First Delivery Advanced HCI

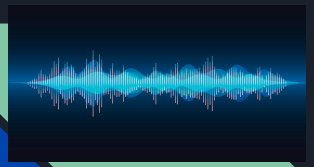
Giuseppe Grisolia, Michele Minniti

Well being digital education

- Inclusive learning
- Mental health issues
- Digital stress
- Augmentative and alternative communication (AAC)

Problem we address

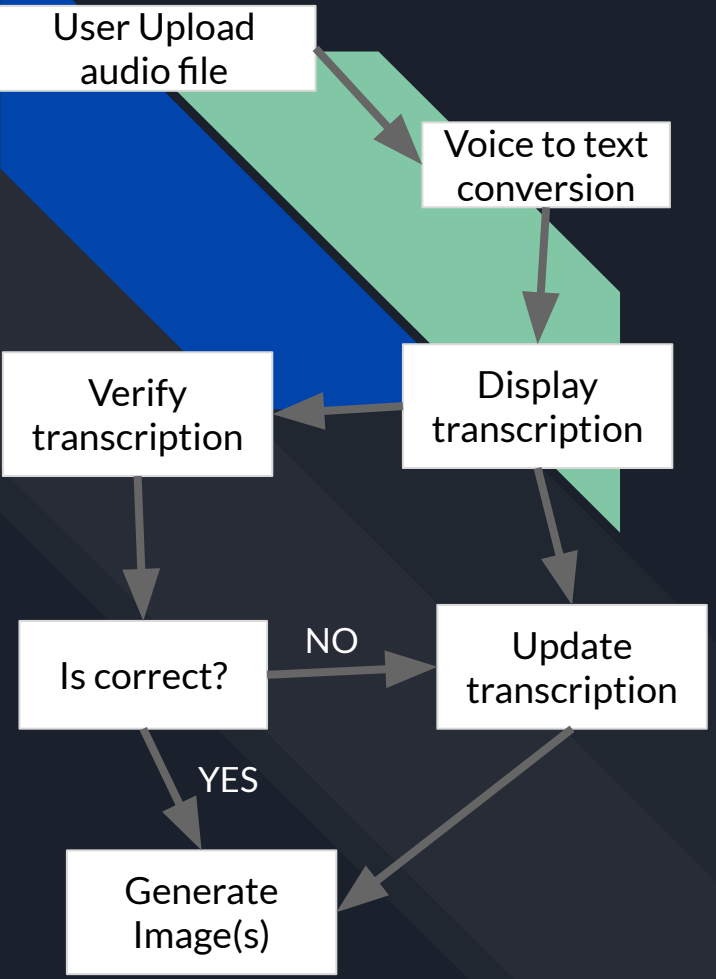
We want to address the problem of augmentative learning for children in the autistic spectrum or with ADHD disturb



“albero”

Speech-to-Image Generation with Stable Diffusion Model

- Dynamic speech-to-text-to-image synthesis
- Improving accessibility also for students with mental illness
- Help with language barriers
- Enhance teaching-learning experience



Speech-to-Image Generation with Stable Diffusion Model 2

- **1° PHASE: TRANSCRIPTION OF AUDIO FILE IN TEXTUAL DATA:**
  -The audio file is loaded.
  -System transcribe the audio content with the google's: "speech recognition service" wich address any potential errors. (unknown value or request errors)
  ->FINAL RESULT: text input

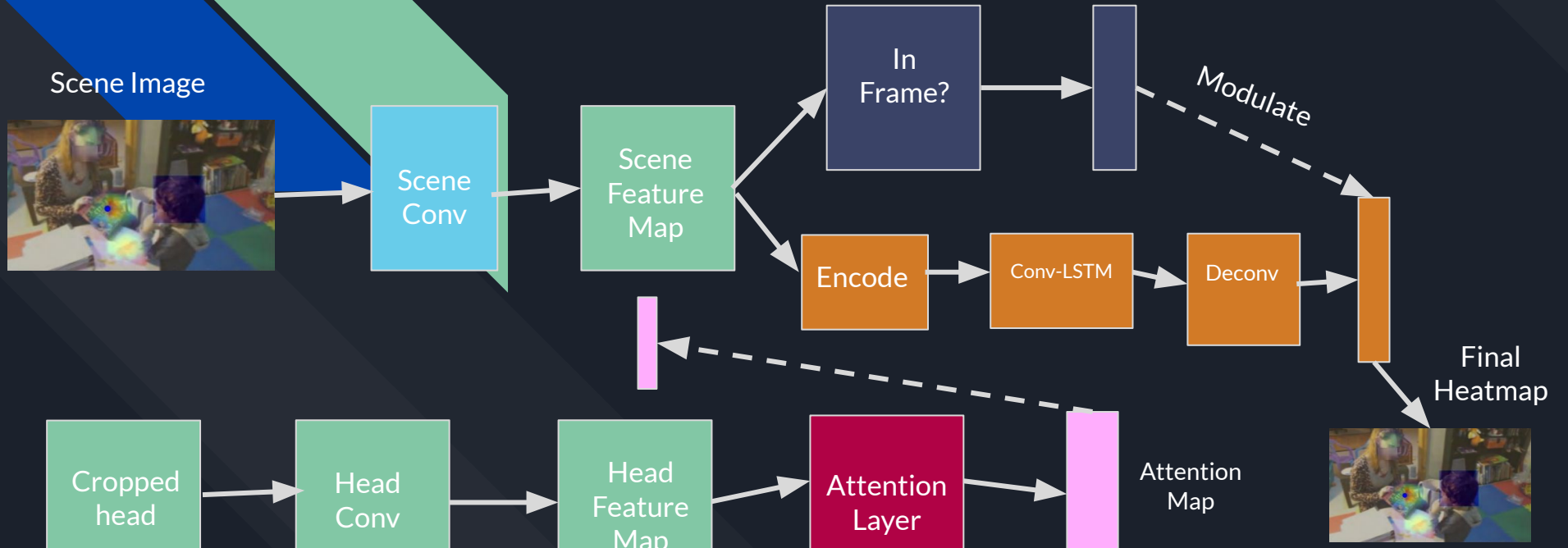
- **2° PHASE: TEXT-TO-IMAGE GENERATION:**
  -The text we extracted, is passed to `Stable Diffusion Pipeline` from the `diffusers` library.
  -This pipeline, pre-trained generate images from the provided textual input, obtained in the previous phase.
  ->FINAL RESULT: image output

- **3° PHASE: DISPLAYING IMAGES GENERATED:**
  After the image generation process, the resultant images are displayed to the user for review and evaluation.

Evaluating Gaze Detection for Children with Autism Using the ChildPlay-R Dataset

- Open Source dataset of children and adults autistic or non autistic
- Analysis runned with the Modified Spatiotemporal Gaze Detection (M-STGD) model and the Spatiotemporal Gaze Architecture (STGA)
- Children within the spectrum use gaze in a different way in respect to their peers
- 35 video clips from 15 videos collected from open-source videos from various environments, such as therapy centers and kindergartens, via YouTube

Evaluating Gaze Detection for Children with Autism Using the ChildPlay-R Dataset - 2



Emotion recognition using facial expressions

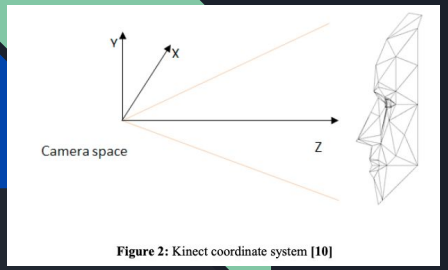


Figure 2: Kinect coordinate system **[10]**

Subject	MLP	3-NN
1	0.94	0.97
2	0.96	0.96
3	0.90	0.98
4	0.74	0.90
5	0.96	0.96
6	0.93	0.97
Average	**0.90**	**0.96**

Table 2: The results of the subject-dependent classification

1) EMOTION CLASSIFICATION: subject-dependent

- Collecting data with Microsoft Kinect
- Features extracted and calculated in 3D face model
- Extract AU:action units (some calculus with features)
- Classifications done with k-NN classifier and MLP
- Recognize seven emotions (neutral, joy, surprise, anger, sadness, fear and disgust)

VERY GOOD RESULTS!

Emotion recognition using facial expressions - 2

ES	neutral	joy	surprise	anger	sadness	fear	disgust
							
AU0	0.21	0.77	-0.10	0.30	0.17	-0.11	0.91
AU1	-0.06	0.09	0.60	-0.07	-0.04	0.20	0.13
AU2	-0.25	1.00	-0.49	0.06	-0.37	-0.60	0.88
AU3	-0.21	0.00	-0.13	0.04	-0.09	-0.17	0.00
AU4	-0.04	-0.47	0.58	-0.19	-0.02	0.28	-0.32
AU5	-0.23	-0.30	0.10	-0.34	-0.27	-0.02	-0.39

No	Subject-Session	MLP	3-NN
1	1-A	0.74	0.67
2	1-B	0.76	0.57
3	2-A	0.76	0.68
4	2-B	0.85	0.70
5	3-A	0.65	0.64
6	3-B	0.76	0.63
7	4-A	0.60	0.36
8	4-B	0.55	0.31
9	5-A	0.80	0.77
10	5-B	0.78	0.72
11	6-A	0.81	0.80
12	6-B	0.67	0.68
	Average	**0.73**	**0.63**

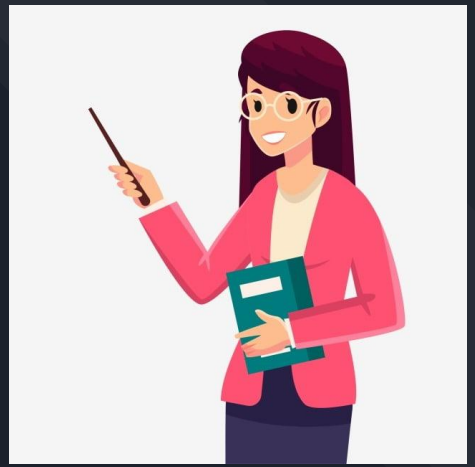
Table 6: The accuracy of subject-independent classification for "natural" division of data

2) EMOTION CLASSIFICATION: subject-independent (for all users together)

- same extraction of AU from kinect image, but from different people
- MLP results > KNN results

RESULTS EVIDENTLY DECREASE

First solution: Studio Assistant



- Multimodal studio assistant to help with AAC students with special needs
- Image generation through speech to visualize the subjects of the explanation
- Gaze control and emotion recognition to assess difficulties or give positive feedback with gamification setup
- Enhancing teaching-learning experience and overcome educational barriers

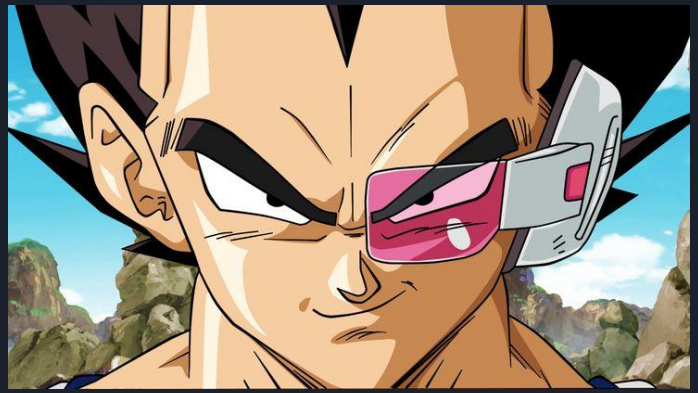
Functionalities Studio Assistant

- SPEECH TO IMAGE GENERATION: based on the first paper, using a diffusion model that convert speech audio in to a textual input through a text encoder; then the text will be used as input and guidance for the image generation, correlated to the lesson or teacher's explanation.

- CONCENTRATION AID: based on the second and third paper, using MLP or k-nn classifier to recognize emotion by a camera and gaze detection, with the aim of better understand the emotions, and understand when it could be distracted through gaze, and therefore help him to better understand the concepts, in relation to his state of mind and concentration.

Second solution: Super Glasses

- adjusting the volume of the audio input to increase concentration
- extracting emotions from facial expressions, of people who are looking at the child with disturb
- visual focus on what is being read, or part of the board that is being looked at.



Functionalities Super Glasses

:) -> Happiness

- **Audio volume adjustment**: based on where your gaze is directed, the system increases the volume of the person you are looking at (who is speaking to you) and reduces the volume of all other voices or noises coming from outside.
- **Extracting emotions from facial expressions:** also based on the third paper, recognition of facial expressions that allow the special child to understand the emotions of the people (students/teachers) in front of him to better interface with them.
- **Visual focus:** technology that focuses on what you are reading, on a book or blackboard, and darkens what is in the background, increasing concentration and limiting distractions.

ARCADE

	Studio Assistant	Super Glasses
Access information and experience	Textual tutorial in display	Audio guide at the start of application
Represent the choice situation	System synthesize image with audio sample	Hardware shaped as glasses with headset
Combine and compute	Sound to address student that doesn't pay attention	Colored led to signal activation of hardware
Advice about processing	Interface to choose input between textual or speech	Audio advice to play at request
Design the domain	Visual feedback only for face and not gesture	Visual feedback in face recognition
Evaluate on behalf of the chooser	Predominantly audio after a couple of interactions	Autofocus if setted to do so

Issues

- **EXPANSIVE HARDWARE** choices for the second solution = Glasses with microphone and screen to display video feedback, audio aid with headset etc.)
- Realize speech-to-image model in **REAL-TIME**
- Concentration aid based on **MULTIPLE PSYCHOLOGICAL STUDIES**
- Implementing emotion recognition is difficult on a small embedded device with **LOW COMPUTATIONAL RESOURCES**

References

1. *Akamsha Timande, Pallavi Borse, Vaishnavi Lande, & A.G.Sharma. (2024). Speech to Image Generation by Stable Diffusion Model. SSGM Journal of Science and Engineering, 2(1), 89–91. Retrieved from https://ssgmjournal.in/index.php/ssgm/article/view/116*

2. *Boluk, Nursena ; Kose, Hatice. / **Evaluating Gaze Detection for Children with Autism Using the ChildPlay-R Dataset**. 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition, FG 2024. Institute of Electrical and Electronics Engineers Inc., 2024. (2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition, FG 2024).*

3. *Paweł Tarnowski, Marcin Kołodziej, Andrzej Majkowski, Remigiusz J. Rak, Emotion recognition using facial expressions, Procedia Computer Science Volume 108, 2017. Retrieved from https://www.sciencedirect.com/science/article/pii/S1877050917305264*