Knowledge-Driven Dialogue and Visual Perception for Smart Orofacial Rehabilitation

Jacobo López-Fernández(✉), Luis Unzueta, Meritxell Garcia, Maia Aguirre, Ariane Méndez, and Arantza del Pozo

Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009 Donostia-San Sebastián, Spain
{jlopez,lunzueta,mgarciap,magirre,amendez,adelpozo}@vicomtech.org

Abstract. This paper addresses the problem of accomplishing Orofacial Rehabilitation (OR) with the assistance of artificial intelligence. The main challenges involve accurately monitoring and interacting with the trainees, while preserving user experience. We analyse different approaches to solving these challenges and propose a methodology to build smart knowledge-driven OR systems that focus on automated interaction. Our proposal leverages the combination of vision-based micro and macro facial expression recognition and skill-based dialogue systems, which facilitate encapsulating the knowledge of rehabilitation professionals into natural language interactions. Experimental results of spoken keyword spotting and micro and macro facial expression recognition algorithms are provided. The OR expressions image dataset employed in our experiments is also published to support further research in the field.

Keywords: Orofacial Rehabilitation · Dialogue Systems · Facial Expression Recognition

1 Introduction

Within an ageing society it is common for individuals to develop physical or cognitive detriment. These impairments typically impact on life quality and people frequently need to be provided with rehabilitation services. Orofacial Rehabilitation (OR) is the branch of Physiatry dedicated to mitigating physical impairments in the orofacial system, which is the set of organs responsible for the physiological functions of breathing, sucking, swallowing, speaking and phonation, including all kinds of facial expressions [21]. Examples of facial gestures for OR are: *bite lower lip*, *bite upper lip*, *blink*, *blow cheeks*, *blow left cheek*,

Supported by SHAPES – Smart and Health Ageing through People Engaging in Supportive Systems - is funded by the Horizon 2020 Framework Programme of the European Union for Research Innovation. Grant agreement number: 857159 - SHAPES – H2020 – SC1-FA-DTS – 2018–2020.

A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 397–411, 2023.
https://doi.org/10.1007/978-3-031-34586-9_26

blow right cheek, *close eyes* (stronger than *blink*), *look left*, *look right*, *frown*, *hide lower lip*, *hide upper lip*, *kiss*, *kiss left* (moving the mouth to the left), *kiss right*, *open eyes* (more than normal), *open mouth*, *press lips*, *rise eyebrows*, *rise nose* (as if it would smell bad), *show teeth* (not smiling), *smile* (without showing teeth), *smile left* (rise left mouth corner), *smile right*, *tongue forward*, *tongue left* and *tongue right*. During an OR session, the trainee would exercise these gestures several times, starting from a neutral facial expression until the maximum gesture intensity is reached and then relaxing again.

The OR process is highly demanding in terms of expert supervision, especially when concerning elderly people [13]. Consequently, minimising the need for caregiver support during the rehabilitation process with the less possible impact on user experience is one of the main open challenges in the field to date [22]. Dialogue Systems (DS) allow human-machine natural language communication through text or speech, just as humans interact with each other. Despite their potential to automate caregiver support resembling the traditional way, their application has not been fully explored for OR. One of the main barriers to exploit DS in the rehab setting is the linguistic expertise required to model each rehabilitation process in terms of intents, entities and dialogue rules [17]. This makes it difficult for professionals to codify their knowledge in the form of dialogue. In addition, the feasibility of automated spoken interaction with mild speech impairments related to orofacial disorders has not yet been tested.

On the other hand, a smart OR system requires Facial Expression Recognition (FER) algorithms capable of efficiently spotting macro and micro expressions (i.e., gestures), together with their degree of achievement to a canonical reference in order to provide real-time feedback and evaluate the progression. This is challenging because state-of-the-art FER methods typically handle fewer facial gestures than those mentioned above. Moreover, the most accurate methods tend to have a higher complexity that might hinder their deployment in devices with limited computational resources, such as smartphones [23]. Besides, each person's neutral expression varies from person to person and, therefore, might prevent FER models from generalizing well to all facial appearances.

An additional challenge in the field of smart rehabilitation in general is the variety of smart devices such as smartphones, tablets and smart speakers that are progressively growing in use for ubiquitous rehabilitation applications [20]. In parallel, traditional client-server architectures are also transitioning towards more distributed architectures [6], demanding to minimise the transfer of sensitive data over the Internet.

In order to address the challenges described above, this work proposes a novel combination of skill-based DS and FER algorithms towards engaging AI-powered automatic OR user experience. More specifically, our contributions can be summarised as follows:

– A smart OR system architecture that allows blending the output of edge-deployed spoken interaction and computer vision modules capable of recognizing spoken keywords and orofacial expressions, with natural language interaction dialogues derived from knowledge encoded directly by rehabilitation professionals in dialogue skills.

- A spoken keyword spotting (KWS) phrase and model experimentally shown to be robust to mild speech impairment.
- A FER method to measure the degree of achievement of macro and micro facial gestures compared to a canonical reference effectively and efficiently.
- The OROFACE dataset to support further research in this field. The dataset can be downloaded from here: https://datasets.vicomtech.org/di24-oroface-dataset/oroface.zip

The remaining sections of the paper are structured as follows: Sect. 2 analyses previous work done under the scope of mixed automated rehabilitation approaches involving FER and DS; Sect. 3 introduces the proposed knowledge-driven dialogue and visual perception system architecture for smart OR; Sect. 4 evaluates this against other state-of-the-art (SOTA) prototypes and shows experimental results for the KWS and FER algorithms; finally, Sect. 5 draws upon the conclusions and potential lines for future work.

2 Related Work

2.1 Smart Rehabilitation Systems and Dialogue

Medical rehabilitation is a procedure executed by professionals including physiatrists, physiotherapists, nursing personnel, occupational therapists or diverse medics that can provide a diagnosis involving rehabilitation. Typically either patients or medical field experts need to be relocated on site for a rehabilitation session. When relocation is not an option, telehealth mechanisms flourish with the intention of performing secure virtual rehabilitation activities remotely [19]. Virtual Reality and friendly graphical user-interfaces have been added to remote rehabilitation systems to perform cognitive and physical recovery exercises supervised by a trainer [8]. Auto data gathering techniques through guided questionnaires were innitially proposed to end-users but, more recently, robotics and computer vision technologies have became more prominent to monitor and register patient performance and complete electronic health reports, which are later evaluated by competent healthcare experts [5]. However, patients still prefer to have nursing personnel next to their remote controlled rehabilitation machinery [11], highlighting the need for more natural interaction mechanisms with smart telerehabilitation systems that resemble the traditional face-to-face recovery procedure.

Dialogue Systems (DS) have been introduced in diverse knowledge fields, to provide human-like interactions with decision making capabilities independent from expert personnel involvement [14]. To prevent negative patient experience, smart rehabilitation systems shall not only collect user generated data and keep track of the exercises performed on behalf of the caregivers, but also interact as humanly as possible in order to engage individuals: informing, guiding, recommending and motivating with personalized content [4]. In this sense, speech technologies can narrow the user experience gap introducing spoken input and responses [15] as in a traditional rehabilitation process, while replacing more

common but less natural visual user interfaces. A key technology in voice-based interaction is Spoken Keyword Spotting (KWS), which enables triggering attention from the system using a custom spoken word or short phrase. It also allows to initiate interactions only when users want to, enhancing user experience and acceptance [10]. Although spoken interaction is only feasible for patients with orofacial disorders not affecting speech production or leading to mild speech impairments, a comparable user experience can be achieved through chat-like text interactions in natural language. In addition, DS can fill the interactive conversational role of health-care professionals, exploiting domain knowledge in order to provide correct answers [7]. Unfortunately, the development of DS still requires considerable manual effort and expert linguistic knowledge. In order to address this problem, customisable conversation structures or dialogue skills that can adapt to each patient and conversational agent [12] have started to be exploited.

2.2 Facial Expression Recognition in Orofacial Rehabilitation

FER approaches are composed of two main phases: (1) facial image pre-processing and (2) facial expression feature classification. Image pre-processing typically includes the following stages [23]: (1) facial region detection and alignment, (2) frame normalization, and (3) motion magnification. The first stage involves extracting the facial image and landmarks from the input image, then reducing the variation in face scale and in-plane rotation. Deep Neural Networks (DNNs) currently obtain the best results for these two tasks [9]. In the second stage, meaningful frames are automatically selected from the entire gesture sequence, typically by aligning the input samples into the same number of frames through temporal interpolation [3]. The third stage usually involves manipulating the sequence transformed to the frequency domain to detect subtle motion changes of micro-expressions [16]. Finally, facial expression categories are inferred from the pre-processed facial image through a Machine Learning method (e.g., another DNN).

Orofacial rehabilitation requires recognizing more macro and micro facial gestures than those usually considered by state-of-the-art FER methods and published datasets [2]. Besides, they should be measured with respect to the specific neutral expression of each person. This means that more accurate methods and specific datasets are needed. Moreover, temporal interpolations and operations in the frequency domain might be unfeasible to get real-time feedback while performing the exercises on devices with limited computational resources. Thus, for our goal, we need to deploy lightweight DNNs, and create procedures for frame normalization and motion magnification adapted to our context.

3 Proposed Approach

3.1 Architecture Design and Workflow

The proposed Smart OR System Architecture is illustrated in Fig. 1. It entails a DS as responsible for the follow-up process, relegating the specialist to a supervis-

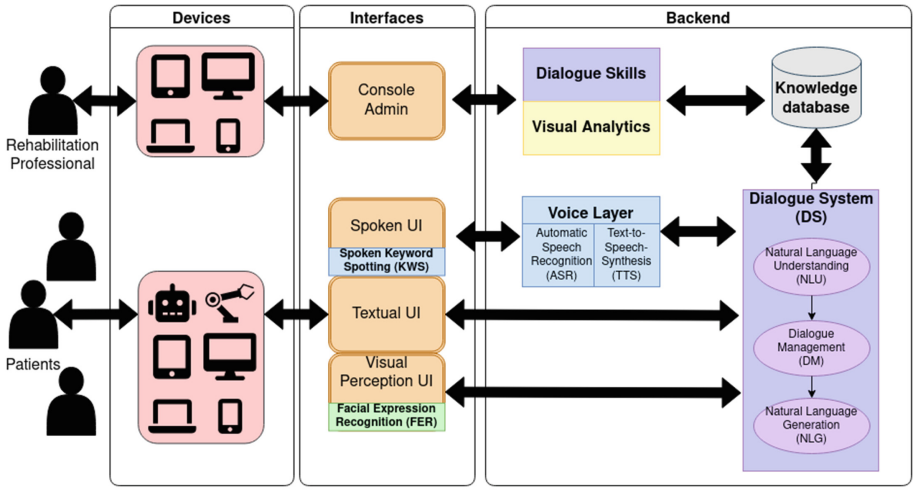


Fig. 1. Smart OR System Architecture

ing role consisting on introducing clinical knowledge through dialogue templates or skills to guide natural language interaction with the patient. This opposes from previous approaches that had professionals driving the whole process through screens with limited operational flexibility and full attention demand on the task. The underlying assumption is that the DS shall be capable of drawing more attention from the user than filling forms, while keeping patients active and engaged improving user experience.

The clinical knowledge encoded in the dialogue skills is stored in a knowledge database, which is then used to automatically instantiate the following DS modules:

- A Natural Language Understanding (NLU) module that performs intent classification and entity extraction on natural language input from the user.
- A Dialogue Management (DM) module that decides which is the next state in the dialogue and stores input information that may influence decision making in the dialogue memory.
- A Natural Language Generation (NLG) module that creates the appropriate natural language response to be returned by the system.

The DS supports both, text and spoken interactions through an optional voice layer that includes automatic speech recognition and speech synthesis technologies, plus a spoken keyword spotting (KWS) module. This way, OR patients without or with mild speech problems are able to interact with the system as they would do with their rehabilitation caregivers.

FER algorithms provide of real-time orofacial rehabilitation information which can be checked against medically accurate exercise templates in order to provide both trainees and instructors from useful feedback about the rehabilitation process. The output of the FER module is fed directly to the DM,

which takes the visual perception information received into consideration for the decision making process in the next natural language dialogue interaction with the user.

In line with current trends, the architecture design contemplates the use of a wide variety of devices (e.g. computers, smartphones, tablets, smart speakers, robots, etc.). To ensure optimum performance, input devices shall include microphones prepared for beamforming, echo cancellation and noise reduction to deliver precise high quality audio and cameras with stabilised image and precise focusing to capture every gesture.

User information gathered from dialogue conversations and video capture peripherals is also stored in the knowledge database, which is subject to the terms of privacy and data protection for the sake of patients confidence. In order to minimise the amount of personal data to be stored in the knowledge database, the KWS and FER algorithms have been devised to be deployed on the edge. This way, only spoken interactions addressed to the system and FER performance results shall be collected. For added security and data protection, the system back-end could be deployed both on premise in a local server or in a restricted cloud network with limited access.

Additionally, the results obtained by completing the whole rehabilitation process can be checked by the expert on an intuitive multi-platform visual analytics component.

In practice, the proposed Smart OR System Architecture involves two separate workflows for rehabilitation professionals and patients:

- **Rehabilitation professionals**: use dialogue skills to define rehabilitation sessions for patients including e.g. the set and sequence of gesture exercises they should perform. Such expert knowledge is then exploited to automatically instantiate DS that guide and interact with the users throughout the rehabilitation session using natural language. Once the sessions are completed, clinicians can consult a visual analytics panel to check how the sessions went, evaluate the results achieved and plan next rehabilitation steps.
- **Patients**: open the smart OR rehabilitation application in their preferred device (i.e. computer, smartphone, tablet, robot, etc.). Calibrate the camera of the FER module to their neutral face position. The system guides them through the rehabilitation session, proposing sequences of gesture exercises adapted to their needs and providing automatic feedback on their performance in natural language. Patients without or with mild speech impairments can even interact using their voice, just as they would with their rehabilitation caregivers. After the session is completed, patients can also receive feedback from their caregivers and a new set of exercises to perform.

3.2 Dialogue Skills and KWS for OR

The most convenient method for the automatic generation of dialog rules are the so-called Dialogue Skills. To this end, each Skill is provided with the ability to manage dialogues that follow specific patterns and respect a specific domain

logic. For all dialogues that partially or completely follow the default dialog structure in the Skill, the interaction rules are automatically instantiated and the Dialog Manager (DM) module gets ready to be used.

OR Dialogue Skills enable interaction capabilities such as repeating, passing, changing, completing, exiting, pausing, continuing or getting additional information about rehabilitation exercises or contacting professionals for help, following a particular conversation structure. These capabilities are linked to their respective NLU semantic tags (i.e., intents and entities) defined for the OR Skill to classify patient input and allow the DM to keep track of the dialogue state by saving and updating information related e.g. to the current exercise, the current step, the exercise number, the step number and any additional requirements (if any) as dialogue attributes. Then, the expert rules created for the Skill are able to handle dialogue state changes based on the NLU semantic tags detected. In response, users are encouraged to keep going with the OR process while receiving positive feedback or being alerted if the FER module does not detect the expected performance. Once the OR Dialogue Skill is developed, professionals only need to fill in a graphical user interface including the sequence of gesture exercises to be completed by the patients.

As mentioned before, Spoken Keyword Spotting (KWS) allows enhancing user experience and acceptance of voice-based interfaces and, thus, has been included in the proposed Smart OR System Architecture. However, a KWS phrase suitable for the OR environment precises to meet certain characteristics: it should be chosen to be as language agnostic as possible avoiding the inclusion of language-specific phonemes; and it should have a length of three to four syllables in order to avoid similarity with other words in short phrases and the complexity of long phrases. For the experimental purposes in this work, the KWS phrase "Hey Nari" has been chosen following those principles. In addition, a dataset recorded by 42 speakers of different languages has been compiled for model training and testing purposes, as described in Sect. 4. Each speaker was asked to record 12 positive audio samples containing the desired phrase and 12 negative audio samples containing diverse speech, for a total set of 1008 audio samples.

3.3 Efficient Facial Expression Recognition for OR

We need to tackle the following tasks to design an appropriate FER method for our goal:

1. Build a balanced dataset with different facial appearances in the wild, performing several trials of all the required facial gestures for orofacial rehabilitation, including neutral expressions. This dataset will allow us to build the canonical reference for the trainee's accomplishment measurements.
2. Design an effective and efficient facial image processing strategy for frame normalization and motion magnification, especially for micro gestures.
3. Design appropriate metrics for facial gesture accomplishment, tailored to each person's neutral expression.

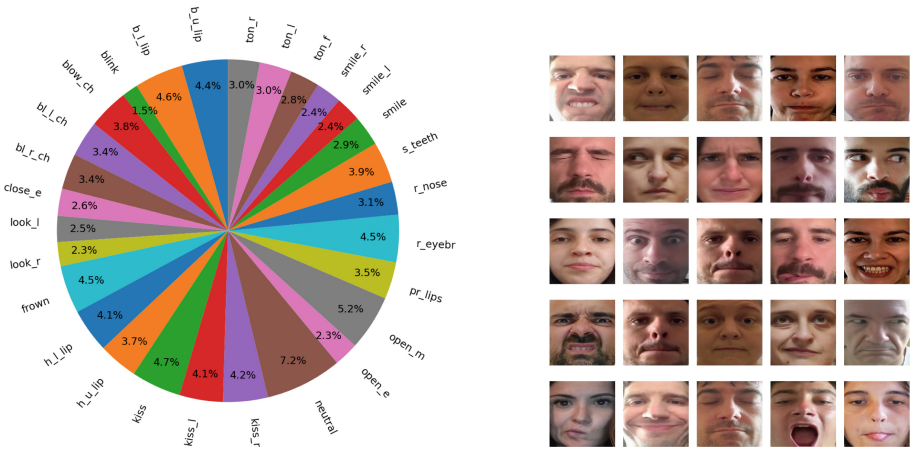


Fig. 2. Data distribution and image samples of the OROFACE dataset.

4. Analyze the visual discriminability of the required facial gestures to train an effective and efficient facial expression feature classification model.

For the first task, we first involve a professional expert in orofacial rehabilitation to define a training session. This expert is recorded from a frontal viewpoint while performing the training session, including all required trials of all the considered facial gestures. Then, similarly, other people are recorded replicating this session while watching it as guidance, as in a mirror workout session. All these people should perform all the exercises with sufficient precision to act as a further reference to others. Finally, we segment the recorded videos, extracting the frame sequences corresponding to the full gesture action, from the starting neutral expression to the ending neutral expression, but without including it. Finally, we segment separate sequences, including the neutral expression (e.g., the moment before all trials start). Following this approach, we have built the OROFACE dataset, recording 20 individuals performing the facial gestures mentioned in the introduction (28 in total) 2–3 times each, as explained above. After segmenting the videos and removing the less successful trials, the dataset contains 17,133 images, distributed as shown in Fig. 2, with the following abbreviations: b_l_lip=bite lower lip, b_u_lip=bite upper lip, blow_ch=blow cheeks, bl_l_ch=blow left cheek, bl_r_ch=blow right cheek, close_e=close eyes, look_l=look left, look_r=look right, h_l_lip=hide lower lip, h_u_lip=hide upper lip, kiss_l=kiss left, kiss_r=kiss right, open_e=open eyes, open_m=open mouth, pr_lips=press lips, r_eyeb=rise eyebrows, r_nose=rise nose, s_teeth=show teeth, smile, smile_l=smile left, smile_r=smile right, ton_f=tongue forward, ton_l=tongue left, and ton_r=tongue right.

For the second task, we propose using contrast-enhanced normalized differential images (CENDIs), computed as shown in Algorithm 1. The five input parameters are: (1) the incoming aligned facial image **I** (like the samples shown

in Fig. 2), (2) an aligned facial image of the user with neutral expression of the user $\mathbf{I}_{\text{neutral}}$ obtained during an initial calibration step, (3) the grade of difference α (in the range of [0,1]), and the parameters for Contrast Limited Adaptive Histogram Equalization (CLAHE) [24] clip limit c (4) and tiles grid size g (5). CENDIs enhance the gesture's relevant areas by including the contrast between the user's actual and the neutral expression, with a small computing overhead. Figure 3 shows examples of CENDIs for different values of α.

Algorithm 1. Contrast-enhanced normalized differential image calculation.

procedure CALCCENDI($\mathbf{I}$, $\mathbf{I}_{\text{neutral}}$, α, c, g)
  $\mathbf{I}_{\text{diff}} \leftarrow \mathbf{I} - \alpha \cdot \mathbf{I}_{\text{neutral}}$ *(in single-precision floating-point format)*
  $\text{val}_{\text{min}}, \text{val}_{\text{max}} \leftarrow \text{getMinMaxValues}(\mathbf{I}_{\text{diff}})$
  $\mathbf{I}_{\text{diff}}^{\text{norm}} \leftarrow 255 \cdot (\mathbf{I}_{\text{diff}} - \text{val}_{\text{min}})/\text{val}_{\text{max}}$
  $\mathbf{I}_{\text{HSV}} \leftarrow \text{convert2HSV}(\mathbf{I}_{\text{diff}}^{\text{norm}})$ *(in 8-bit precision format)*
  $\mathbf{H}, \mathbf{S}, \mathbf{V} \leftarrow \text{splitInChannels}(\mathbf{I}_{\text{HSV}})$
  $\mathbf{V}_{\text{enhanced}} \leftarrow \text{applyCLAHE}(\mathbf{V}, c, g)$ [24]
  $\text{CENDI} \leftarrow \text{convert2RGB}(\text{mergeChannels}(\mathbf{H}, \mathbf{S}, \mathbf{V}_{\text{enhanced}}))$
  return CENDI
end procedure

Fig. 3. Examples of CENDIs with $\alpha = 0, 0.25, 0.5, 0.75,$ and 1 for the bite lower lip micro gesture, compared to the incoming aligned facial image (right).

For the third and fourth tasks, we consider using a DNN for facial expression feature classification trained with CENDIs generated from a dataset like OROFACE. Thus, the output of the DNN is a vector of scores of all the considered gestures. The closer the captured gesture's performance to the trained reference is, the higher the score for the corresponding class will be. However, even though gestures are perfectly performed, some gestures could be visually very similar (e.g., *blink* and *close eyes*), and the classifier could find difficulties in discriminating between them. Therefore, in some cases, we might require fusing some gestures, retraining, and retesting the DNN iteratively until we obtain an effective classifier, even though, in the end, we might request the user to distinguish between them for the exercises. Confusion matrices of testing data help us decide which gestures should be fused if required during this process. Nevertheless, our goal is to measure the degree of achievement to a canonical reference, and this approach is sufficient for that. In our context, we should select

lightweight DNNs for this classification and facial region and landmarks detection with a good trade-off between accuracy and computational complexity for devices with limited computational resources.

4 Experiments and Evaluation

This section includes a qualitative evaluation of the proposed Smart OR System Architecture, benchmarking it against methodologies with similar characteristics found in the literature. In addition, a quantitative evaluation of the developed KWS and FER models is also presented.

Table 1 summarizes the main features of the systems analysed, which have been chosen to address smart rehabilitation or healthcare support through dialogue and/or computer vision components. The main characteristics that have been compared across systems are: whether they use dialogue to interact with the users (DS); whether dialogue skills are exploited to facilitate the development of dialogue interfaces adapted to each user (Dialogue Skills); whether users can communicate with the system using spoken interaction (Spoken Interaction); whether computer vision algorithms are exploited to automatically monitor user performance (CV); whether they consider the use of smart interconnected devices by users and propose distributed artificial intelligent deployments (IoT Edge) with several embedded machine learning (ML) algorithms.

As it can be observed, it is hard to find a solution that combines dialogue and visual perception to address the specific problem of orofacial rehabilitation. Some approaches are based on the use of smart wearable IoT devices [20] or computer vision technology alone which are not smart nor applied to facial rehabilitation [1]. Other systems exploit DS and spoken interaction [4,15], but do not target rehabilitation applications nor blend visual perception components with the dialogue. None of the published works has proposed to apply conversational skills to facilitate the inclusion of personalized expert knowledge into system-patient interactions. Regarding smart devices deployment, some approaches reflect an IoT architecture but fail to address modern limitations by embedding lightweight ML algorithms on edge devices [4].

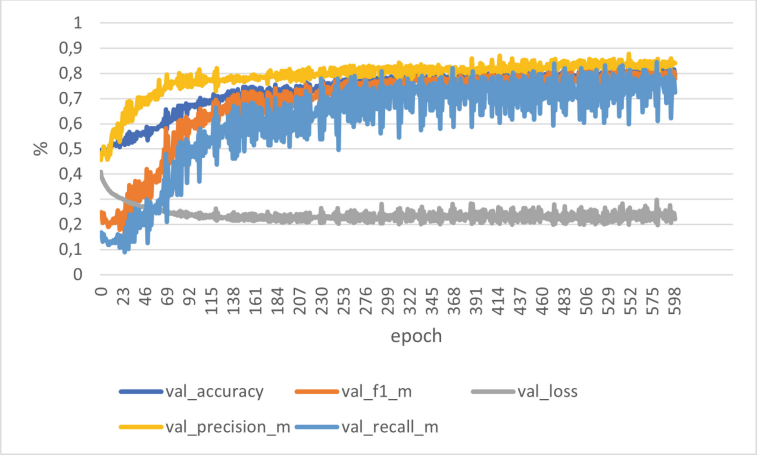
Quantitative evaluation of the KWS approach described in 3.2 has been carried out by training a model of the defined phrase on the compiled training dataset. Figure 4 shows the results achieved on the validation set for different training epochs, the best ones being those obtained with 600 epochs (final model). A small test set has been created to evaluate final model performance over different speakers and conditions as is later discussed in Table 2. As it can be seen, the maximum Accuracy obtained is 0.79, which can be considered acceptable. Overall, the Precision (0.84) of the model is higher than its Recall (0.72) leading to an F1 measure of 0.78 and 0.22 Loss.

Additionally, we have also preliminarily compared the performance of the KWS model on speech samples with and without mild speech impairment. Table 2 shows the average probabilities returned by the model for each case for a set of given test audio samples. Audio samples containing speech and noises that

Table 1. Feature Qualitative Check

Reference	Short Description	Features[a]				
		DS	***Dialogue Skills***	***Spoken Interaction***	***CV***	***IoT Edge***
[20]	Programmable device to guide rehabilitation patients	×	×	×	×	✓
[4]	Health dialog systems for patients and consumers	✓	×	✓	×	×
[15]	A dialogue monitoring scheme for a virtual doctor	✓	×	✓	×	✓
[1]	Wize Mirror - a smart, multisensory cardio-metabolic risk monitoring system	×	×	×	✓	×
Ours	Knowledge-Driven Dialogue and Visual Perception for Smart Orofacial Rehabilitation	✓	✓	✓	✓	✓

[a]Based on published research conference papers and journals.



Epoch Number	Metrics[a]				
	Acc	***F1***	***Loss***	***Precision***	***Recall***
600	0.79	0.78	0.22	0.84	0.72

[a]Based on micro metrics from validation phase.

Fig. 4. KWS Quantitative Analysis

differ from the KWS phrase are added to highlight the value of the results. The KWS model is inferred on overlapped fixed-size audio frames (smaller than the complete sample) trying to find a high probability value. Average probabilities reflect a clear distance between positive non-mild-speech-impairment samples (0.60) and negative samples (0.16) with positive mild-speech-impairment sam-

ples (0.45) finding their probability space somewhere in the middle of them closer to positive values. Therefore, the probability gap between those three groups appoints that this smart OR system architecture is preliminarily KWS capable on patients with mild speech impairments provided that the keyword probability detection threshold is flexibly tuned (e.g. 0.30).

Table 2. KWS Performance on Samples with and without Mild Speech Impairment

Voice conditions	Non-impaired speech	Non-impaired speech	Mild impaired speech
Wake-Up Word Audio **Sample Type**	Positive	Negative[b]	Positive
Average Probability **Achieved**[a]	0.600493349	0.157178358	0.449133078

[a] Averages are calculated based on several audio frame probabilities [0.0–1.0].
[b] Negative non-impaired speech results extrapolable to negative mild impaired speech results.

Finally, to test the convenience of CENDIs for FER in our context, we have trained ten lightweight DNNs for the first five subjects of OROFACE performing all the recorded gestures (i.e., two recognition models per subject with $\alpha = 0$ and 0.5). The data used for training each model includes the rest of the dataset's subjects, except the targeted one used for testing. We have chosen the EfficientNet-lite-0 DNN architecture [18] for these models, as it is appropriate for deploying in devices with limited computational resources.

Table 3 shows that the models trained with $\alpha = 0.5$ obtain a better recognition accuracy, as expected, because $\alpha = 0$ does not include a contrast between the user's actual and the neutral expression.

Table 3. Average gesture recognition accuracy (%) for the first five subjects of OROFACE with EfficientNet-lite-0 trained with the 28 gestures.

α **value**	**Subject 1**	**Subject 2**	**Subject 3**	**Subject 4**	**Subject 5**	**Overall**
0	75.01	66.57	68.47	63.95	57.45	**66.29**
0.5	80.83	76.53	72.58	70.78	66.78	**73.50**

In contrast, $\alpha = 0.5$ does. Figure 5 shows the normalized average confusion matrix for $\alpha = 0.5$. It reveals that in this configuration, EfficientNet-lite-0 confuses some gestures. This could be because the model is not discriminative enough, but also because in practice, some users might perform some requested gestures also moving other facial parts unconsciously (e.g., *open eyes* also rising the eyebrows, in a similar way to the *rise eyebrows* gesture).

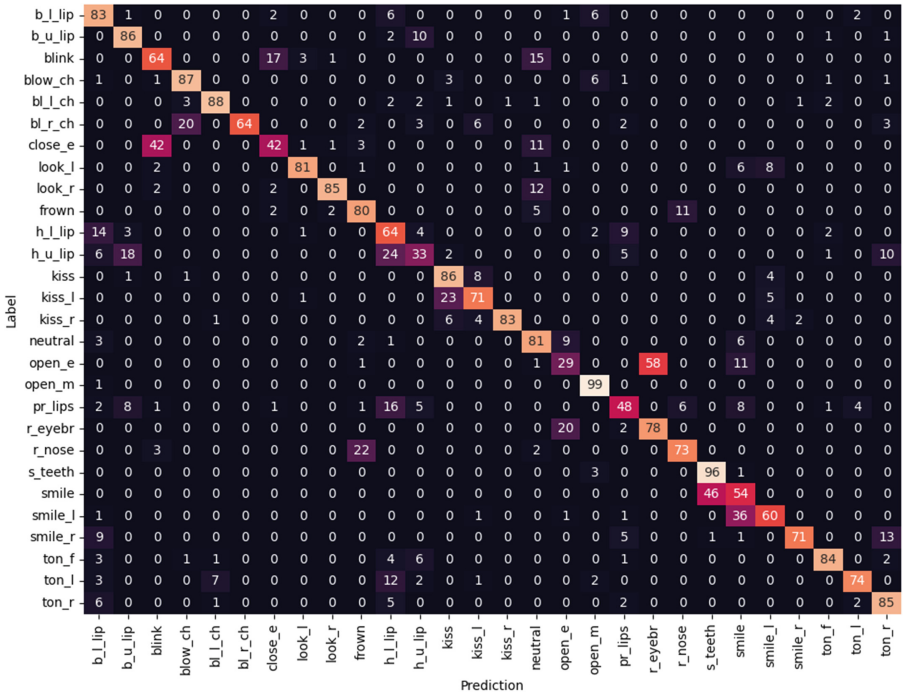


Fig. 5. Normalized average confusion matrix for $\alpha = 0.5$, for EfficientNet-lite-0 trained with all the gestures and tested with OROFACE's first five subjects.

Moreover, these results are computed per frame in the cropped sequences, where the highest gesture intensity typically happens around the middle, and starting and ending frames might not represent the gesture properly in some cases. Thus, following the proposed approach, the problematic gestures should be fused, and then the DNN retrained and retested iteratively until sufficient accuracy is obtained.

5 Conclusions and Future Work

This paper proposes a Smart System Architecture to automate OR accurately while preserving user experience through facial expression recognition (FER) and natural language dialogue. Both, textual and spoken interactions are supported, allowing patients to communicate with the system as they would with their caregivers. In addition, dialogue skills are introduced as a mechanism to facilitate the inclusion of personalized expert professional knowledge in the system. The presented architecture also supports the use of a variety of smart devices and integrates spoken interaction and visual perception components on the edge, with the aim of minimising the transfer of sensitive data over the Internet.

The main features of the proposed Smart OR System Architecture have been benchmarked against approaches with similar characteristics found in the literature verifying that, although other published solutions use some of the same components, none of them combines conversational skills and visual perception to address the specific problem of orofacial rehabilitation. In addition, spoken keyword spotting (KWS) and FER models have also been experimentally evaluated. The developed KWS module has achieved acceptable accuracy and robustness to mild speech impairment. Regarding FER, the trained model has obtained an overall recognition accuracy of 73.50 and the OR expressions image dataset employed for experimentation is shared to support further research in the field.

Future research should further develop and confirm these initial findings by implementing a concrete use case and piloting with real users. In addition, the feasibility of using spoken interaction with a wider range of speech impairments caused by orofacial disorders should also be more thoroughly explored. The same applies to FER where further investigation should be carried out on gesture overlapping. Finally, interesting questions for future research can be derived from working towards embedding all the technological components and algorithms on smart devices on the edge.

References

1. Andreu, Y., et al.: Wize mirror - a smart, multisensory cardio-metabolic risk monitoring system. Computer Vision and Image Understanding, 148:3–22. Special issue on Assistive Computer Vision and Robotics - "Assistive Solutions for Mobility, Communication and HMI" (2016)
2. Ben, X., et al.: Video-based facial micro-expression analysis: a survey of datasets, features and algorithms. IEEE Trans. Pattern Anal. Mach. Intell. pp. 1–1 (2021)
3. Ben, X., Zhang, P., Yan, R., Yang, M., Ge, G.: Gait recognition and micro-expression recognition based on maximum margin projection with tensor representation. Neural Comput. Appl. **27**(8), 2629–2646 (2016)
4. Bickmore, T., Giorgino, T.: Methodological review: health dialog systems for patients and consumers. J. Biomed. Inform.-JBI (2021)
5. Bouteraa, Y., Abdallah, I.B., Alnowaiser, K., Ibrahim, A.: Smart solution for pain detection in remote rehabilitation. Alexandria Eng. J. **60**(4), 3485–3500 (2021)
6. Chaparro, J.D.: The shapes smart mirror approach for independent living, healthy and active ageing. Sensors, **21**(23) (2021)
7. Chen, H., Liu, X., Yin, D., Tang, J.: A survey on dialogue systems: recent advances and new frontiers. SIGKDD Explor. Newsl. **19**(2), 25–35 (2017)
8. Thumm, P.C., Giladi, N., Hausdorff, J.M., Mirelman, A.: Tele-rehabilitation with virtual reality: a case report on the simultaneous, remote training of two patients with Parkinson disease. Am. J. Phys. Med. Rehabil. **100**(5) (2021)
9. Gogic, I., Ahlberg, J., Pandzic, I.S.: Regression-based methods for face alignment: a survey. Signal Process. **178**, 107755 (2021)
10. Kepuska, V., Breitfeller, J.: Wake-up-word speech recognition application for first responder communication enhancement. In: Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense V, vol. 6201, pp, 431–438. SPIE (2006)

11. Kim, J., Lim, S., Yun, J., Kim, D.H.: Telerehabilitation needs: a bidirectional survey of health professionals and individuals with spinal cord injury in south Korea. Telemedicine Journal and e-health : the Official Journal of the American Telemedicine Association, 18(9), 713–717 (2012)
12. Liu, B., Mazumder, S.: Lifelong and continual learning dialogue systems: learning during conversation. In: Proceedings of the AAAI Conference on AI, **35**(17) (2021)
13. Maags, C.: Hybridization in china's elder care service provision. Soc. Pol. Adm. **55**(1), 113–127 (2021)
14. Major, L., Warwick, P., Rasmussen, I., Ludvigsen, S., Cook, V.: Classroom dialogue and digital technologies: a scoping review. Educ. Inf. Technol. **23**(5), 1995–2028 (2018)
15. Mallios, S., Bourbakis, N.: A dialogue monitoring scheme for a virtual doctor. In: 2015 National Aerospace and Electronics Conference (NAECON), pp. 249–253 (2015)
16. Le Ngo, A.C., Johnston, A., Phan, R.C.W., See, J.: Micro-expression motion magnification: Global lagrangian vs. local Eulerian approaches. In: 13th IEEE International Conference on Automatic Face & Gesture Recognition, pp. 650–656. IEEE Computer Society (2018)
17. Okur, E., Sahay, S., Nachman, L.: Data augmentation with paraphrase generation and entity extraction for multimodal dialogue system (2022)
18. Tan, M., Le, Q.V.: Efficientnet: rethinking model scaling for convolutional neural networks. In: Chaudhuri, K., Salakhutdinov, R., (eds.), Proceedings of the 36th International Conference on Machine Learning ICML, vol. 97 of Proceedings of Machine Learning Research, pp. 6105–6114. PMLR (2019)
19. Terrell, E.A., Bopp, A., Neville, K., Scala, D., Zebley, K.: Telerehabilitation policy report: Interprofessional policy principles and priorities. Int. J. Telerehabilitation, 13(2) (2021)
20. Tradigo, G., Vizza, P., Guzzi, P.H., Fragomeni, G., Ammendolia, A., Veltri, P.: A programmable device to guide rehabilitation patients: design, testing and data collection. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1487–1491 (2020)
21. Williams, M., Evans, P.L., Serriah, M.A.: Modern maxillofacial rehabilitation, pp. 381–420. Springer International Publishing, Cham (2022)
22. Zak, M., et al.: Frailty syndrome-fall risk and rehabilitation management aided by virtual reality (VR) technology solutions: a narrative review of the current literature. Int. J. Environ. Res. Publ. Health, **19**(5), 2985 (2022)
23. Zhou, L., Shao, X., Mao, Q.: A survey of micro-expression recognition. Image Vis. Comput. **105**, 104043 (2021)
24. Zuiderveld, K.: Contrast Limited Adaptive Histogram Equalization, pp. 474–485. Academic Press Professional Inc, USA (1994)