the convergence between all data sources was generally strong. It seems that objective data sometimes capture emotional states of which learners are not always aware, and thus may not verbalize in self-perception reports. The classes lasted twice 1 hour and 30 minutes, which are long periods during which emotions fluctuate greatly. Students interviewed after the class may not remember everything that happened during those 3 hours. This is why the measurement of their EFE and HR variation, as well as classroom observations, were triangulated with their FLE and FLB scores and their responses to open-ended questions to gain a more holistic and precise view of the emotions experienced over such a long period.

The moments of convergence between all data and the three components of emotion

are primarily related, to significant moments in the class that the student remembers and that are often characterized by strong enjoyment linked to stimulating and collaborative group activities. These findings align with those of Nozawa et al. [93], which demonstrated positive emotional dynamics linked to pair-work.

It also seems that boredom, which is reflected in a decrease in student activity during class, corresponds to the presence of fewer SMs identified in HR variation and EFE. This supports the findings of Li et al. [150], which showed the very negative effects of boredom on language learners' motivation and their WTC.

Anxiety was detected in the AMs during oral activities performed in front of other group members during the oral phase. This confirms the finding of Gregersen et al. [90] and Dewaele and MacIntyre [114] that oral presentations highly anxiety-inducing in the FL classroom.

The three students in this study found the activities related to the NLA method boring and too repetitive because their level in French FL was much more advanced than that of the other students in the group. Agrawal et al. [110] and Li [151] explained that boredom occurs with repetitive, under-challenging tasks when learners feel that they do not learn anything new. However, we observed that during oral activities and when discussing emotionally intense topics (as in Unit 3 of the semester, which involved talking about events that marked our lives through anecdotes and unforgettable moments that shaped our past), the students were very engaged, and the results across the three components of the study converged during these moments.

The multimodal datastream allowed us to capture moments of convergence between the significant moments detected by physiological measurements and classroom observations, with learners' self-reports through questionnaires helping us interpret these results. These reflected episodes of positive emotional contagion between Oksana and Mitch during oral task activities and between Oksana and Zeynep [126, 127]. They were highly engaged together in the different tasks proposed by the teachers. This confirms the findings in Dewaele and MacIntyre [114] and Bielak and Mystkowska-Wiertelak [120] that working with a partner can be a powerful source of enjoyment and can create a sense of solidarity and empathy. The ability to communicate with peers and overcoming the fear of making mistakes in a positive environment is vital [119]. This positive emotional contagion experienced within a small group can lead to significant engagement in the task and even to a state of flow [93]. We could argue that where Nozawa et al. [93] caught evidence of brain synchronization between partners, we found evidence of "heart synchronization".



Positive emotions can help sustain motivation [152]. More specifically, FLE and levels of motivation flourish together [153, 154], while FLCA has the opposite relationship [153].

What matters to learners is that the activity allows them to collaborate with students they enjoy working with and enables them to express genuine emotions and feelings [120]. However, the activity must be sufficiently stimulating and challenging for them to engage with it. Any activity deemed too easy and repetitive risks being considered uninteresting, leading to a drop in engagement.

The multimodal method developed in the present study for tracking student engagement is both simple and easy to deploy, effectively addressing the complexities of classroom environments. By using low-cost sensors to capture HR signals and EFE, we aimed to create a dataset and to explore a pipeline capable of identifying significant moments based on unsupervised clustering at the student level. In other words, we focused on considering the individual context of each student by analyzing their physiological and behavioral variations in relation to their baseline emotional state, rather than grouping all students together and deriving the SM on each teaching session. This method is designed to be accessible and reproducible. For the analysis of HR signals, based on literature [146, 6], we selected time windows ranging from 60 to 210 seconds with a 50% overlap to capture the temporal variations in physiological responses. We employed unsupervised GMM clustering for each student individually, taking into account cultural and gender differences to reveal distinct emotions.

This approach not only improved the detection of SM but also enhanced the overall robustness of our engagement tracking system. The silhouette scores used to evaluate the cohesion and separation of clusters confirmed the effectiveness of our clustering strategy, particularly highlighting the importance of HR acceleration and speed parameters in distinguishing levels of engagement.

In contrast to the temporal processing of HR signals, the analysis of EFE focused on image-level recognition. Using a pre-trained ViT for emotion recognition, we clustered the resulting feature vectors with GMM to identify significant moments. This demonstrated the feasibility of using advanced neural network architectures for real-time engagement tracking in educational settings.

Overall, the proposed multimodal pipeline, combining HR signals and EFE, and cross-referenced with questionnaires and classroom observations, would provide a comprehensive framework for understanding students’ emotions . The experimental results highlight the potential of multimodal pattern recognition to enhance our understanding of the emotions

experienced by students during FL learning. Further research may consider lip emotion recognition models which can be employed as a substitute for EFE and have the advantage of maintaining participants' anonymity.

However, this study is not without its limitations. Firstly, longitudinal research in a naturalistic setting involving a large amount of multimodal data cannot be carried out on large groups and all longitudinal research inevitably suffers from attrition over time [134]. Moreover, the use of multiple tools increases the risk of technological malfunction which can lead to further loss of participants in the experiment [92]. However, small sample sizes are the norm in research inspired by Complex Dynamic Systems Theory [155] where the aim is to collect rich and detailed longitudinal data about unique individuals.

Secondly, utilizing HR signals allows us to track individual emotions continuously throughout the session, even when visual cues might be obscured due to students' head poses or obstructions by the teacher. Additionally, these two modalities provide distinctly different types of information: EFE reflect behavioural responses to emotional states, while HR signals represent physiological reactions that offer more objective measures for emotion recognition systems [140, 6]. However, EFE may be unreliable as individuals can consciously control these physical manifestations to hide their true emotions, a phenomenon known as social masking [145, 6]. Therefore, while HR monitoring is more intrusive, it is also more effective in accurately tracking student emotions. The integration of these two modalities has proven to be more reliable in determining emotional states.

The study also involved a small number of students because implementing such a protocol in the classroom presents various technical challenges and can sometimes disrupt the learner. We should point out that, due to the lack of data, we opted for unsupervised clustering to identify SM based on the two modalities stated previously. In a scenario where data is abundant and live tracking of student emotions is preferred, we would suggest isolating an initial session to establish a baseline for normal and SM for each student, followed by using a sliding time window in subsequent sessions for real-time emotion tracking. In a future study, it would also be interesting to show the videos to the students and discuss with them some time after the class (as is the case in idiodynamic studies), showing them the SM detected both from the HR signal and by the camera for the EFE.

## 4.6 Conclusion

The originality of the present study lies in its interdisciplinary approach and in the development of new tools to capture fleeting emotions. To the best of our knowledge, no previous study has integrated emotional facial expressions (EFE), heart rate (HR) signals, classroom observations, and self-reports—over an extensive duration and across multiple sessions, while considering the complexities inherent in real-world teaching environments. The novelty of this multimodal approach, combined with the absence of comparable studies using all modalities in similar contexts, precludes direct comparisons with state-of-the-art methods. Despite this limitation, our pilot study provide valuable insights into student’s emotions and offer new perspectives for future research in this area.

The decision to combine applied linguistics, language and culture didactics, artificial intelligence studies, computer engineering, automation, and signal processing allowed us to expand the range of dependent variables and to shed light on the complex dynamic system of language learners’ emotions at work in their classrooms. The rich stream of multimodal data collected from authentic interactions between three learners in one intact classroom over 16 sessions guaranteed ecological validity. The large quantity of data also allowed us to zoom in on episodes of particular interest, namely peaks and drops in the various emotions and especially moments of convergence between heart rates, facial expressions and self-reported data. These moments could be interpreted in light of the tasks being performed and the empathy with the partner.

To conclude, this pilot study provides researchers with new tools to capture the many manifestations of dynamic FL learner emotions and represents a move away from exclusive reliance on learners’ self-reports. The well-known phenomenon of emotional contagion could thus be observed in real-time across modalities and its sources could be identified.

# CONCLUSION AND PERSPECTIVES

---

## 5.1 Methodological Contributions to Multimodal Pattern Recognition

In this dissertation, we advanced state-of-the-art methods in multimodal pattern recognition for short utterance scenarios, spanning supervised to unsupervised methods applied to real-life applications. The contributions of this thesis can be categorized into three main fields: *From Unimodal to Multimodal Supervised Biometrics Recognition*, *Supervised Multimodal Deepfake Detection*, and *Unsupervised Multimodal Student Engagement Detection*.

Initially, we focused on unimodal biometrics identification using voice-based methods. Our study demonstrated the depth and invariance-scale importance of the sparse-based method, WST, in generating invariant features. This approach significantly enhanced the identification contrast between two speakers reading the same short sentence. In scenarios involving a large number of speakers, we introduced a novel integration of WST with x-vectors, resulting in an architecture with fewer trainable parameters yet competitive identification accuracy compared to state-of-the-art methods.

To further validate the efficacy of integrating multiple modalities for improved identification accuracy, we proposed a multimodal late fusion approach combining depth videos and audio signals. At the depth video level, we developed an efficient feature extraction method that reduced computational costs compared to existing methods. This late fusion method utilized 2D decomposition of spatiotemporal information, where each view was processed using a pre-trained ResNet-18 on ImageNet with an added self-attention module to balance features extracted from each view. For the audio modality, we employed pre-trained x-vectors architecture on VoxCeleb 1 & 2 to compress the audio signal into a single feature vector. This spatiotemporal architecture effectively extracted features from both modalities, overcoming challenges posed by short utterances and varying recording conditions. This research underscored the advantages of integrating depth information

---

with audio, advocating for a robust multimodal approach in speaker identification.

The emergence of deepfake methods poses a significant challenge to biometrics recognition systems. In Chapter 3, we fine-tuned our previous late fusion architecture by re-training only the fusion layers, not the entire architecture. We investigated the role of each view and modality in detecting deepfake generation methods' weak points, noting the superiority of the audio modality. Given the lack of interpretability and comprehensiveness of deep learning methods, we proposed a late fusion architecture at the decision level, using two hand-crafted methods to separately detect audio authenticity by focusing on the correlation between negative and positive samples. On the visual side, we employed a temporal anomalies detection method concentrating on the lips region, which exhibits significant variations across audio generation.

Building upon our findings in the biometrics identification field, we moved on to unsupervised multimodal student engagement detection in Chapter 4. We explored the effectiveness of multimodal data in tracking student engagement in classrooms using low-cost sensors to capture heart rate (HR) signals and facial expressions. By focusing on student-dependent clustering, we developed a reproducible dataset and a democratized pipeline for identifying abnormal moments of disengagement, considering cultural differences. Facial expressions analysis, using a pre-trained Vision Transformer (ViT) for emotion detection, complemented HR signal processing. Cross-referencing with expert observations validated this multimodal approach, demonstrating its potential for real-time monitoring of student engagement to enhance educational outcomes through timely interventions.

## 5.2 Future Perspectives

This doctoral research emphasizes the integration of different modalities in various applications, ranging from biometrics identification to student engagement. The proposed methodologies, involving shallow and deep learning methods, were simple yet effective. However, some limitations should be investigated in the future:

### 5.2.1 Biometrics Identification

Our multimodal and multi-view approach, which decomposes a video into three views and uses late fusion at the features level, could be made more robust by integrating a self-attention module at each extraction level. Additionally, our deepfake detection systems have primarily focused on GAN-based methods, but the rise of diffusion models and other

---

deepfake generation methods presents a potential threat. To address this challenge, one may develop and implement a more generalized architecture.

### **5.2.2 Student Engagement Detection**

The lack of sessions in our NeuroCam dataset can pose several limitations. Therefore, acquiring a larger number of sessions and using a supervised method based on HR signals and facial expressions could be more understandable and generalizable for each student. Additionally, we focused less on facial expressions and more on heart rate variations at the decision level due to the lower performance of deep learning models based on facial expressions and the fact that heart rate signals are more reliable in these situations. However, a more sophisticated model incorporating face-based emotion recognition could be developed in the future, giving equal importance to both RGB facial data and heart rate signals. Additionally, we only used SOTA handcrafted features developed for emotion recognition and adapted them for student engagement detection, but one may propose new features designed specifically for this purpose.

---

## 5.3 Publications

### Journal Articles

- **Abderrazzaq Moufidi**, David Rousseau, and Pejman Rasti, "Attention-Based Fusion of Ultrashort Voice Utterances and Depth Videos for Multimodal Person Identification", *Sensors*, 23.13 (2023), p. 5890.
- **Abderrazzaq Moufidi**, David Rousseau, and Pejman Rasti, "Toward Comprehensive Short Utterances Manipulations Detection in Videos", *Multimedia Tools and Applications* (2024): 1-14.
- Delphine Guedat-Bittighoffer, **Abderrazzaq Moufidi**, Jean-Marc Dewaele, David Rousseau, Hugo Voyneau, Pejman Rasti, "Heart rates, facial expressions and self-reports: A multimodal longitudinal approach of learners' emotions in the Foreign Language classroom" *Computers and Education*, (**Submitted**).

### Proceedings

- **Abderrazzaq Moufidi**, David Rousseau, and Pejman Rasti, "Wavelet Scattering Transform Depth Benefit: An Application for Speaker Identification," *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, Springer, 2022, pp. 97–106.
- **Abderrazzaq Moufidi**, David Rousseau, and Pejman Rasti, "Multimodal Deepfake Detection for Short Videos," *IMPROVE*, 2024, pp. 67–73.
- Delphine Guedat-Bittighoffer, **Abderrazzaq Moufidi**, David Rousseau, and Pejman Rasti, "Regards interdisciplinaires croisés sur les émotions éprouvées par trois apprenants de Français langue étrangère en contexte universitaire," *Colloque Langage et éMOTions*, 2024.

# WAVELET SCATTERING TRANSFORM

---

In Chapter 2, the concept of WST [32] method is to apply iteratively the wavelet transform  $\psi$  [75] and modulus as a non-linearity function (shift invariance) and an average Gaussian filter  $\phi$ .

An audio signal  $x$  is convoluted  $\star$  with dilated wavelets  $\psi_\lambda$  that are generated from a Morlet mother wavelet to produce a sparse representation (i.e: less coefficients):

$$\psi_{\lambda_i}(t) = \lambda_i \psi(\lambda_i t), \quad (\text{A.1})$$

where  $\lambda_i = 2^{\frac{j}{Q_i}}$ ,  $j \in \mathbb{Z}$  and  $Q_i$  is the quality factor or the number of wavelets per octave and  $i > 0$  refers to the order index.  $\psi_{\lambda_i}$  are centered at  $\lambda_i$  and have a bandwidth of  $\frac{\lambda_i}{Q_i}$  in the frequency domain.

At the zero-order, the scattering coefficients are given by  $S_0(x) = x \star \phi(t)$  (usually null for voice signals). At the  $i$ -th order ( $i \geq 1$ ), the signal is down-sampled by using the mean pooling at every scale, and the scattering coefficients are calculated as follows until a fixed maximum order  $m$

$$S_i x(t, \lambda_1, \dots, \lambda_i) = |||x \star \psi_{\lambda_1} | \star \dots | \star \psi_{\lambda_i} | \star \phi(t). \quad (\text{A.2})$$

These coefficients  $(S_i x)_{i=1, \dots, m}$  are often log-normalised (equation A.3) in order to reduce redundancy and increase translation invariance.

$$\begin{aligned} \tilde{S}_1 x(t, \lambda_1) &= \log \left( \frac{S_1 x(t, \lambda_1)}{|x \star \phi(t) + \epsilon|} \right) \\ &\quad \forall i \geq 2 \\ \tilde{S}_i x(t, \lambda_1, \dots, \lambda_i) &= \log \left( \frac{S_i x(t, \lambda_1, \dots, \lambda_i)}{S_{i-1} x(t, \lambda_1, \dots, \lambda_{i-1}) + \epsilon} \right) \end{aligned} \quad (\text{A.3})$$

where  $\epsilon$  is a silence detection threshold.





# TECHNICAL DESCRIPTION OF THE OXIMETER USED IN OUR EXPERIMENTS

---

In Chapter 4, to build our NeuroCam dataset for assessing student engagement in real-world scenarios, we captured RGB video of the classroom and tracked each student's heart rate using a wearable device, fingertip oximeter model PC-60FW depicted in Fig. B.1, it is from Pacific Medical Australia [156]. This fingertip oximeter is designed to measure pulse rate and the oxygen saturation ( $SpO_2$ ) through the user's finger. It is suitable for spot-checking  $SpO_2$  and pulse rate in both adult and pediatric patients, making it ideal for use in homes and medical clinics. It has the following characteristics.

## B.1 Fingertip Pulse Oximeter Features:

- Special splash proof and drop resistant design
- Display of Oxygen Saturation ( $SpO_2$ ), Pulse Rate (PR), Perfusion Index (PI), pulse bar and waveform
- Spot check and continuous pulse oximeter measurement modes
- Pulse rate analysis for spot check measurement
- Up to 12 groups  $SpO_2$  data storage
- Audible and visual alarm with low battery indication
- Automatic power on/off
- Four direction display
- Artefact removal and anti-motion
- Accurate and sensitive
- Wireless function (i.e., the oximeter is Bluetooth paired with ViHealth app on the phone)



Figure B.1 – Left: Oximeter used to record the heart rate beats of the students. Right: ViHealth application that receive the data from the oximeter.

## B.2 Fingertip Pulse Oximeter Specification:

### *Oxygen saturation ( $SpO_2$ )*

- Transducer: Dual-wavelength LED sensor
- Measuring range: 35 – 100%
- Measuring accuracy:  $\leq 2\%$  range from 70 – 100%

### *Pulse Rate*

- Measuring range: 30 – 240 bpm
- Measuring accuracy:  $\pm 2$ bpm or  $\pm 2\%$  (whichever is the greater)

### *Perfusion Index*

- Display range: 0-20%

### *Default Alarm Limit*

- $SpO_2$  low limit: 90%
- Pulse rate high limit: 120 bpm
- Pulse rate low limit: 50 bpm

---

### ***Measuring Mode***

- 30 seconds spot check mode : The measurement begins automatically once the finger is correctly placed in the finger clip. The process lasts for 30 seconds with a countdown displayed. At the end of the 30 seconds, the oxygen saturation ( $SpO_2$ ) and pulse rate (PR) readings freeze, and the pulse rhythm analysis is displayed. Once the finger is removed, the display clears, and the oximeter automatically shuts down.
- Continuous mode : The measurement starts automatically when the finger is inserted properly into the finger clip. The measurement continues indefinitely, with **real-time updates** of the oxygen saturation ( $SpO_2$ ) and **pulse rate (PR)** readings, until the finger is removed, at which point the oximeter shuts down automatically.

### ***Dimensions, Weight and Power***

- Dimensions:  $56 \times 34 \times 30$  mm ( $L \times W \times H$ )
- Weight: 52 g
- Power: 2  $\times$  AAA alkaline battery

# BIBLIOGRAPHY

---

- [1] Wenzhong Guo, Jianwen Wang, and Shiping Wang, « Deep multimodal representation learning: A survey », *in: Ieee Access* 7 (2019), pp. 63373–63394.
- [2] Jabeen Summaira, Xi Li, Amin Muhammad Shoib, Songyuan Li, and Jabbar Abdul, « Recent advances and trends in multimodal deep learning: A review », *in: arXiv preprint arXiv:2105.11087* (2021).
- [3] Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Jen-Chun Lin, Yu Tsao, Hsiu-Wen Chang, and Hsin-Min Wang, « Audio-visual speech enhancement using deep neural networks », *in: 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, IEEE, 2016, pp. 1–6.
- [4] Yi Xiao, Felipe Codevilla, Akhil Gurram, Onay Urfalioglu, and Antonio M López, « Multimodal end-to-end autonomous driving », *in: IEEE Transactions on Intelligent Transportation Systems* 23.1 (2020), pp. 537–547.
- [5] Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty, « MOMENTA: A multimodal framework for detecting harmful memes and their targets », *in: arXiv preprint arXiv:2109.05184* (2021).
- [6] Yujin Wu, « Multimodal emotion recognition from physiological signals and facial expressions », PhD thesis, Université de Lille, 2023.
- [7] Lin Shu, Yang Yu, Wenzhuo Chen, Haoqiang Hua, Qin Li, Jianxiu Jin, and Xiangmin Xu, « Wearable emotion recognition using heart rate data from a smart bracelet », *in: Sensors* 20.3 (2020), p. 718.
- [8] Miroslav Minović and Miloš Milovanović, « Real-time learning analytics in educational games », *in: Proceedings of the first international conference on technological ecosystem for enhancing multiculturalism*, 2013, pp. 245–251.

- 
- [9] Sinem Aslan, Nese Alyuz, Cagri Tanriover, Sinem E Mete, Eda Okur, Sidney K D’Mello, and Asli Arslan Esme, « Investigating the impact of a real-time, multimodal student engagement analytics technology in authentic classrooms », *in: Proceedings of the 2019 chi conference on human factors in computing systems*, 2019, pp. 1–12.
- [10] MGNSA Bueckle and Katy Börner, « Empowering instructors in learning management systems: Interactive heat map analytics dashboard », *in: Retrieved Nov 2 (2017)*, p. 2017.
- [11] John Hattie and Helen Timperley, « The power of feedback », *in: Review of educational research 77.1 (2007)*, pp. 81–112.
- [12] Roger A Federici and Einar M Skaalvik, « Students’ Perceptions of Emotional and Instrumental Teacher Support: Relations with Motivational and Emotional Responses. », *in: International education studies 7.1 (2014)*, pp. 21–36.
- [13] Anil K Jain, Lin Hong, and Yatin Kulkarni, « A multimodal biometric system using fingerprint, face and speech », *in: 2nd Int’l Conf. AVBPA*, vol. 10, 1999.
- [14] Arun Ross and Anil K Jain, « Multimodal biometrics: An overview », *in: 2004 12th European signal processing conference*, IEEE, 2004, pp. 1221–1224.
- [15] Robert W Frischholz and Ulrich Dieckmann, « Biold: a multimodal biometric identification system », *in: Computer 33.2 (2000)*, pp. 64–68.
- [16] Yifei Zhang, Désiré Sidibé, Olivier Morel, and Fabrice Mériaudeau, « Deep multimodal fusion for semantic image segmentation: A survey », *in: Image and Vision Computing 105 (2021)*, p. 104042.
- [17] Yipin Zhou and Ser-Nam Lim, « Joint audio-visual deepfake detection », *in: Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14800–14809.
- [18] Hafsa Ilyas, Ali Javed, and Khalid Mahmood Malik, « AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio-visual deepfakes detection », *in: Applied Soft Computing 136 (2023)*, p. 110124.
- [19] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, « X-Vectors: Robust DNN Embeddings for Speaker Recognition », *in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

- 
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, « Deep residual learning for image recognition », *in: Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [21] Shervin Minaee, Amirali Abdolrashidi, Hang Su, Mohammed Bennamoun, and David Zhang, « Biometrics recognition using deep learning: A survey », *in: Artificial Intelligence Review* (2023), pp. 1–49.
- [22] Yassin Kortli, Maher Jridi, Ayman Al Falou, and Mohamed Atri, « Face recognition systems: A survey », *in: Sensors* 20.2 (2020), p. 342.
- [23] Kien Nguyen, Hugo Proença, and Fernando Alonso-Fernandez, « Deep learning for iris recognition: A survey », *in: arXiv preprint arXiv:2210.05866* (2022).
- [24] Žiga Emeršič, Vitomir Štruc, and Peter Peer, « Ear recognition: More than a survey », *in: Neurocomputing* 255 (2017), pp. 26–39.
- [25] Debbrota P Chowdhury, Ritu Kumari, Sambit Bakshi, Manmath N Sahoo, and Abhijit Das, « Lip as biometric and beyond: a survey », *in: Multimedia Tools and Applications* 81.3 (2022), pp. 3831–3865.
- [26] Petar S. Aleksic, « Lip Movement Recognition », *in: Encyclopedia of Biometrics*, ed. by Stan Z. Li and Anil Jain, Boston, MA: Springer US, 2009, pp. 904–908.
- [27] Zhongxin Bai and Xiao-Lei Zhang, « Speaker recognition based on deep learning: An overview », *in: Neural Networks* 140 (2021), pp. 65–99.
- [28] Mirco Ravanelli and Yoshua Bengio, « Speaker Recognition from Raw Waveform with SincNet », *in: 2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021–1028.
- [29] Wajdi Ghezaiel, Luc Brun, and Olivier Lézoray, « Hybrid Network For End-To-End Text-Independent Speaker Identification », *in: 2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 2352–2359.
- [30] Hannah Muckenhirn, Mathew Magimai-Doss, and Sébastien Marcel, « On Learning Vocal Tract System Related Speaker Discriminative Information from Raw Signal Using CNNs. », *in: Interspeech*, 2018, pp. 1116–1120.
- [31] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, « Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification », *in: arXiv preprint arXiv:2005.07143* (2020).

- 
- [32] Joakim Andén and Stéphane Mallat, « Deep Scattering Spectrum », *in: IEEE Transactions on Signal Processing* 62.16 (2014), pp. 4114–4128.
- [33] Stéphane Mallat, « Group invariant scattering », *in: Communications on Pure and Applied Mathematics* 65.10 (2012), pp. 1331–1398.
- [34] John S Garofolo, « Timit acoustic phonetic continuous speech corpus », *in: Linguistic Data Consortium, 1993* (1993).
- [35] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu, « Human action recognition from various data modalities: A review », *in: IEEE transactions on pattern analysis and machine intelligence* (2022).
- [36] Lea Schönherr, Dennis Orth, Martin Heckmann, and Dorothea Kolossa, « Environmentally robust audio-visual speaker identification », *in: 2016 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2016, pp. 312–318.
- [37] Haoran Wei, Pranav Chopada, and Nasser Kehtarnavaz, « C-MHAD: Continuous Multimodal Human Action Dataset of Simultaneous Video and Inertial Sensing », *in: Sensors* 20.10 (2020).
- [38] Maycel Isaac Faraj and Josef Bigun, « Motion features from lip movement for person authentication », *in: 18th International Conference on Pattern Recognition (ICPR'06)*, vol. 3, IEEE, 2006, pp. 1059–1062.
- [39] Shtwai Alsubai, Monia Hamdi, Sayed Abdel-Khalek, Abdullah Alqahtani, Adel Binbusayyis, and Romany F Mansour, « Bald eagle search optimization with deep transfer learning enabled age-invariant face recognition model », *in: Image and Vision Computing* 126 (2022), p. 104545.
- [40] Dengshi Li, Yu Gao, Chenyi Zhu, Qianrui Wang, and Ruoxi Wang, « Improving Speech Recognition Performance in Noisy Environments by Enhancing Lip Reading Accuracy », *in: Sensors* 23.4 (2023), p. 2053.
- [41] Hsu Mon Lei Aung, Charnchai Pluempitiwiriyawej, Kazuhiko Hamamoto, and Somkiat Wangsiripitak, « Multimodal Biometrics Recognition Using a Deep Convolutional Neural Network with Transfer Learning in Surveillance Videos », *in: Computation* 10.7 (2022).
- [42] Krzysztof Wrobel, Rafal Doroz, Piotr Porwik, Jacek Naruniec, and Marek Kowalski, « Using a probabilistic neural network for lip-based biometric verification », *in: Engineering Applications of Artificial Intelligence* 64 (2017), pp. 112–127.



- 
- [43] Guido Borghi, Stefano Pini, Roberto Vezzani, and Rita Cucchiara, « Driver face verification with depth maps », *in: Sensors* 19.15 (2019), p. 3361.
- [44] Abderrazzaq Moufidi, David Rousseau, and Pejman Rasti, « Wavelet Scattering Transform Depth Benefit, An Application for Speaker Identification », *in: IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, Springer, 2022, pp. 97–106.
- [45] Abderrazzaq Moufidi, David Rousseau, and Pejman Rasti, « Attention-Based Fusion of Ultrashort Voice Utterances and Depth Videos for Multimodal Person Identification », *in: Sensors* 23.13 (2023), p. 5890.
- [46] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al., « SpeechBrain: A general-purpose speech toolkit », *in: arXiv preprint arXiv:2106.04624* (2021).
- [47] <https://huggingface.co/speechbrain/spkrec-xvect-voxceleb>.
- [48] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, « Voxceleb2: Deep speaker recognition », *in: arXiv preprint arXiv:1806.05622* (2018).
- [49] Md Rashidul Hasan, Mustafa Jamil, MGRMS Rahman, et al., « Speaker identification using mel frequency cepstral coefficients », *in: variations* 1.4 (2004), pp. 565–568.
- [50] Naomi Harte and Eoin Gillen, « TCD-TIMIT: An audio-visual corpus of continuous speech », *in: IEEE Transactions on Multimedia* 17.5 (2015), pp. 603–615.
- [51] Vijeta Sharma, Manjari Gupta, Ajai Kumar, and Deepti Mishra, « Video processing using deep learning techniques: A systematic literature review », *in: IEEE Access* 9 (2021), pp. 139489–139507.
- [52] Faisal Khan, Shahid Hussain, Shubhajit Basak, Joseph Lemley, and Peter Corcoran, « An efficient encoder–decoder model for portrait depth estimation from single images trained on pixel-accurate synthetic data », *in: Neural Networks* 142 (2021), pp. 479–491.
- [53] Guido Borghi, Marco Venturelli, Roberto Vezzani, and Rita Cucchiara, « Poseidon: Face-from-depth for driver pose estimation », *in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2017, pp. 5494–5503.

- 
- [54] Rui Min, Neslihan Kose, and Jean-Luc Dugelay, « KinectFaceDB: A Kinect Database for Face Recognition », *in: Systems, Man, and Cybernetics: Systems, IEEE Transactions on* 44.11 (Nov. 2014), pp. 1534–1548.
- [55] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool, « Random Forests for Real Time 3D Face Analysis », *in: Int. J. Comput. Vision* 101.3 (Feb. 2013), pp. 437–458.
- [56] Tao Zhang, « Deepfake generation and detection, a survey », *in: Multimedia Tools and Applications* 81.5 (2022), pp. 6259–6276.
- [57] Tzu-hsien Huang, Jheng-hao Lin, and Hung-yi Lee, « How far are we from robust voice conversion: A survey », *in: 2021 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2021, pp. 514–521.
- [58] Venkatachalam Kandasamy, Štěpán Hubálovský, and Pavel Trojovský, « Deep fake detection using a sparse auto encoder with a graph capsule dual graph CNN », *in: PeerJ Computer Science* 8 (2022), e953.
- [59] Hasam Khalid, Minha Kim, Shahroz Tariq, and Simon S Woo, « Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors », *in: Proceedings of the 1st workshop on synthetic multimedia-audiovisual deepfake generation and detection*, 2021, pp. 7–15.
- [60] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang, « Wild-deepfake: A challenging real-world dataset for deepfake detection », *in: Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 2382–2390.
- [61] Davide Salvi, Honggu Liu, Sara Mandelli, Paolo Bestagini, Wenbo Zhou, Weiming Zhang, and Stefano Tubaro, « A Robust Approach to Multimodal Deepfake Detection », *in: Journal of Imaging* 9.6 (2023), p. 122.
- [62] Jun Ling, Xu Tan, Liyang Chen, Runnan Li, Yuchao Zhang, Sheng Zhao, and Li Song, *StableFace: Analyzing and Improving Motion Stability for Talking Face Generation*, 2022, arXiv: 2208.13717 [cs.CV].
- [63] Deepak Dagar and Dinesh Kumar Vishwakarma, « A literature review and perspectives in deepfakes: generation, detection, and applications », *in: International journal of multimedia information retrieval* 11.3 (2022), pp. 219–289.

- 
- [64] Ruben Tolosana, Sergio Romero-Tapiador, Ruben Vera-Rodriguez, Ester Gonzalez-Sosa, and Julian Fierrez, « DeepFakes detection across generations: Analysis of facial regions, fusion, and performance evaluation », *in: Engineering Applications of Artificial Intelligence* 110 (2022), p. 104673.
- [65] Vrizlynn LL Thing, « Deepfake Detection with Deep Learning: Convolutional Neural Networks versus Transformers », *in: arXiv e-prints* (2023), arXiv-2304.
- [66] Ziyue Jiang, Jinglin Liu, Yi Ren, Jinzheng He, Chen Zhang, Zhenhui Ye, Pengfei Wei, Chunfeng Wang, Xiang Yin, Zejun Ma, et al., « Mega-TTS 2: Zero-Shot Text-to-Speech with Arbitrary Length Speech Prompts », *in: arXiv preprint arXiv:2307.07218* (2023).
- [67] Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik, « Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward », *in: Applied intelligence* 53.4 (2023), pp. 3974–4026.
- [68] Jia Wen Seow, Mei Kuan Lim, Raphael CW Phan, and Joseph K Liu, « A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities », *in: Neurocomputing* 513 (2022), pp. 351–371.
- [69] Abderrazzaq Moufidi, David Rousseau, and Pejman Rasti, « Toward Comprehensive Short Utterances Manipulations Detection in Videos », *in: Multimedia Tools and Applications - Under review* (2024).
- [70] Abderrazzaq Moufidi, David Rousseau, and Pejman Rasti, « Multimodal Deepfake Detection for Short Videos. », *in: IMPROVE, 2024*, pp. 67–73.
- [71] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen, « Mesonet: a compact facial video forgery detection network », *in: 2018 IEEE international workshop on information forensics and security (WIFS)*, IEEE, 2018, pp. 1–7.
- [72] Alessandro Pianese, Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva, *Deepfake audio detection by speaker verification*, 2022, arXiv: 2209.14098 [cs.SD].
- [73] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann, « MediaPipe: A Framework for Perceiving and Processing Reality », *in: Third Workshop on*

- 
- Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [74] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman, « Synthesizing obama: learning lip sync from audio », *in: ACM Transactions on Graphics (ToG)* 36.4 (2017), pp. 1–13.
- [75] Mallat Stephane, *A wavelet tour of signal processing*, 1999.
- [76] 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*, vol. 34, 2, Wiley Online Library, 2015, pp. 193–204.
- [77] Luisa Verdoliva, « Media forensics and deepfakes: an overview », *in: IEEE Journal of Selected Topics in Signal Processing* 14.5 (2020), pp. 910–932.
- [78] Ross Cutler and Larry Davis, « Look who’s talking: Speaker detection using video and audio correlation », *in: 2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)*, vol. 3, IEEE, 2000, pp. 1589–1592.
- [79] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, « Imagenet: A large-scale hierarchical image database », *in: 2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.
- [80] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo, « FakeAVCeleb: A novel audio-video multimodal deepfake dataset », *in: arXiv preprint arXiv:2108.05080* (2021).
- [81] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis, « Fast face-swap using convolutional neural networks », *in: Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3677–3685.
- [82] Yuval Nirkin, Yosi Keller, and Tal Hassner, « Fsgan: Subject agnostic face swapping and reenactment », *in: Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7184–7193.
- [83] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar, « A lip sync expert is all you need for speech to lip generation in the wild », *in: Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 484–492.

- 
- [84] 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 », *in: Advances in neural information processing systems* 31 (2018).
- [85] C Sanderson, *Vidimit audio-video dataset*, 2001.
- [86] Pavel Korshunov and Sébastien Marcel, « Deepfakes: a new threat to face recognition? assessment and detection », *in: arXiv preprint arXiv:1812.08685* (2018).
- [87] Conrad Sanderson and Brian C Lovell, « Multi-region probabilistic histograms for robust and scalable identity inference », *in: Advances in Biometrics: Third International Conference, ICB 2009, Alghero, Italy, June 2-5, 2009. Proceedings 3*, Springer, 2009, pp. 199–208.
- [88] Qiangfu Yu, « A review of foreign language learners' emotions », *in: Frontiers in Psychology* 12 (2022), p. 827104.
- [89] Jérémie Sauvage and Déborah Nourrit, « Humanités numériques et pensée complexe », *in: LHUMAINE 1.1* (2022).
- [90] Tammy Gregersen, Peter D MacIntyre, and Mario D Meza, « The motion of emotion: Idiodynamic case studies of learners' foreign language anxiety », *in: The Modern Language Journal* 98.2 (2014), pp. 574–588.
- [91] Peter MacIntyre and Nathan Ducker, « The idiodynamic method: A practical guide for researchers », *in: Research Methods in Applied Linguistics* 1.2 (2022), p. 100007.
- [92] Katie Hoemann, Jolie B Wormwood, Lisa Feldman Barrett, and Karen S Quigley, « Multimodal, idiographic ambulatory sensing will transform our understanding of emotion », *in: Affective Science* 4.3 (2023), pp. 480–486.
- [93] Takayuki Nozawa, Mutsumi Kondo, Reiko Yamamoto, Hyeonjeong Jeong, Shigeyuki Ikeda, Kohei Sakaki, Yoshihiro Miyake, Yasushige Ishikawa, and Ryuta Kawashima, « Prefrontal inter-brain synchronization reflects convergence and divergence of flow dynamics in collaborative learning: A pilot study », *in: Frontiers in Neuroergonomics* 2 (2021), p. 686596.
- [94] Güray Tonguç and Betül Ozaydın Ozkara, « Automatic recognition of student emotions from facial expressions during a lecture », *in: Computers & Education* 148 (2020), p. 103797.

- 
- [95] Lisa Feldman Barrett, *How emotions are made: The secret life of the brain*, Pan Macmillan, 2017.
- [96] Pamela Gobin, Véronique Baltazart, and Aurélie Simoës-Perlant, « Les émotions », *in: Emotions et Apprentissages* (2021), pp. 19–49.
- [97] Richard S Lazarus, *Emotion and adaptation*, vol. 557, Oxford University Press, 1991.
- [98] Kees De Bot, Wander Lowie, and Marjolijn Verspoor, « A dynamic systems theory approach to second language acquisition », *in: Bilingualism: Language and cognition* 10.1 (2007), pp. 7–21.
- [99] Sidney D’Mello, Andrew Olney, Claire Williams, and Patrick Hays, « Gaze tutor: A gaze-reactive intelligent tutoring system », *in: International Journal of human-computer studies* 70.5 (2012), pp. 377–398.
- [100] Paul Ekman, « An argument for basic emotions », *in: Cognition & emotion* 6.3-4 (1992), pp. 169–200.
- [101] Maria Gendron, Lisa Feldman Barrett, and Laurent Bury, « La perception des émotions: une synchronie conceptuelle », *in: Sensibilités 2* (2018), pp. 70–83.
- [102] Elaine K Horwitz, Michael B Horwitz, and Joann Cope, « Foreign language classroom anxiety », *in: The Modern language journal* 70.2 (1986), pp. 125–132.
- [103] Peter MacIntyre and Tammy Gregersen, « Emotions that facilitate language learning: The positive-broadening power of the imagination », *in: Studies in second language learning and teaching* 2.2 (2012), pp. 193–213.
- [104] Oriane Petiot and Jérôme Visioli, *Les émotions en contexte scolaire*, De Boeck Supérieur, 2022.
- [105] Elaine K Horwitz, « On the misreading of Horwitz, Horwitz, and Cope (1986) and the need to balance anxiety research and the experiences of anxious language learners », *in: New insights into language anxiety: Theory, research and educational implications* 31 (2017), p. 47.
- [106] Eloise Botes, Jean-Marc Dewaele, and Samuel Greiff, « The foreign language classroom anxiety scale and academic achievement: An overview of the prevailing literature and a meta-analysis », *in: Journal for the Psychology of Language Learning* 2.1 (2020), pp. 26–56.

- 
- [107] Jim King, « Silence in the second language classrooms of Japanese universities », *in: Applied linguistics* 34.3 (2013), pp. 325–343.
- [108] M Pawlak, M Kruk, and J Zawodniak, *Teachers reflecting on boredom in the language classroom*, 2024.
- [109] Chengchen Li, Jean-Marc Dewaele, and Yanhong Hu, « Foreign language learning boredom: Conceptualization and measurement », *in: Applied Linguistics Review* 14.2 (2023), pp. 223–249.
- [110] Mayank Agrawal, Marcelo G Mattar, Jonathan D Cohen, and Nathaniel D Daw, « The temporal dynamics of opportunity costs: A normative account of cognitive fatigue and boredom. », *in: Psychological review* 129.3 (2022), p. 564.
- [111] Jean-Marc Dewaele and Chengchen Li, « Teacher enthusiasm and students' social-behavioral learning engagement: The mediating role of student enjoyment and boredom in Chinese EFL classes », *in: Language Teaching Research* 25.6 (2021), pp. 922–945.
- [112] Chengchen Li, « Foreign language learning boredom and enjoyment: The effects of learner variables and teacher variables », *in: Language Teaching Research* (2022), p. 13621688221090324.
- [113] C Li, W Li, and G Jiang, « Emotions in second language learning: Retrospect and prospect », *in: Modern Foreign Languages* 1 (2024), pp. 63–75.
- [114] Jean-Marc Dewaele and Peter D MacIntyre, « The two faces of Janus? Anxiety and enjoyment in the foreign language classroom », *in: Studies in second language learning and teaching* 4.2 (2014), pp. 237–274.
- [115] Mihaly Csikszentmihalyi, *Flow*, Munksgaard København, 1991.
- [116] Elouise Botes, Jean–Marc Dewaele, and Samuel Greiff, « The development and validation of the short form of the foreign language enjoyment scale », *in: The Modern Language Journal* 105.4 (2021), pp. 858–876.
- [117] Carmen Boudreau, Peter MacIntyre, and Jean-Marc Dewaele, « Enjoyment and anxiety in second language communication: An idiodynamic approach », *in: Studies in Second Language Learning and Teaching* 8.1 (2018), pp. 149–170.
- [118] Majid Elahi Shirvan and Nahid Talebzadeh, « Exploring the fluctuations of foreign language enjoyment in conversation: an idiodynamic perspective », *in: Journal of Intercultural Communication Research* 47.1 (2018), pp. 21–37.

- 
- [119] Majid Elahi Shirvan, Tahereh Taherian, and Elham Yazdanmehr, « The dynamics of foreign language enjoyment: An ecological momentary assessment », *in: Frontiers in Psychology* 11 (2020), p. 1391.
- [120] Jakub Bielak and Anna Mystkowska-Wiertelak, « Emotions and emotion regulation in L2 classroom speaking tasks: A mixed-methods study combining the idiodynamic and quantitative perspectives », *in: The Modern Language Journal* (2024).
- [121] Jean-Marc Dewaele and Liana Maria Pavelescu, « The relationship between incommensurable emotions and willingness to communicate in English as a foreign language: a multiple case study », *in: Innovation in Language Learning and Teaching* 15.1 (2021), pp. 66–80.
- [122] Elouise Botes, Jean-Marc Dewaele, and Samuel Greiff, « Taking stock: A meta-analysis of the effects of foreign language enjoyment », *in: Studies in Second Language Learning and Teaching* 12.2 (2022), pp. 205–232.
- [123] Jean-Marc Dewaele and Peter MacIntyre, « Do flow, enjoyment and anxiety emerge equally in English foreign language classrooms as in other foreign language classrooms? », *in: Revista Brasileira de Linguística Aplicada* 22 (2022), pp. 156–180.
- [124] Freerkien Waninge, Zoltán Dörnyei, and Kees De Bot, « Motivational dynamics in language learning: Change, stability, and context », *in: The Modern Language Journal* 98.3 (2014), pp. 704–723.
- [125] Chengchen Li and Jean-Marc Dewaele, « How classroom environment and general grit predict foreign language classroom anxiety of Chinese EFL students », *in: Journal for the Psychology of Language Learning* 3.2 (2021), pp. 86–98.
- [126] Sharona Moskowitz and Jean-Marc Dewaele, « Is teacher happiness contagious? A study of the link between perceptions of language teacher happiness and student attitudes », *in: Innovation in Language Learning and Teaching* 15.2 (2021), pp. 117–130.
- [127] Nahid Talebzadeh, Majid Elahi Shirvan, and Gholam Hassan Khajavy, « Dynamics and mechanisms of foreign language enjoyment contagion », *in: Innovation in Language Learning and Teaching* 14.5 (2020), pp. 399–420.



- 
- [128] Chengchen Li and J Dewaele, « Understanding, measuring, and differentiating task enjoyment from foreign language enjoyment », *in: Individual differences and task-based language teaching* (2024), pp. 87–114.
- [129] Jonghwa Kim and Elisabeth André, « Emotion recognition based on physiological changes in music listening », *in: IEEE transactions on pattern analysis and machine intelligence* 30.12 (2008), pp. 2067–2083.
- [130] Michael W Wukitsch, Michael T Petterson, David R Tobler, and Jonas A Pologe, « Pulse oximetry: analysis of theory, technology, and practice », *in: Journal of clinical monitoring* 4 (1988), pp. 290–301.
- [131] Y Iyriboz, S Powers, J Morrow, D Ayers, and G Landry, « Accuracy of pulse oximeters in estimating heart rate at rest and during exercise. », *in: British journal of sports medicine* 25.3 (1991), pp. 162–164.
- [132] Elouise Botes, Lindie Van der Westhuizen, Jean-Marc Dewaele, Peter MacIntyre, and Samuel Greiff, « Validating the short-form foreign language classroom anxiety scale », *in: Applied Linguistics* 43.5 (2022), pp. 1006–1033.
- [133] Jean-Marc Dewaele, « Collecting and Analyzing Individual Differences Data in Foreign Language Learning », *in: Current Approaches in Second Language Acquisition Research: A Practical Guide* (2023), pp. 215–232.
- [134] Zoltan Dornyei, *Research methods in applied linguistics*, Oxford university press, 2007.
- [135] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel, « Backpropagation applied to handwritten zip code recognition », *in: Neural computation* 1.4 (1989), pp. 541–551.
- [136] Mohamad Alansari, Oussama Abdul Hay, Sajid Javed, Abdulhadi Shoufan, Yahya Zweiri, and Naoufel Werghi, « Ghostfacenets: Lightweight face recognition model from cheap operations », *in: IEEE Access* 11 (2023), pp. 35429–35446.
- [137] Sefik Serengil and Alper Ozpinar, « A Benchmark of Facial Recognition Pipelines and Co-Usability Performances of Modules », *in: Bilisim Teknolojileri Dergisi* 17.2 (2024), pp. 95–107.

- 
- [138] Sefik Ilkin Serengil and Alper Ozpinar, « HyperExtended LightFace: A Facial Attribute Analysis Framework », *in: 2021 International Conference on Engineering and Emerging Technologies (ICEET)*, IEEE, 2021, pp. 1–4.
- [139] Sefik Ilkin Serengil and Alper Ozpinar, « LightFace: A Hybrid Deep Face Recognition Framework », *in: 2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, IEEE, 2020, pp. 23–27.
- [140] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al., « A systematic review on affective computing: Emotion models, databases, and recent advances », *in: Information Fusion* 83 (2022), pp. 19–52.
- [141] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain, « A review of affective computing: From unimodal analysis to multimodal fusion », *in: Information fusion* 37 (2017), pp. 98–125.
- [142] Alexey Dosovitskiy, « An image is worth 16x16 words: Transformers for image recognition at scale », *in: arXiv preprint arXiv:2010.11929* (2020).
- [143] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al., « Challenges in representation learning: A report on three machine learning contests », *in: Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*, Springer, 2013, pp. 117–124.
- [144] [https://huggingface.co/dima806/facial\\_emotions\\_image\\_detection](https://huggingface.co/dima806/facial_emotions_image_detection).
- [145] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang, « A review of emotion recognition using physiological signals », *in: Sensors* 18.7 (2018), p. 2074.
- [146] Sylvia D Kreibig, « Autonomic nervous system activity in emotion: A review », *in: Biological psychology* 84.3 (2010), pp. 394–421.
- [147] Shan Li and Weihong Deng, « Deep facial expression recognition: A survey », *in: IEEE transactions on affective computing* 13.3 (2020), pp. 1195–1215.
- [148] Peter J Rousseeuw, « Silhouettes: a graphical aid to the interpretation and validation of cluster analysis », *in: Journal of computational and applied mathematics* 20 (1987), pp. 53–65.

- 
- [149] Kazuya Mera and Takumi Ichimura, « Emotion analyzing method using physiological state », *in: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, Springer, 2004, pp. 195–201.
- [150] Chengchen Li, Jean-Marc Dewaele, Mirosław Pawlak, and Mariusz Kruk, « Classroom environment and willingness to communicate in English: The mediating role of emotions experienced by university students in China », *in: Language Teaching Research* (2022), p. 13621688221111623.
- [151] Chengchen Li, « A control–value theory approach to boredom in English classes among university students in China », *in: The Modern Language Journal* 105.1 (2021), pp. 317–334.
- [152] Peter D MacIntyre, Jean-Marc Dewaele, Nicole Macmillan, and Chengchen Li, « The emotional underpinnings of Gardner’s attitudes and motivation test battery », *in: Contemporary language motivation theory* 60 (2019), pp. 57–79.
- [153] Jean-Marc Dewaele and Anna Lia Proietti Ergün, « How different are the relations between enjoyment, anxiety, attitudes/motivation and course marks in pupils’ Italian and English as foreign languages? », *in: Journal of the European Second Language Association* 4.1 (2020), pp. 45–57.
- [154] Haihua Wang, Lin Xu, and Jiaxin Li, « Connecting foreign language enjoyment and English proficiency levels: The mediating role of L2 motivation », *in: Frontiers in psychology* 14 (2023), p. 1054657.
- [155] Phil Hiver, Ali H Al-Hoorie, and Reid Evans, « Complex dynamic systems theory in language learning: A scoping review of 25 years of research », *in: Studies in Second Language Acquisition* 44.4 (2022), pp. 913–941.
- [156] <https://pacificmedical.com.au/product/fingertip-pulse-oximeter/>.



**Titre :** Intégration multimodale basée sur l'apprentissage automatique pour l'identification biométrique et la détection d'engagement à partir des durées courtes

**Mot clés :** Multimodalité, Wavelet Scattering Transform, x-vectors, Détection de Deepfake, Détection d'engagement des élèves, Rythme Cardiaque

**Résumé :** Le progrès rapide et la démocratisation de la technologie ont conduit à l'abondance des capteurs. Par conséquent, l'intégration de ces diverses modalités pourrait présenter un avantage considérable pour de nombreuses applications dans la vie réelle, telles que la reconnaissance biométrique ou la détection d'engagement des élèves. Dans le domaine de la multimodalité, les chercheurs ont établi des architectures variées de fusion, allant des approches de fusion précoce, hybride et tardive. Cependant, ces architectures peuvent avoir des limites en ce qui concerne des signaux temporels d'une durée courte, ce qui nécessite un changement de paradigme vers le développement de techniques d'apprentissage automatique multimodales qui promettent une précision et une efficacité pour l'analyse de ces données courtes. Dans cette thèse, nous nous appuyons sur l'intégration de la multimodalité pour relever les défis précédents, allant de l'identification biométrique supervisée à la détection non supervisée de l'engagement des étudiants. La première contribution de ce doctorat porte sur l'intégration de la Wavelet Scattering Transform à plusieurs couches avec une architecture profonde appelée x-vectors, grâce à laquelle nous avons amélioré la performance de l'identification du locuteur dans des scénarios impliquant des énoncés courts tout en réduisant le nombre de paramètres nécessaires à l'entraînement. En s'appuyant sur les avantages de la multimodalité, on a proposé une architecture de fusion tardive combinant des vidéos de la profondeur des lèvres et des

signaux audios a permis d'améliorer la précision de l'identification dans le cas d'énoncés courts, en utilisant des méthodes efficaces et moins coûteuses pour extraire des caractéristiques spatio-temporelles. Dans le domaine des défis biométriques, il y a la menace de l'émergence des "deepfakes". Ainsi, nous nous sommes concentrés sur l'élaboration d'une méthode de détection des "deepfakes" basée sur des méthodes mathématiques compréhensibles et sur une version finement ajustée de notre précédente fusion tardive appliquée aux vidéos RVB des lèvres et aux audios. En utilisant des méthodes de détection d'anomalies conçues spécifiquement pour les modalités audio et visuelles, l'étude a démontré des capacités de détection robustes dans divers ensembles de données et conditions, soulignant l'importance des approches multimodales pour contrer l'évolution des techniques de deepfake. S'étendant aux contextes éducatifs, la thèse explore la détection multimodale de l'engagement des étudiants dans une classe. En utilisant des capteurs abordables pour acquérir les signaux du rythme cardiaque et les expressions faciales, l'étude a développé un ensemble de données reproductibles et un plan pour identifier des moments significatifs, tout en tenant compte des nuances culturelles. L'analyse des expressions faciales à l'aide de Vision Transformer (ViT) fusionnée avec le traitement des signaux de fréquence cardiaque, validée par des observations d'experts, a mis en évidence le potentiel du suivi des élèves afin d'améliorer la qualité d'enseignement.

---

**Title:** Machine Learning-Based Multimodal integration for Short Utterance-Based Biometrics Identification and Engagement Detection

**Keywords:** Multimodality, Wavelet Scattering Transform, x-vectors, Deepfake Detection, Student Engagement Detection, Heart Rate

**Abstract:**

The rapid advancement and democratization of technology have led to an abundance of sensors. Consequently, the integration of these diverse modalities presents an advantage for numerous real-life applications, such as biometrics recognition and engagement detection. In the field of multimodality, researchers have developed various fusion architectures, ranging from early, hybrid, to late fusion approaches. However, these architectures may have limitations involving short utterances and brief video segments, necessitating a paradigm shift towards the development of multimodal machine learning techniques that promise precision and efficiency for short-duration data analysis. In this thesis, we lean on integration of multimodality to tackle these previous challenges ranging from supervised biometrics identification to unsupervised student engagement detection. This PhD began with the first contribution on the integration of multiscale Wavelet Scattering Transform with x-vectors architecture, through which we enhanced the accuracy of speaker identification in scenarios involving short utterances. Going through multimodality benefits, a late fusion architecture combining lips depth videos and audio signals further improved

identification accuracy under short utterances, utilizing an effective and less computational methods to extract spatiotemporal features. In the realm of biometrics challenges, there is the threat emergence of deepfakes. Therefore, we focalized on elaborating a deepfake detection methods based on, shallow learning and a fine-tuned architecture of our previous late fusion architecture applied on RGB lips videos and audios. By employing hand-crafted anomaly detection methods for both audio and visual modalities, the study demonstrated robust detection capabilities across various datasets and conditions, emphasizing the importance of multimodal approaches in countering evolving deepfake techniques. Expanding to educational contexts, the dissertation explores multimodal student engagement detection in classrooms. Using low-cost sensors to capture Heart Rate signals and facial expressions, the study developed a reproducible dataset and pipeline for identifying significant moments, accounting for cultural nuances. The analysis of facial expressions using Vision Transformer (ViT) fused with heart rate signal processing, validated through expert observations, showcased the potential for real-time monitoring to enhance educational outcomes through timely interventions.