The Improvement of the Speech Recognition Technology (SRT) by Artificial Intelligence (AI)

Matej Cigale, Mitja Lustrek, Matjaz Gams
Jozef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenija
matej.cigale@ijs.si, mitja.lustrek@ijs.si, matjaz.gams@ijs.si

Torsten Kramer, Meike Engelhardt, Peter Zentel
Pädagogische Hochschule, Heidelberg
Institut fur Sonderpadagogik
Keplerstrabe 87D
Heidelberg, Germany
kraemer@ph-heidelberg.de, engelhardt@phheidelberg.de, zentel@ph-heidelberg.de

ABSTRACT: *People with Profound Intellectual and Multiple Disabilities (PIMD) stand for a broad and very heterogeneous spectrum of people that are characterised by some common aspects like a severe intellectual disability usually in combination with a lack of conventional and symbolic communication abilities, coupled with the need for high levels of support due to comorbidities or other possible disabilities (i.e., motor or sensorial impairments). Supporting these individuals is extremely challenging, as their communication signals are atypical and idiosyncratic. Therefore, a plethora of these behaviours are not or not easily readable for the caregivers. Without background information on a specific person with PIMD, it is hard for a caregiver, even a trained professional, to interpret the desires and mental state of the person they are interacting with, which leads to a stressful interaction for both. With advances in computer vision (CV), speech recognition technology (SRT) and artificial intelligence (AI), we are making the first steps in codifying these behaviours and attempting to mechanically extract the meaning of the communication. The INSENSION project aims to use these advancements to catalogue the actions of persons with PIMD and the environment and thus provide feedback to caregivers and enable individuals to control their surroundings. A similar system could be used to analyse the behaviour of healthy individuals so that the generalised and personalised expressions of body language could be codified and compared across cultures and individuals.*

Keywords: Communication, Gesture Recognition, PILD, AI

Received: 3 May 2018, Revised 12 June 2018, Accepted 20 June 2018

© 2018 DLINE. All Rights Reserved

1. Introduction

People with PIMD experience a lot of trouble when they are attempting to communicate to the outside world. Generally, a profound intellectual disability, which complicates the learning of new skills for them, is combined with other sensory or physical impairments, which lead to an unusual communication in comparison to people without disabilities.

On this occasion, the task of their caregivers (parents, therapists, etc.) is to interpret their communication attempts and teach them to make further attempts more distinct. The problem arises from the fact that the communication attempts of the people with PIMD are indistinct, mostly unique to each individual, and hard to distinguish or interpret for people who are not close to them. The differences stem from different abilities of each person, different reception of the gestures from their caregivers and other external factors.

At present, only close caregivers are able to interpret the desires of people with PIMD in the right way, nevertheless often combined with feelings of insecurity. This makes it difficult to easily expand the circle of communication partners. The INSENSION project faces this issue by aiming to use the advances in computer vision, specifically new ways to extract posture and facial expressions from video to codify them. In the second stage of the processing, the context of the expression will be extracted, i.e. what is the state of the environment around the person. This way, the expression can be coupled with the interaction and the intent of the communication can be interpreted. This would hopefully provide improvements for both sides: for the people in contact with people with PIMD to have a window into their internal states and see their attempts and, of course, for the people with PIMD themselves. However due to their disabilities, the communication attempts of people with PIMD seem simpler and their internal working models are often assumed to be limited to temporally and spatially neighbouring desires that makes the interpretation context smaller.

This research has broader implications as a similar, but more complex, system could be used to interpret communications of individuals so that their behaviour could be objectively determined enabling a more rigid research into the communication of people and their internal state.

The rest of the paper is organised as follows: In Section 2 we look at the state of the people with PIMD and their communication attempts. In Section 3 we take a brief look at the underlying technical advancements that can facilitate the extraction of posture and facial expressions and the vocalisations of the person. In Section 4 we present the annotations that will be the input for the Machine Learning (ML) system. In Section 5 we discuss the implications of this system and present some caveats to the system. In section 6 we look at broader implications of the system as it could be used on a more general population to systematically codify the interactions as well as further fields of inquiry that could be developed based on this system.

2. People with PIMD

People with PIMD, as the name implies, have multiple disabilities, which makes it even harder for them to participate in the large number of non-barrier-free parts of our society. Generally, PIMD means a profound intellectual disability combined with other sensory (blindness, deafness) or physical impairments (lack of mobility, problems with fine hand movements, etc.). These factors severely influence the person's ability to live without any care, support or therapy of others [1]. Individuals with PIMD have an above average risk to get additional diseases, and frequently require regular medication that also implies administrative aid. Individuals are assumed to attempt to communicate but are often not able to do so successfully, because of the inherent and external limitations [2].

A common denominator in the population we are dealing with is the limited ability to communicate coherently with their caregivers or the other way around. They usually communicate on a presymbolic level and their understanding of speech is severely limited. Some individuals have the ability to form joined attention with their communication partner [3] but this is not universal. While they are capable of learning, the acquisition of new skills takes significantly more time and requires frequent repetition. People with PIMD tend to exhibit not or not easily readable behaviours in order to communicate their (dis-)pleasure or to get attention. Examples of these are pushing unwanted objects away, loud vocalisations or banging to gain attention. In general, the communication attempts of people with PIMD are relying on caregivers who have been trained to understand their communications by interpreting their whole body behaviour or specific personal expressions [4].

Communication attempts are very multifaceted based on the specific individual. Some persons are capable of vocalising simple words, such as saying "Hi", but do not consistently use them in a correct manner or they grab towards toys and individuals they want to interact with. Making eye contact is possible for some individuals, which can be an orientation towards desired objects as communication attempt. Others do not have any coherent vocalisations, lack motor skills and require help holding items [3], [5].

Further complicating their behaviour or the interpretation of is stereotypy. These are actions that do not contain communication attempts but can be considered "ticks" that do not carry meaning. There is a correlation between stereotypical behaviours and low level of social interaction and stimulation. This behaviour can escalate over time to aggressive behaviour and sometimes even self-injuring. The level of these problematic behaviours seems to be correlated with communication problems [5] and would presumably point to this being an expression of frustration. These behaviours can range from hand wringing, to hitting legs, head or nearby objects, from purposeful breath holding to screaming etc.

Several attempts were made to bridge the gap of communication of people with PIMD. Some individuals have access to switches that produce specific sounds enabling an easier communication for less experienced communication partners [4], [6]. These switches can take several forms form simple push buttons to systems that attach to the individual muscles. Other systems, such as Picture Exchange Communication where the individuals with PIMD are expected to provide a picture, usually on a card, for the desired communication or Simplified Signalling that draws the inspiration form natural gestures that are taught to the people with PIMD so that their communication is in line with general public [6]. There is also research in using Brain Computer Interface for communication [7] that uses Electroencephalography (EEG) to map the activity of the brains to interpret the desires. The system requires adaptation based on the individual and training of the individual.

3. Analysing the Human

Enabling computers to interpret the desires of its users and their state of mind is a longstanding goal of computer science. The most developed systems focus on speech recognition[8], [9], but other systems are also explored. Eye tracking [10] is becoming more and more robust, moving from specialised hardware to simple web cameras [11].

More advanced systems enable facial features extraction [12], and from this psychological state of the individual can be extrapolated [13]. These systems work extremely well on typical individuals in good conditions, for example in good lighting and direct camera position [14]. In more dynamic conditions, such as unstable lighting these results are less certain but still reach acceptable levels of accuracy. Additional information, such as voice inflection or contact sensor provides additional information and greater accuracy. Research has been conducted in extracting stress of students using only the smartphone, carried by the person[15].

Another advancement in computer vision is the possibility of mapping the body parts of one or more persons from video [16]. This enables extracting of the limb and torso position, the position of fingers and facial markers. This enables researchers to qualify the position of a person and increases the robustness and ease of analysing behaviours of humans. An example of this can be seen in Figure 1.

The position of a person is only part of the puzzle of finding out what the person wants. In order to determine the context of the interaction there is another piece of the puzzle. The environment of the person must be taken into account. There are several systems that can take a video and return the objects present in the scene. One of the fastest open source solutions is YOLO [17]. Other commercial systems also exist, such as Google or Amazon Object recognition API [18].

4. The Expressions of the People with PIMD

In our research we are working with six people with PIMD in order to provide a robust data set that will be used to train the system and extract the state of the individual and the communication they are trying to accomplish. Based on this we hope to extract some rudimentary information, such as their psychological state (pleasure, displeasure or neutral) and the mode of communication they are exhibiting such as protest, demand or comment. Based on this information and the context – the activity that was happening before, the objects that are available for interaction and the estimation of internal needs based on models such as hunger, thirst etc. – the system will propose the action the caregiver should take.

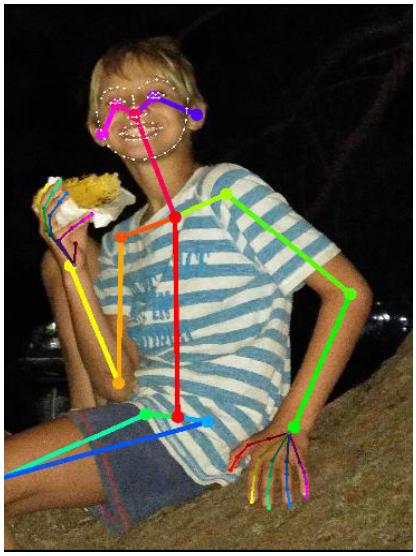

Figure 1. An example of the image processed by OpenPose. The Fingers, arms and facial characteristics are extracted and returned as points

The first step for this is collecting the data. In our case this video is manually annotated in order to provide information for the system. We use several cameras, an infrared camera, a wristband to collect physiological parameters and microphones that collect sounds. The videos are collected in several interactions ranging from meal time, playing (Figure 2) and physical therapy and even some life-skills training.

The videos are annotated to indicate the position of the individuals so that the system can be trained to return the desired information such as arm position, facial expressions and actions such as rubbing parts of the body, interacting with objects or people, presence of people in the scene and any external disruptions, such as loud noises etc. In Figure 3 we can see an example of the resulting annotation.

Figure 2. An image of video recording of a play session with a caregiver

5. Overviewof the Proposed System

Our system will take the information collected from the camera, microphone and other sensors, extract the objects in the scene and the information about the communication attempts of the users and, based on this, provide a guess on the presumable mood and communicated content.

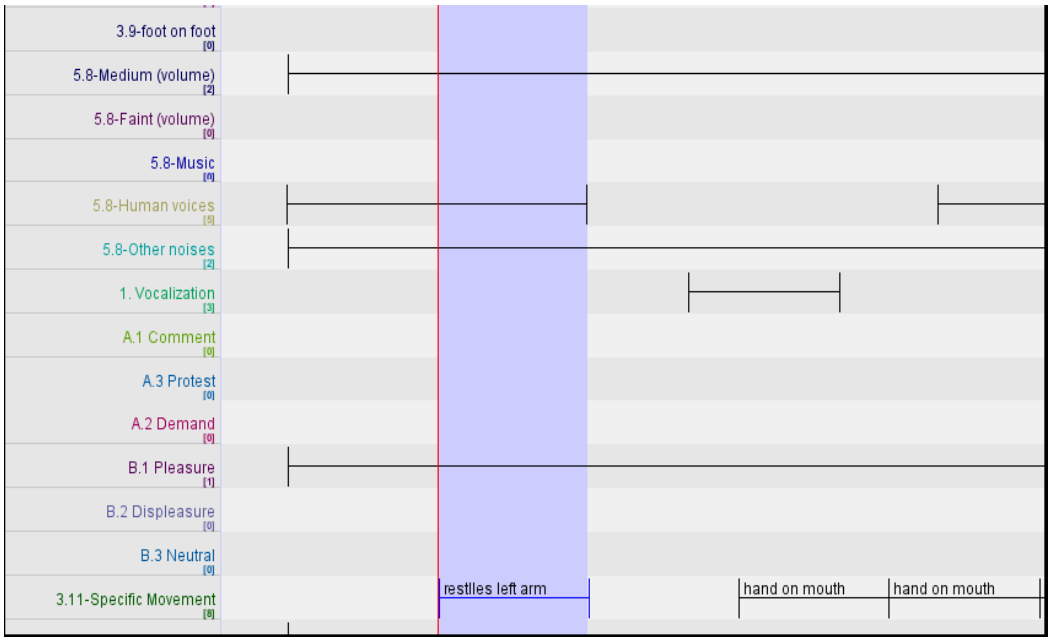

Figure 3. An image of video recording of a play session with a caregiver

In the first round, communication classes will be extracted using unsupervised learning that will return common actions performed by the individual, as annotated or provided by the movement analysing system. This will return meaningful communication clusters. These will be simple like moving the hand from position A (on head) to position B (pointing), interacting with the desk, vocalisation or a combination of these, such as pointing and producing a distinct sound.

All these communication clusters can then be mapped onto the state and communication system. At this stage supervised learning will be used. Some actions will for instance indicate that the user is displeased - providing the information on the internal state of the individual. While others will indicate communication attempts such as demand for something. Together they can indicate to the caregivers what action they should take to provide care for their charges.

The first problem we have to solve is to remove the behaviour that does not carry any communication information. In general, it is expected that action that happen regardless of the state of the individual will be filtered out. This could become problematic as some actions can have several meanings. As such the efficacy of this is considered ongoing research.

In addition, the context, such as presence of strangers or objects in the vicinity, will be used to provide information on the desires and aversions of the individual, creating an internal database of likes and dislikes that can be taken into consideration.

6. Conclusions

As the system is not yet operational, there are several problems, that are still part of research. For instance, it is unknown which role stereotypy will play, or how accurately we can extract the information if the user is interacting with a specific object or person. The information that the user is interacting with one object may not be enough to infer the actual desire of the user.

Once this, and other problems are resolved the system can be extended to general public. Enabling researchers and others to analyse the behaviour of the individuals and extract the communication desires and psychological state of the individual to

further their understanding of the motivations and desires of people, providing standardised analysis of their movements and philological states. This will in turn enable greater rigidity of inquiry providing a faster and reproducible way of analysing behaviour of people.

However, the system could be misused. Organisations and individuals could use the system to determine the state of the individual and use this to manipulate her or him for their gain. A system that can extract the information could be used to associate some products and people with this feeling, thus steering the people to the competition. Furthermore, body language could be important for determining not the information but relationships of people, creating an internal map that could result in an interaction map that could be used to influence certain groups of people or simply to determine the connections between them and their reactions to certain events.

7. Acknowledgments

Funding: This work was supported by the European Union's Horizon 2020 research and innovation program [grant agreement No 780819 (INSENSION)].

References

[1] Nakken, H., Vlaskamp. C. (2007). A Need for a Taxonomy for Profound Intellectual and Multiple Disabilities, *J. Policy Pract. Intellect. Disabil.*, 4 (2), p 83–87, June 2007.

[2] Atkin, K., Lorch, M. P. (2016). An ecological method for the sampling of nonverbal signalling behaviours of young children with profound and multiple learning disabilities (PMLD), *Dev. Neurorehabil.*, 19 (4), p 211– 225.

[3] Yoder, P. J., Warren, S. F. (2001). Relative treatment effects of two prelinguistic communication interventions on language development in toddlers with developmental delays vary by maternal characteristics., *J. Speech. Lang. Hear. Res.*, 44, (1), p 224–237.

[4] Harding, C., Lindsay, G., O'Brien, A., Dipper, L., Wright, J. (2011). Implementing AAC with children with profound and multiple learning disabilities: A study in rationale underpinning intervention, *J. Res. Spec. Educ. Needs*, 11 (2), p 120–129.

[5] Me. Laura Roche. (2017). Evaluating and Enhancing Communication Skills in Four Adolescents with Profound and Multiple Disabilities, Victoria University.

[6] Goldbart. J., Caton, S. (2010). Communication and People with the most complex needs: what works and why this is essential, *What is communication and what are complex communication needs?*, July. p 3–4.

[7] Adeli, H., Rahimian, P., Tabrizi, N. (2016). Communicating with People with Profound Intellectual Disabilities Using Brain Computer Interface, *J. Technol. Pers. with Disabil.*, 4, p 133–145.

[8] Benzeghiba, M., *et al.* (2007). Automatic speech recognition and speech variability: A review, *Speech Commun.*, 49, 10–11, pp. 763–786.

[9] Anusuya, M., Katti, S. (2009). Speech recognition by machine: A review, *Int. J. Comput. Sci. Inf. Secur.*, vol. 6, no. 3, pp. 181–205.

[10] Morimoto, C. H., Mimica, M. R. M. (2005). Eye gaze tracking techniques for interactive applications, *Comput. Vis. Image Underst.*, 98 (1), p 4–24.

[11] San Agustin, J., *et al.*(2010). Evaluation of a low-cost opensource gaze tracker, *Proceedings 2010 Symp. Eye-Tracking Res. Appl. - ETRA '10*, p. 77, 2010.

[12] Chu, W. S., De La Torre, F., Cohn, J. F. (2017). Selective transfer machine for personalized facial expression analysis, *IEEE Trans. Pattern Anal. Mach. Intell.*, 39, (3), 529–545.

[13] Yin, Z., Zhao, M., Wang, Y., Yang, J., Zhang, J. (2017). Recognition of emotions using multimodal physiological signals and an ensemble deep learning model, *Comput. Methods Programs Biomed.*, 140, 93–110.

[14] Gunes, H., Hung, H. (2016). Is automatic facial expression recognition of emotions coming to a dead end? The rise of the new kinds on the block, *Image Vis. Comput.*, 55 (1), 6–8.

[15] Gjoreski, M., Gjoreski, H., Lutrek, M., Gams, M. (2015). Automatic Detection of Perceived Stress in Campus Students Using

Smartphones, 2015 *Int. Conf. Intell. Environ.*, p 132–135.

[16] Wei, S.-E., Ramakrishna, V., Kanade, T., Sheikh, Y. (2016). Convolutional Pose Machines, in *Conference on Computer Vision and Pattern Recognition*, 2016.

[17] Redmon, J. S. D. R. G. A. F. (2016). (YOLO) You Only Look Once, *Cvpr*.

[18] Buyya, R., Yeo, C. S., Venugopal, S. (2008). Market-oriented cloud computing: Vision, hype, and reality for delivering IT services as computing utilities, *Proceedings 10th IEEE Int. Conf. High Perform. Comput. Commun. HPCC 2008*, *p* 5–13.