Journal of Intelligent Information Systems (2019) 53:381–404
https://doi.org/10.1007/s10844-019-00547-y

Database of speech and facial expressions recorded with optimized face motion capture settings

Miłosz Kawaler[1] · **Andrzej Czyżewski**[1]

Received: 29 November 2018 / Revised: 19 January 2019 / Accepted: 22 January 2019 /
Published online: 21 February 2019

Abstract
The broad objective of the present research is the analysis of spoken English employing a multiplicity of modalities. An important stage of this process, discussed in the paper, is creating a database of speech accompanied with facial expressions. Recordings of speakers were made using an advanced system for capturing facial muscle motion. A brief historical outline, current applications, limitations and the ways of capturing face muscle motion as well as the problems with recording facial expressions are discussed. In particular, the scope of the present analysis concerns the registration of facial expressions related to emotions of speakers which accompany articulation. The camera system, instrumentation and software used for registration and for post-production are outlined. An analysis of the registration procedure and the results of the registration process was performed. The obtained results demonstrate how muscle movements can be registered employing reflective markers and point at the advantages and limitations of applying FMC (Face Motion Capture) technology in compiling a multimodal speech database. A short discussion pertaining to the usage of FMC as ground truth data source in facial expression databases concludes the paper.

Keywords Motion capture · Facial expressions · Registration and post-production

1 Introduction

Since speech corpus is necessary for developing speech recognition or speech synthesis algorithms, thus multiplicity of speech signal databases was created. The development of automated lip reading technology was preceded by construction of many audio-visual or visual speech corpora. One of such database, called “Modality corpus” was prepared in Gdańsk University of Technology, specifically to assist audio-visual speech recognition systems (AVSR) development (Czyżewski et al. 2017). It contains more than 30 hours of recordings of English speech including high-resolution, high-framerate stereoscopic video

✉ Andrzej Czyżewski
ac@pg.gda.pl

[1] Multimedia Systems Department, Gdańsk University of Technology ETI Faculty, Narutowicza 11/12, Gdańsk, Poland

streams from RGB cameras, depth imaging stream utilizing Time-of-Flight camera accompanied by audio recorded using both: a microphone array and a microphone built in a mobile computer. However, further development of speech recognition moves towards the allophonic level which is necessary for automated speech transcription to IPA (International Phonetic Alphabet). Moreover, advanced speech resynthesis may use components from an existing signal for the purpose of transmitting emotions which voice often conveys. Therefore, to assist research development in this area we decided to extend the Modality corpus accessible at the Web address http://www.modality-corpus.org/ by adding Face Motion Capture recordings presenting facial expressions and speech along with respective audio recordings. Consequently, to the recorded speech sounds a representation was added of facial images reflecting emotions accompanying speech production. Many databases containing facial expressions can be found on the Internet, as listed by Wikipedia in the page entitled: "Facial expression databases", which shows 16 databases (as accessed on Nov. 30th, 2018), together with 18 references to papers related to these resources. However, some subtle movements of facial muscles can be reflected most precisely by capturing motion of reflective markers attached to speaker's face (Le et al. 2013; Reverdy et al. 2015). The recordings acquired in this way might be used as a source of ground-through data, because motion capture systems allow users to track minute changes in the position of objects (in real time or through the analysis of recordings). The possible captures include the general location of the body, the motion of individual limbs, and minor changes in particular muscle groups. Such an accurate mapping allows, for example, the transferring the motion of a living actor to a three-dimensional model in a computer animation, examining bone, muscle or respiratory diseases, as well as designing speech recognition systems based on lip motion. Before the goal of this work is explained, the history and the current development of this technology will be outlined first.

1.1 Historical view

The paper *Emotional head motion predicting from prosodic and linguistic features* (Jiang et al. 2016) presents the possibility of predicting emotional motion of the facial muscles based on such parameters as the length of spoken words, the type of part of speech, tone of voice, or stress. Without the motion capture system, it would be impossible to create such a prediction model. Due to the fact that there are about 206 bones in the human body, eight axes of rotation are used in the simplified model of motion (Zarins and Kondrats 2014). The more complex the muscle tissue is, the greater the possibilities of expression and the diversification of dynamics. In addition, the change in the position of one part of the body affects other and there is a relationship between the mood, condition and attitude of the person (Williams 2001). This means that capturing motion is a complicated and technically demanding process.

In the past, rotoscopy was the primary technique. It was based on the mapping of the motion of real actors playing in the film, by cartoonists analyzing the recordings. One of the most well-known works that was realized in this technique was the animated film from 1937 (Aloff 2013). In the 1980s, a system based on goniometers was created. When changing the angle in the joint, there was a change in the value set in the given potentiometer, which modified the position of the computer limb of the model composed of lines (Calvert et al. 1982). The next step was optical systems such as e.g. Op-Eye. On the basis of the camera image, the position of the light emitting points was detected. There were numerous problems associated with technological limitations due to which it was impossible to transfer a human face to a three-dimensional digital model (Maxwell and Ginsberg 1984). In

1988, Brad deGraf and Michael Wahrman presented the "Mike the Talking Head" project at the SIGGRAPH conference for the first time. It was a spatial model of a human face, animated in real time. "Mike" actively participated in the conference and reacted to events, which aroused great interest (ABCP channel 2018).

Currently, motion capture systems are also used in medicine. The examination of the patient's motion is primarily a non-invasive method. It is possible to diagnose chronic obstructive pulmonary disease (Hasegawa et al. 2016) or bone system damage (Hasegawa et al. 2016). In addition, motion capture systems allow to identify a person based on his or her movements; either the way of walking (Sulovska et al. 2017), or the movement of mouth, e.g. when articulating a password (Hassanat 2014). One of the oldest and currently the most common applications of motion capture systems is the animation of characters in films such as "The Lord of the Rings: The Fellowship of the Ring" (Nature video channel 2018), "Rise of the Planet of the Apes" (WIRED channel 2018), "Avatar" (Media Magik Entertainment channel 2018) or "The Hobbit: The Desolation of Smaug" (ABCP channel 2018). It is not unlikely that in the future, the prizes for the best acting will be awarded for roles performed using motion capture systems (Menache 2011). In addition to cinematography, the video game industry shares great interest in motion capture systems. Current capabilities of generating ultra-realistic graphics combined with authentic animations based on human behavior enable the creation of increasingly immersive and realistic-looking games. The games in which motion capture systems played a key role are e.g. "LA Noire" (Noire 2018), "Beyond: Two Souls" (PlayStation channel 2018), "Call of Duty: Advanced Warfare" (GameCrate channel 2018) and "Hellblade" (2018).

New examples of deep convolutional neural networks applications show cases of accurate recognition of emotional features from speech audio signal (Zhang et al. 2018) and from video of a face (Kahou et al. 2016). Combining sound analysis and video analysis in a multimodal approach can result in increased accuracy of utterance sentiment classification, personalized for a particular speaker (Vryzas et al. 2018). The emotional information extracted from the speech can be creatively utilized in an artistic visualization. An interesting concept of visual changes of the stage synchronized with a live acting was proposed in works of Vryzas et al. (2018).

An interesting trend visible in the literature concerns visual prosody. Since people naturally move their heads when they speak, thus head motion conveys linguistic information. Three-dimensional head and face motion and the acoustics of a talker producing Japanese sentences were recorded and analyzed previously in this context (Munhall et al. 2004). The inherent relationship by building the mapping model between head motions and Chinese speech prosody and linguistic features has been also studied more recently (Minghao et al. 2016).

A long-standing problem in marker-based facial motion capture is what are the optimal facial mocap marker layouts. The subject of optimization of characteristic control points has been studied in the literature before (Le et al. 2013) and it provides also one of topics of the present paper.

1.2 Motion capture technologies

To effectively transfer the movement of the body to the digital three-dimensional space, different solutions were used. Three types of motion capture systems can be distinguished. This division is based on the placement of the sensors and the source of the signal being processed. The external-internal system is equipped with receivers outside the tested object, and the signal sources are attached to this object. Optical systems as such are a type of

system where cameras register the location of markers emitting or reflecting light. The next type is an internal-external system. It is the opposite of the external-internal system. It has emitters located in the space around the object and receivers on it. Electromagnetic systems can be an example. The generator of the electromagnetic field is located "outside" and the sensors measuring the change of its intensity depending on the position, on the body of the actor. The internal-internal system is composed entirely of elements that are located on the registered object. Such systems are often enclosed in a specially designed costume. Sensors are, for example, potentiometers, accelerometers or goniometers, and the human body is the generator that causes changes in the parameters registered (Parent 2009). Over the years, many solutions have been developed for the capturing of object motion. We can distinguish solutions such as: acoustic, mechanical, electromagnetic, magnetic, optical or electromechanical systems. Significant factors influencing the usability of the systems were precision, practicality, number of samples or frames per second and reliability (Nogueira 2011). Below, the most commonly used solutions, i.e. optical and electromechanical, will be briefly discussed.

Optical systems are suitable for working primarily in confined spaces. They belong to external-internal systems. Sensors are in this case special cameras, and active or passive markers act as generators. The signal from the recorders is transferred to the software in the computer, where on the basis of the trajectory of particular points seen by at least two cameras, their location in three-dimensional space is determined (Schulz 2010). Cameras record high resolution image with a high frame rate. They operate in the field of infrared waves, thanks to which one can obtain a good read accuracy of markers, which can be covered with a special coating that reflects such waves well or be emitters in the form of LEDs. The advantages of such systems are the great comfort for the actors and the good precision associated with high resolution and a high frame rate. In the case of facial motion capture, this is currently the most common method of registration due to its effectiveness. Its disadvantage is the problem with the markers visibility. If a tag is covered or too close to another marker, it becomes impossible to determine its position in space. This causes problems when capturing motion of a number of persons, enforces the use of multiple cameras at the same time and the need to arrange markers at possibly large mutual distances (Parent 2009).

Mobile electromechanical systems are also often used. They are used especially when motion capture needs to take place outside the recording studio. They are composed of costumes, on which various types of sensors are mounted in the most important places for correct animation (most often these are joints). Before each capture, calibration is necessary to determine the entire body reference point. At the beginning of the capture, each sensor transmits data about the change relative to the initial state. The advantages of such a system is the ability to capture motion anywhere. Additionally, the risk of signal loss from one of the markers is much smaller than in the case of optical systems. It is possible to simultaneously register the motion of many people at the same time without disturbances. The disadvantages of such a solution are the discomfort associated with the need to mount the active equipment on the suit and reduced accuracy of measurements. Another difficulty is the calibration required before each capture (Barsegyan 2017).

Due to the fact that the optical system is used in the presented studies, further considerations will therefore apply to this type of technique.

1.3 Aim of experimental research

The experiments conducted began with the recording and then post-production of recordings made using a motion capture system (manufactured by Vicon 2018), together with sound.

The recordings reflect facial expressions of people who speak freely or read a passage of text, express emotions, recite, sing, beatbox, or make a pantomime. The aim of the experiment was to show the effectiveness of optical motion capture system, when registering such complex and diverse phenomena as human speech, facial expressions, as well as unconventional ways of using facial muscles. In addition, a template was created for the description for the distribution and unambiguous naming of markers placed on the faces of the persons being recorded. Next, an analysis of the recording process and its results allowing to formulate conclusions regarding the problem of registration of facial expressions with the use of motion capture systems was carried out.

2 Experiment preparation

An important element of the described research experiment was the extraction of changes in the position of particular points on the face of registered persons. To obtain varied and interesting data in terms of research, appropriate design assumptions had to be made. An important issue was to find a way to determine the optimal placement of markers, so that it would be possible to reflect the facial expressions of recorded persons as true as possible. It was also necessary to specify the scenario of each of the recordings and the post-production method.

2.1 The scope of the research experiment

Conducting the experiment covered four stages. The first of them was to create a master record, based on which the template for labeling markers was defined, and the decision about their number and spacing was made. Proper preparation of this part allows to minimize the need for post-production of subsequent recordings and ensures the best mapping of human facial expressions. Part two was the recording of five persons. These persons were asked to present examples of expressing emotions, free speech, reading a given text and presenting skills such as beatbox, singing and pantomime. Then, in the third part, post-production of recordings took place in order to remove redundant data, as well as to repair and interpolate the data that are needed to reproduce the facial expressions. Then, the data obtained were analysed.

2.2 Labeling and placement of markers

The high complexity of a human face does not allow for perfect mapping it in virtual reality. For this reason, the most important points are chosen which are decisive in the transmission of emotions or speech. When the recording is made (in this case using the Vicon system (Vicon 2018) available in our laboratory), each marker receives a clear labelling. There is also the possibility of creating links between individual markers in the form of lines, in order to enable a convenient visual analysis. A big problem in the process of position extraction during the entire recording are the temporary loss of markers in the fields of camera visibility or too close proximity of two neighboring markers. This results in temporary loss of data, which causes the markers to miss their label assignment for the remaining recording time. By means of manual editing, it is possible to recover them, and using the interpolation process, based on two closest recorded positions (preceding and following) one can determine their approximate location in the moments of visibility loss. However, this reduces the quality of the recording. Therefore, it is necessary to create such a pattern of distribution,

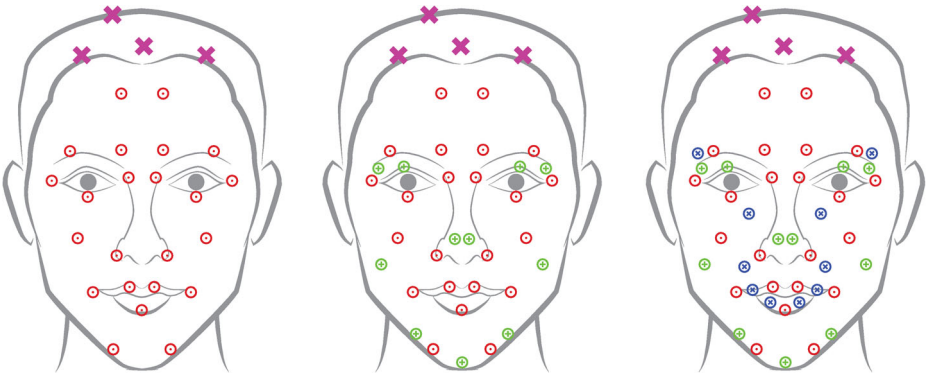

Fig. 1 Three templates for the distribution of markers on the face **a**) layout of 23 sticking markers on the face plus 4 attached (with Velcro) on the cap, **b**) layout of 34 glued markers on the face plus 4 on the cap, **c**) layout of 44 glued markers on the face plus 4 on the cap

so that it would allow for good reproduction of speech and emotions and it will not induce problems in post-production. A higher number of markers enables better nuances of mimic, but it increases the likelihood of marker fading and error occurrence. Taking into account the results of the work entitled: *Marker Optimization for Facial Motion Acquisition and Deformation* (Le et al. 2013), three templates of distribution of markers on the face were designed, which are presented in Fig. 1.

The layouts use in turn: 23, 34 and 44 markers, which should be placed on the face of the person being recorded, allowing for the elementary mapping of facial expressions. Four additional markers, which are fastened with Velcro to the cap, are also included in each of the templates. To unify the nomenclature, the layout of Fig. 1a was named 23 + 4, in Fig. 1b is defined as 34 + 4, and in Fig. 1c, as 44 + 4. The first number corresponds to markers glued to the face and the second number to the markers placed on the cap for the recordings. In the version with 34 and 44 markers, some markers are additionally placed to reproduce the movement of the upper eyelid. The placement takes into account the most important places on the face in terms of expressing emotions, such as eye sockets and lips. For this reason,

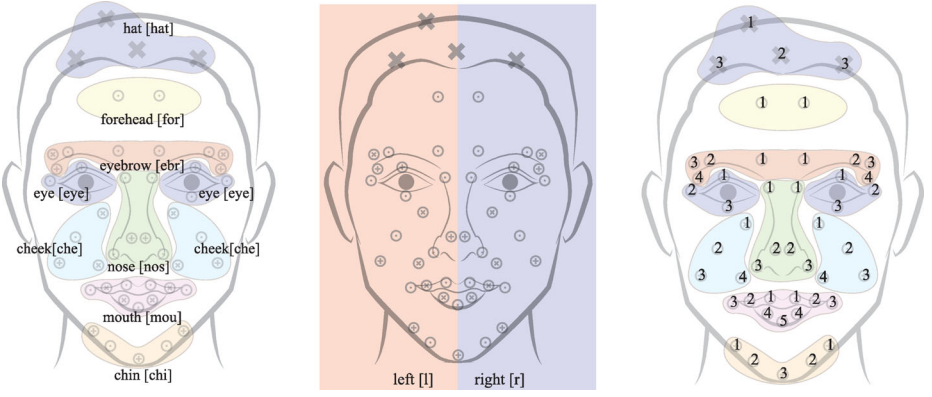

Fig. 2 Basic assumptions of naming of markers **a**) division into areas of the face, **b**) division into sides of the faces, **c**) example numbering of markers

templates with a larger number of markers, reflect the movement in these areas more accurately. The decision on the template used will be taken based on the quality of the motion mapping and the complexity of the problems faced during the post-production of the first, reference recording. Each marker must have a label assigned that uniquely identifies it. To obtain a clear division of these labels, the face has been divided into areas that can be identified on the basis of a three-letter abbreviation. This has been presented in Fig. 2a. This division was combined with a distinction between the face (left [l], right [p]), presented in Fig. 2b. In the case of markers that belong to both sides, because they are located in the middle, it was necessary to abandon this designation. To determine each of the markers, it was necessary to additionally number the markers within particular areas and sides. An example numbering is presented in Fig. 2c.

Thanks to this division, examples of the signs: for_l_1, ebrebr_r_3, moumou_5 and chichi_3 can be obtained. The final names were given after selecting the appropriate number and location of all markers.

2.3 The course of recordings

In order to create recordings that will contain interesting content in terms of the possibility of analysis, it was necessary to prepare a recording plan. It has been divided into four stages and presented in Table 1.

Each recording should contain a part shared by all speakers as well as a characteristic element. The first stage is to present the person, which allows one to observe how facial expressions work throughout the face when information is passed freely. Then, to acquire a similar result for each person in terms of content, each of them should read a part of the same text. This allows to compare discrepancies in the articulation of individuals when articulating the same content. The next stage is the presentation of five emotional expressions. They are: neutrality, joy, sadness, surprise and anger. At the end, each person presents different ways of expressing emotions or content. This stage is important due to the introduction of a sufficiently large variation in the recorded data. Planned activities that will be performed by the persons being recorded are: reciting a text from memory and reading it by a person who previously did not know this content, singing fragments of songs, beatbox or pantomime. The first two recordings are meant to find how the content presented depends on the effect of sentence memorizing. Registration of singing or beatbox allows one to compare effects to the free speech, because singing or beatbox is a specific use of the human speech apparatus. The pantomime (facial gestures recorded without speech) is compared to the five exemplary expressions from the third stage.

2.4 Post-production

The material in the Vicon system was recorded as a video image from six cameras working in the infrared spectrum and one in the visible for humans wavelength range. It was

Table 1 Recording plan

	Stage 1	Stage 2	Stage 3	Stage 4
Activity to perform	Free speech	Reading text	Presentation of five emotional expressions	Characteristic element for the examined person

necessary to reconstruct the markers. Thanks to this, on the basis of the view from several different cameras, their position in three-dimensional space is determined. After completing, each marker must be explicitly named by assigning labels. If a template was previously defined, one can use it to speed up the work. Thanks to this, the program automatically assigns corresponding names to markers in the positions defined by the template. In this project, this template will be drawn up in accordance with the assumptions contained in sub-chapter 2.2. The next stage is finding the moments in which the markers were invisible to the system. This causes them to lose their assignment to their labels. After re-entering the field of visibility of the camera set, they are also treated as completely new, other markers. This requires assigning the label again. When all markers containing such moments are properly machined, interpolation of positions is necessary based on the position before and after the loss of visibility. The longer this moment and the more dynamic the movement, the greater the likelihood of poor mapping. Therefore, it is important to analyze the recordings immediately after registration. The earlier creation of a labeling template significantly simplifies the whole process, because in the case of losing the visibility of a given marker, the next one that appears in its place has an automatically assigned label. However, this does not solve the problem of the lack of data at the given moment.

3 Implementation

This chapter aims to present implementation work carried out on the basis of a reference recording. The types of errors that may appear during post-production and how one can avoid or fix errors are described here.

3.1 Software and hardware used

The recordings were made using the Vicon system (Vicon 2018) and the Blade software included in it, as well as a sound recorder with a microphone and markers. The layout of the arrangement and the connection of the entire setup is presented in Fig. 3.

The Vicon system consists of a special, stable frame located in front of the person to be recorded, on which cameras are installed, of which there are a total of seven; six of them are Vicon Vero cameras with a 2.2 Mpx matrix (2048 × 1088 px), working at the maximum frame rate of 330 frames per second. The cameras record infrared waves, which facilitates the separation of markers from the background without distracting the registered persons. Other systems available on the market offer cameras with frame rate up to 360 fps (Motion Analysis 2018; OptiTrack 2018; Tracklab 2018). This number is about 15 times greater than the refreshing value, during which a person perceives motion as fluid. This redundancy, however, brings benefits in post-production, through increasing the use of recorded data. Too low frame rate in the extreme case can cause that very fast movement will be registered only in the initial and in final phase. The seventh centrally located camera is a Vicon VUE with a 2.1 MPX (1920 × 1080 px) matrix with a maximum frame rate of 120 frames per second that records the image in the visible wavelength range for humans. The cameras are connected to a device that synchronizes their work and transmits the image to a desktop computer. The image registration stand has been presented in Fig. 4.

Registration and post-production are carried out using Blade software, produced by Vicon. It allows both observation of the image from each camera (Capture view), reconstruction, as well as analysis of the recordings made. The perspective view makes it possible to observe the mapping of points in three-dimensional space, the Graph view shows changes

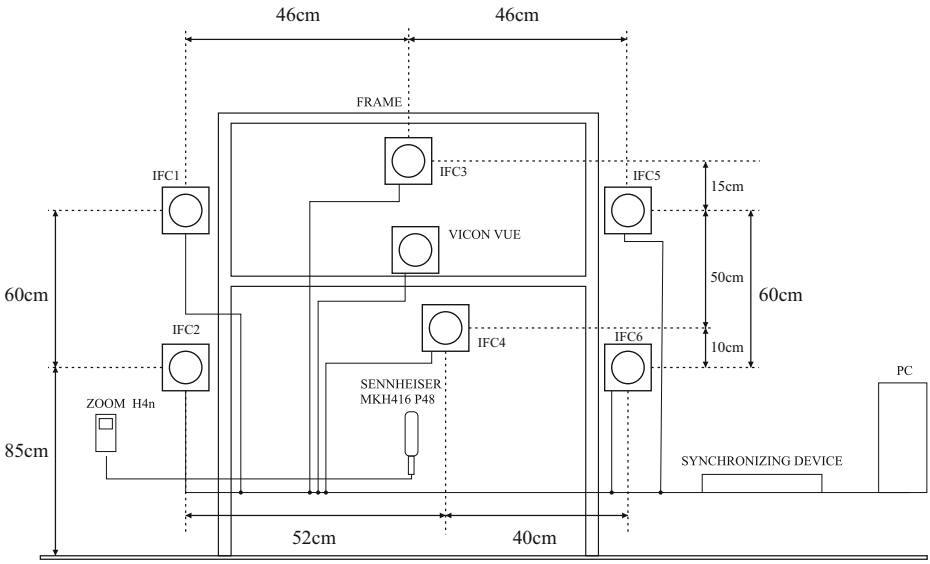

Fig. 3 Arrangement and connection of devices used for experiments. IFC 1-6 represent infrared cameras for tracking reflective markers

in the position or rotation of individual markers in time using charts. The Data Health window allows to check the continuity of markers recordings in time. In addition to recording the vision, sound was also recorded. A condenser microphone, interference microphone Sennheiser MKH416 P48 and a recorder Zoom H4n were used. A device for sonic synchronization of the start of image and sound recording was also used, as well as an instrument

Fig. 4 A system of seven cameras to capture motion (Vicon 2018)

to calibrate the entire system and to establish the start of a virtual coordinate system (the so-called wand with active LED markers). The OptiTrack markers applied to the face had a diameter of 4 mm and a shape of half-sphere. On one side, they were covered with reflective foil, which reflected the light well, and on the other with a cosmetic glue that allowed them to be easily applied and peeled off. To capture moves of the whole character, bigger tags are used mounted on Velcro handles. They can be attached to the costume, which can be full or partial. One of the elements of such a costume is, for example, a cap, which was used during the experiment.

3.2 Standard recording

The recording took place using the three templates shown in Fig. 1. The examined person talked about himself freely and read the same text three times with different layouts of markers on the face. This allowed to compare the quality and repeatability of recorded data. The number of markers that temporarily disappeared during the recording was taken into account, which resulted in the separation between several other markers with different labels and those that appeared on one frame, referred to as blinks. The number of correctly registered markers was also analyzed.

The first type of errors, i.e. the momentary disappearance, caused the physical registration of one marker, using several different markers in three-dimensional space, as shown in Fig. 5. It shows a graph of the dependence of the change in the position of the markers on the subsequent frames recorded using the Vicon system. The X axis corresponds to the frame numbers of the recording, which has 5000 frames. The Y axis corresponds to the position values of the markers, which vary from 40 to 50centimeters horizontally with respect to the origin of the coordinate system. This designation also appears in Figs. 6, 7 and 8.

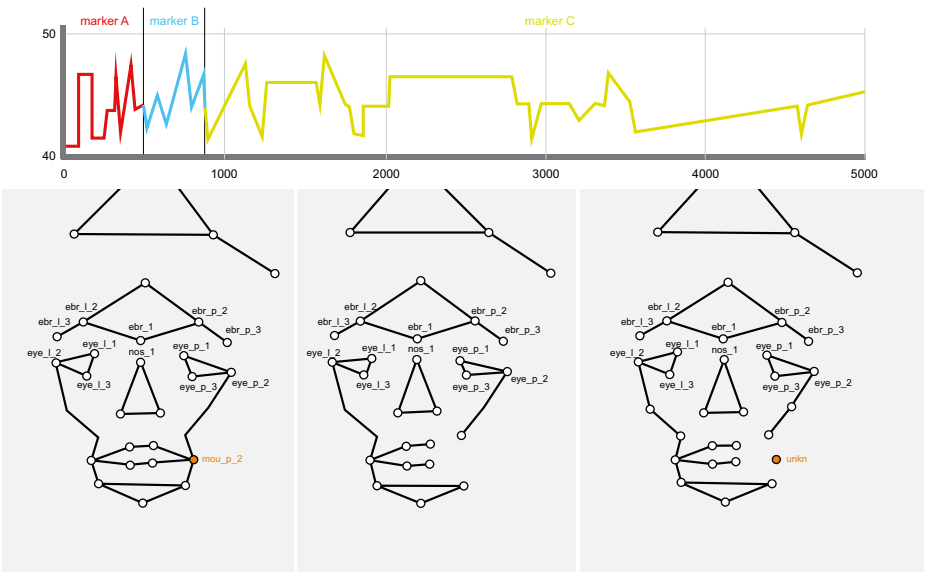

Fig. 5 Example of the discontinuity of marker registration. Top: marker position versus frame number. Bottom: previously visible marker mou_p_2 (left) is lost (middle). After the recovery of the marker (right) its name is unknown and must be manually fixed or derived from the template. The missing trajectory is interpolated

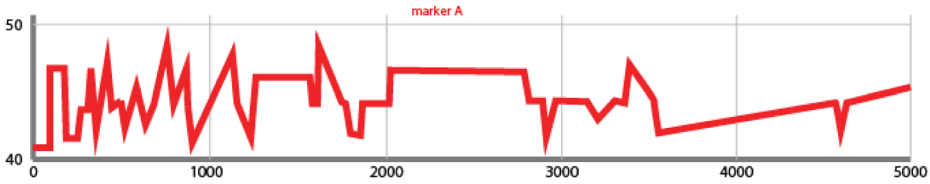

Fig. 6 Example of correct marker registration (marker position versus frame number)

In the range of frames with numbers from 0 to approx. 500 the marker A was a marker physically located on the cheek. Due to the temporary interruption of visibility, it was changed by the marker B. As a result, the label assigned to the marker A will only function for approx. 500 frames of recording. The same situation took place during the registration of the 900^{th} frame for the marker C. Repairing this type of errors consists of assigning the same label to all three markers (A, B, C), and then performing the interpolation of the position of the markers during the missing frames of the recording. Thanks to the pre-defined labelling template, there is a possibility of automatic naming, which greatly speeds up the post-production process. For comparison, the movement of a correctly registered marker was reproduced in Fig. 6.

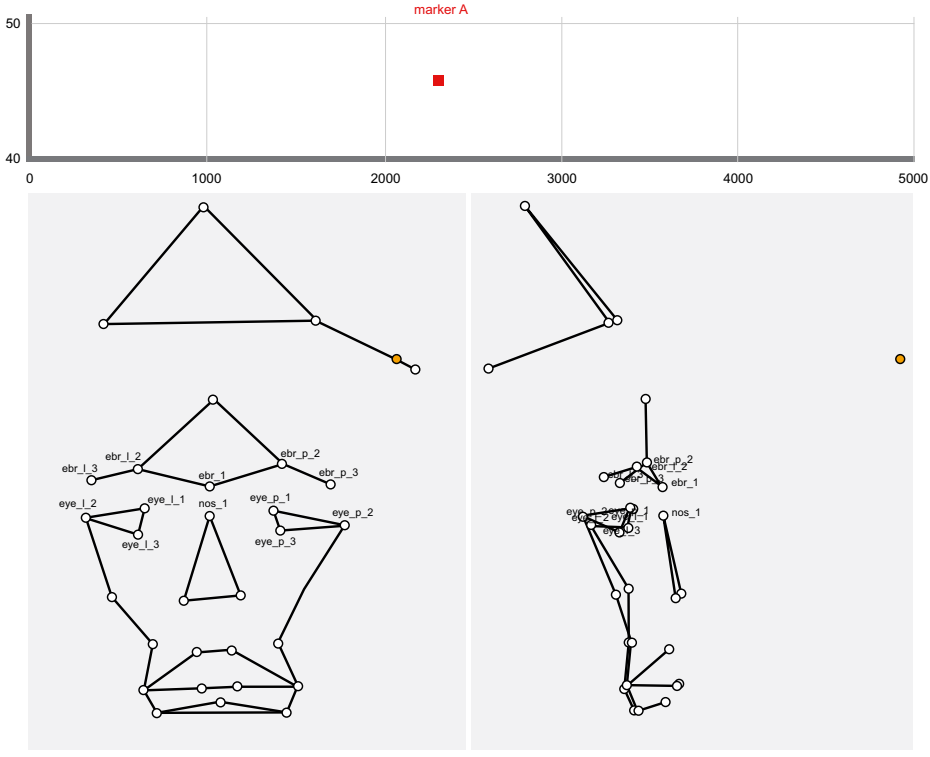

Fig. 7 Glint example (position versus frame number). Frontal and side view of the marker layout, glint marked in orange

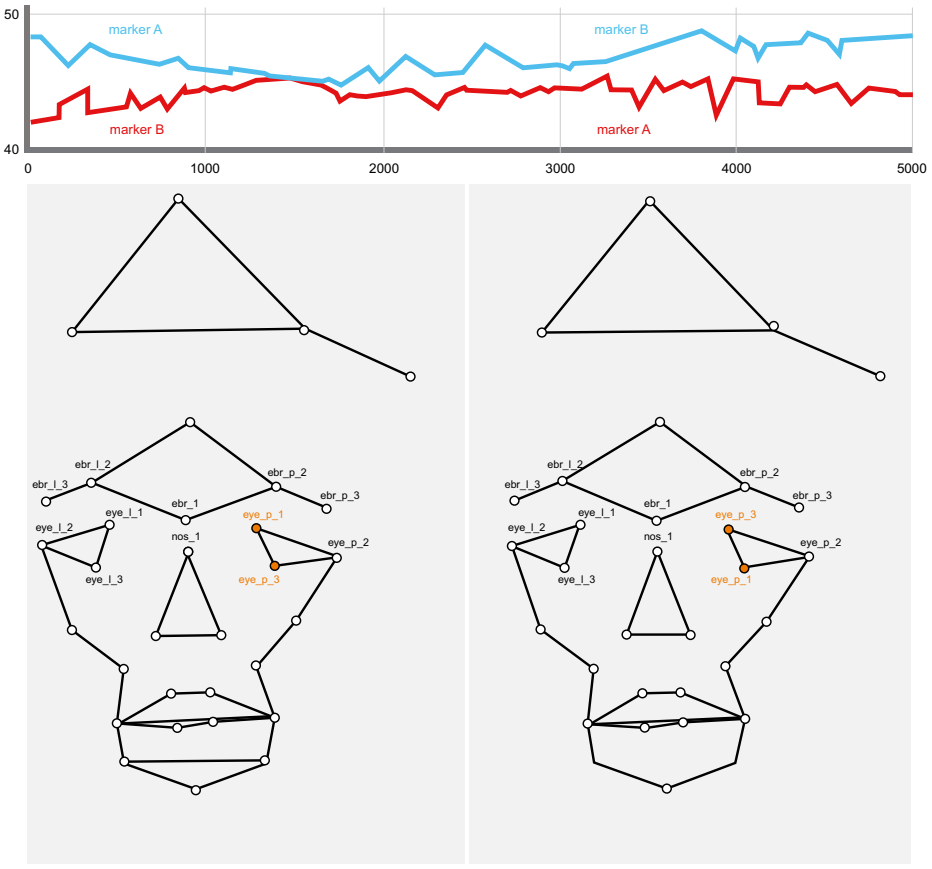

Fig. 8 Example of substitution of labels for two left eyelids markers (orange)

In such a recording there is no interruption of continuity and it is enough to only assign a label to the marker once. It will then function throughout the entire recording.

Another type of errors are reflections that appear in single frames, which are mistaken as a marker. Representation of such a phenomenon (a glint) is shown in the graph in Fig. 7.

This causes more markers to appear than there actually are. In Fig. 7 (frontal and side views) an example of a single glint is presented. It can be observed that the glint appeared in a considerable distance from the face and that it was registered by several cameras (it is highlighted in orange).

In addition to the errors already discussed, there are also substitution of labels. They occur when the two markers appear too close to each other. This causes that the previously assigned label changes to the neighboring marker, as it is depicted in Fig. 8.

This is a particular type of error, because it requires careful examination of the recording itself. It cannot be observed in the charts, because theoretically the label is present at all times. Detection is facilitated by lines connecting pairs of markers, because they are carried out in an inappropriate way in the event of such errors. During the implementation of the master record, such irregularities did not occur.

Recorded results yielded the summary presented in Table 2: the lowest number of discontinuous and glints was achieved for $34 + 4$ layout. Thus $34 + 4$ layout turned out to be optimal.

Table 2 The number of incorrectly and correctly registered markers in the reference recording

The number of markers	Layout 23 + 4	Layout 34 + 4	Layout 44 + 4
All registered	52	41	51
Being single-frame glints	12	0	3
Discontinuous	13	2	3
Correctly registered	27	39	45

This layout caused the least problems with post-production. In the case of a reduced number of markers, there were a lot of glints as well as temporary disappearances. The $44 + 4$ layout was also highly accurate. In his case, the biggest problems occurred at the beginning of the recording, because in the initial range of frames there were markers that were not physically present. For the correctness of the recorded data, it was also important how exactly the markers reflect the movements of the muscles on the face. In the case of the largest number of markers, some of them turned out to be unnecessary. It was possible to replace pairs of neighboring markers with a single, equally effective one. Therefore, the final template for the distribution of markers should be based on the $44 + 4$ layout reduced as much as possible, in consequence reducing the actor preparation time. It was assumed that the final number of markers should not exceed about 30 (ca. 1 minute per marker resulting in up to half an hour total time). Both in the case of speech and the expression of emotions, the areas of the eyes and lips showed the largest changes in the position.

3.3 Optimal placement of markers

Based on the completed reference recordings, the final pattern for the placement of the markers has been established. Their number was minimized in such a way that they would reflect facial expressions well. An additional recording was also used, in which five types of highly expressed emotions were presented. This allowed to determine approximate, maximum ranges of deflections of individual markers. It is important to properly position the marker representing the blinking eyelid. If the marker is located too close to the eyebrow, it is covered and the registration is interrupted. In Fig. 9 the final spacing of all markers is presented. There are 28 markers attached directly to the face and 4 attached with Velcro. The layout is denoted as $28 + 4$.

In addition to reducing the number of markers, their position has been corrected. Eye_r_1 and eye_l_1 have been moved to a place where they are better visible under the eyebrow. In this way, the prepared template can be adapted to enable a convenient analysis by combining markers. It was decided that the joins would be implemented in such a way that they would help interpret the obtained results, as it was presented in Fig. 10.

In this way, the prepared labelling template was created and exported to a file containing the first person's recordings, in turn, allowing for the acceleration and automation of post-production.

It must be stressed here that the manual physical placement of the facial markers to a new actor will always cause small geometrical deviations from the model, depending on the person face geometry.

In fact, it is required to adapt the model to each new face and to follow general guidelines of markers placement, in order to match positions of mimetic muscles and their

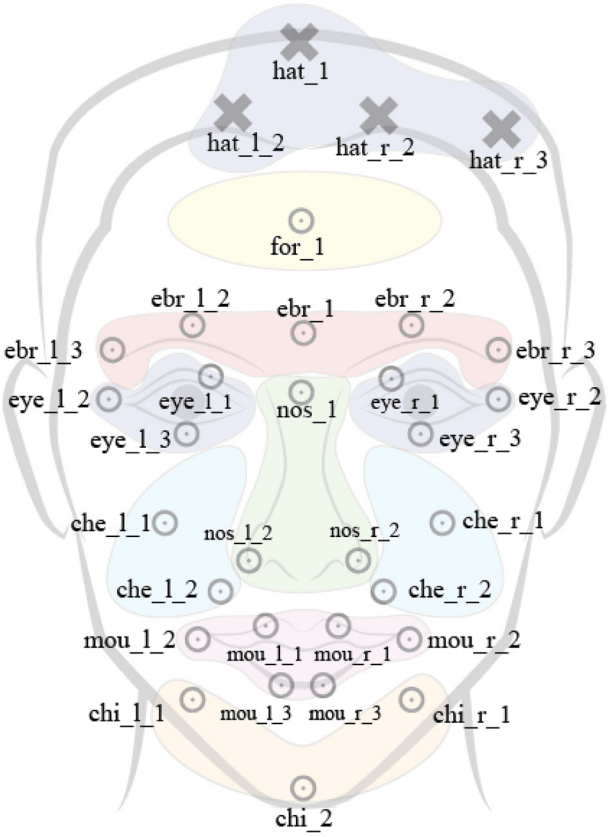

Fig. 9 The final placement of markers on the face in a 28 + 4 layout

deformations facial expressions accompanying speech production. For example, the upper eyelid marker position must be personalized to correctly match the geometry of eyebrow fold, to allow for a proper registration of wide eye opening and squinting for an actor.

4 People taking part in the recordings

People invited to the experiment show various anatomical features, skills and the ability to express themselves. They all are non-native, but fluent English speakers. In order to normalize the nomenclature in the further part of the work, the descriptions of individual subjects are used, that are presented in Table 3. Face anatomy was outlined based on its size and shape. Taking these factors into account is important because the labelling template is created on the basis of the first registered person and then transferred to other of a different anatomy. This allows an analysis aimed to determining how the change in the proportion of distances between individual markers affects the effectiveness of recognizing and labelling them in the software. Each person, apart from different anatomy, has features important for registration employing the motion capture system, such as prominent eyebrow folds, which

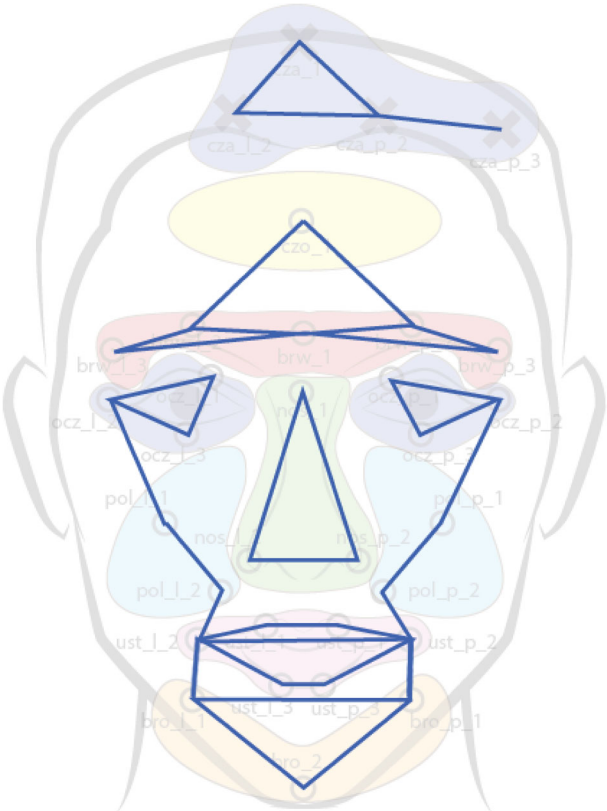

Fig. 10 Distribution of lines connecting pair of markers in 28 + 4 layout

are able to cover the marker responsible for the upper eyelid, large eyebrows that cause the markers to cover over them, or intense facial hair causing problems when applying markers to the skin.

Table 3 Persons taking part in the recordings

Person	Anatomy of the face	Characteristic features relevant to face registration	Presented skill in the 4th stage of the recording
1	Long shape, medium-sized	Big eyebrows	Reading text recited by a person 2
2	Oval shape, large	Eminent eyebrow arches	Recitation of the text read by the person 1
3	Square shape, medium size	None	Beatbox and singing
4	Long shape, large	Full facial hair, small eyes	Pantomime and singing
5	Oval shape, fine	Eminent eyebrow arches	Pantomime

5 Experiment course

The experiment of the recording of facial expressions during articulation of speech was made in accordance with the information contained in the previous chapters of the paper.

5.1 Preparations for recordings

Prior to the recording, each person was informed about the rules and about the course of the experiment, and also agreed in writing to participate in it. Markers were then applied. In the case of the person 4 who wears a beard, it was decided that the place where there was thick beard, and on which the chi_2 marker had to be affixed, a layer of adhesive tape is added. The remaining markers were located next to the facial hair. Next, a hat with additional markers was put on. In addition to the possibility of attaching reference markers to the Velcro, the cap made the hair of the persons recorded to not cover the markers. All respondents ready for recording are presented on the photograph in Fig. 11.

Before starting the recordings, the correctness of the markings applied in the Blade program was analyzed due to the possibility of live viewing of reconstructed data, as well as data coming directly from all seven cameras. It was also important to ensure that people do not have any jewelry or watches and that no shiny objects appear in the background, because they cause strong reflections of light, which are recorded by the cameras. When the recording person was ready, the level of the signal reaching the sound recorder was checked and the recording was switched on after the correction. In order to synchronize the voice of the actor with his face movements, the sync sound was used when starting the registration by the motion capture system.

5.2 The course of recordings

The recordings were registered in accordance with the assumptions presented in section 2.3. And the correctness of the recorded data was checked on an ongoing basis. Each person was assigned a separate session in the database, and each of the stages was recorded in separate shots. Such division allowed for easier post-production and data management. Registration of all stages in one recording would make work difficult because some tasks cause more errors. Separation of these moments allows for faster reconstruction of data and for a more careful analysis. Before the next stages were realized, the tasks to be performed were discussed again with the person being recorded, and during the third stage the persons who were to express emotions should be instructed on an ongoing basis.

Fig. 11 Persons taking part in the experiment: **a**) person 1, **b**) person 2, **c**) person 3, **d**) person 4, **e**) person 5

The registration time was controlled or not, depending on the stage. In the case of stage 1, the volume of speech was not interfered so that the speech was not forced and artificially prolonged, which is why the recording times of individual people differed significantly. In the case of stage 2, the time depended on the speed of reading the text. Stage 3 was supervised by issuing commands by the lecturer, and in stage 4 the total length of the recorded material differed depending on the activities performed. In the case of person 1, the recording was made using one shot. Person 2 recited the text read by the person 1 during the recording divided into three parts. It almost doubled the speech time. Person 3 presented the technique of beatbox in the first shot, and singing in the second. The skills of person 4 were also divided into two parts (first pantomime, and then singing). The person 5, presenting skills in the form of pantomime was registered in three recordings, approximately one minute each. Approximate recording times in mm:ss format of each stage are presented in Table 4. Stage 4 was presented by means of the sum of times for individual activities, which are listed above in succession.

Despite previous recommendations, during the recording, the head moved or changed the position of the whole body. Such situations occurred during stage 2 for persons 3 and 4 and during stage 4 for all persons except the first one. During this type of insubordination, no recorded data errors were observed, but if the face was rotated sideways or backward, the recording was interrupted. Data from such a recording would be burdened with too many errors and would make proper post-production impossible. Another problem was the application of markers near or directly of facial hair. Even slight facial hair caused situations in which the marker was removed during the recording process. It was therefore necessary to place them in a position at a distance of approx 1 cm from the place that was marked in the template. This was done in the case of markers on the chin of the person 1 and on the chin and at the corners of the mouth of the person 4.

5.3 Post-production of recordings

The recordings were subjected to the post-production process. For this part of the project to run smoothly, it was necessary to first create a labelling template. An error-free recording of stage 2, for the first person, and reconstruction were made, and then a template was created and stored.

In order to obtain the material adapted for analysis, the reconstructions of all recordings were made first. Then, the previously created labelling template was imported into each of them.

The next stage was to note the correctness of the recorded data. In Tables 5, 6, 7, 8 and 9 the number of blinks, all markers and their discontinuities in the individual stages of post-production recordings were presented. The number of blinks was determined based on the

Table 4 Time of recordings of five persons during individual stages

Person	Stage 1	Stage 2	Stage 3	Stage 4
1	00:20	00:40	00:53	00:33
2	00:43	00:38	01:12	00:38 + 00:14 + 00:09 = 01:01
3	00:21	00:42	01:02	00:56 + 01:08 = 02:04
4	00:53	00:54	01:06	00:45 + 00:33 = 01:18
5	00:31	00:56	01:01	01:04 + 01:07 + 01:00 = 03:11

Table 5 Number of registered markers and errors in the recordings of person 1

Stage	The number of all registered markers	The number of registered glints	Number of discontinuities, registered data
1	33	1	0
2	32	0	0
3	32	0	0
4	32	0	0

markers that appeared for a time corresponding to several frames of the recording and were not appropriately named in the process of automatic labelling. The number of discontinuities has been defined on the basis of the number of all markers excluding glints and the number of 32 markers that are defined in the 28 + 4 layout.

In Table 6 the fourth stage was divided into three parts due to the fact that the recording of the recitation was made in three independent recordings. Stage 4.1 corresponds to the first stanza, the second stanza is 4.2 and 4.3 represents the third stanza of the recited song. In Table 7 the fourth stage is divided into two parts. Stage 4.1 is the presentation of beatboxing and stage 4.2 of singing. In Table 8 stage 4 is also divided into two parts. Stage 4.1. is pantomime, and 4.2 is singing. In Table 9, the last stage was divided into three parts, during which the pantomime was presented. Due to the occurrence of multiple errors in stage 3 during the registration of person 4, their exact determination turned out to be impossible.

Then, the data was repaired in the recordings that required it. In the case of a person 1, only one glint occurred in the first stage and it was removed. The location of some markers in the recordings of other persons required the process of interpolation and manual editing due to moments of disappearance from the field of camera visibility. Table 10 shows which markers were burdened with these types of errors in individual stages.

Markers placed on the upper eyelids caused major problems in the post-production process. Difficulties in maintaining the continuity of the material also occurred in the area of the eyebrows or lips.

Due to the physical characteristic of faces, a large number of errors occurred in the recordings of stage 3 of person 3 (errors in eyelids markers for squinting eyes), and stages 3 and stage 3 and 4.1 of the person 4 (errors in eyelids markers, and glints), they required manual editing. To give the registered data an appropriate look, it was necessary to manually refine the labelling of the markers. A frequent mistake in these recordings was the substitution of neighboring marker names and the automatic assignment of some markers being

Table 6 Number of registered markers and errors in the recordings of the person 2

Stage	The number of all registered markers	The number of registered glints	Number of discontinuities, registered data
1	121	89	0
2	33	1	0
3	173	10	131
4.1	131	16	83
4.2	76	44	0
4.3	48	7	9

Table 7 Number of registered markers and errors in the recordings of the person 3

Stage	The number of all registered markers	The number of registered glints	Number of discontinuities, registered data
1	34	0	2
2	33	1	0
3	304	192	80
4.1	51	1	18
4.2	55	14	9

Table 8 Number of registered markers and errors in the recordings of the person 4

Stage	The number of all registered markers	The number of registered glints	Number of discontinuities, in registered data
1	80	1	47
2	41	0	9
3	713	No data found	No data found
4.1	173	15	126
4.2	134	0	102

Table 9 Number of registered markers and errors in the recordings of the person 5

Stage	The number of all registered markers	The number of registered glints	Number of discontinuities, in registered data
1	64	0	32
2	32	0	0
3	63	2	29
4.1	92	10	50
4.2	81	3	46
4.3	96	5	59

Table 10 Marker names that had to be subjected to the interpolation process

Person	Stage 1	Stage 2	Stage 3	Stage 4
1	none	none	none	none
2	none	none	nos_l_2, ebr_l_3, eye_l_1, eye_r_1	nos_l_2, ebr_l_3, eye_l_1, eye_r_1
3	ebr_l_3	none	eye_l_1, eye_r_1, eye_l_3, eye_r_3	nos_1, nos_l_2, mou_3, ebr_l_3
4	eye_l_1, ebr_r_3, ebr_l_3	ebr_r_3	eye_l_1, eye_r_1, eye_r_2, eye_r_2, eye_l_3, eye_r_3, nos_1	nos_1, nos_l_2, mou_l_3, ebr_l_3, ebr_r_3
5	eye_l_1, eye_r_1	none	eye_l_1, eye_r_1	eye_l_1, eye_r_1, mou_r_3

omitted by the algorithm, resulting in long breaks in the visibility of a given marker. Once all the data has been appropriately corrected, the faulty trajectories were removed, resulting in each recording consisting of 32 continuous markers.

6 Analysis of results

A thorough analysis of the recordings obtained allows for determining which factors positively or negatively affect the quality of recorded data. In addition, it allows to compare the results obtained in individual stages by different people and the stages themselves.

There are several factors that can influence the number of errors in the recordings. Due to the fact that the labelling template was created on the basis of the first person's recording, the data collected with its participation in all stages were practically free of errors regardless of the length of the recording. After analyzing the recordings, it was observed that the ebr_l_3 marker made vibrating movements throughout the recording time. This behavior was probably related to the slight covering of the eyebrows by the hair or contamination of the marker. A similar phenomenon was observed in the case of the nos_r_2 marker of the person 2, probably due to dirt. The analysis of the recorded data allowed to state that the shape or size of the face different from the person appearing during the creation of the labeling template slightly affect the quality of the reconstruction. There are far more serious consequences in the case of unfavorable characteristics of motion capture. A large fold of skin on the eyebrow or small eye size caused a large number of errors on the markers on the eyelids. The markers on the upper eyelid were concealed under the eyebrow or they were too close to the markers on the lower eyelid, which changed their names. This situation occurred most often during stage 3, so it is burdened with the most errors. They appeared at the time of strong and long-lasting squinting, eg when expressing anger or sadness. The beard in the case of a fourth person caused several errors around the mouth. Longer registration caused more errors of discontinuity during reconstruction and related complications. Markers, despite automatic labelling, lost their names, although they were present in places appropriate for the template.

Steps 1 to 3 have made it possible to compare the way information is provided by various persons. In stage 1 people spoke freely, which caused the movement of all markers responsible for the mouth, beard, as well as che_l_2 and che_r_2. The movement of both eyelids, small movements of the entire head as well as the eyebrows were recorded. In stage 2, in the case of persons 1 and 3 there was a significant reduction in the movement of the entire face. The change in the position of markers occurred mainly within the mouth and chin. Person 2, however, showed only the movement of the lips themselves. In contrast, people dealing with pantomime (4 and 5) emphasized their facial expressions more than in the case of free speech. To properly intonate individual fragments of the text, they moved the whole head and also activated the muscles around the eyes. During stage 4, each person similarly expressed a feeling of joy. They were expressed by raising the markers on the lips, on the chin and cheeks, and also by winking eyes. The feeling of sadness reduced the corners of the mouth and raised the markers on the chin. Person 3 caused the chi_2 marker to coincide almost with markers on the lower lip of the mouth. Due to facial hair, the person 4 was not as able to present this expression as well as the others. In addition, the ebr_l_1 and ebr_r_1 markers were lowered, and the eyes were almost closed, which caused numerous errors due to the too close-up of the markers that lasted a few seconds. During blinking, the markers were approaching each other for a very short period of time, which meant that they did not cause such problems. The expression of feeling surprised, participants presented through

their mouths and eyelids. An exception is person 2 who lifted one side of the upper lip. The feeling that entailed the opposite behavior of facial expressions was anger. The participants caused that the markers placed on their faces approached each other within individual areas. They clenched their lips, frowned and squinted their eyes. It was an expression that caused a lot of errors. In Fig. 12, five exemplary emotional expressions of the person 3 were presented.

It is worth noting that every person, despite activating similar facial muscles, did so to varying degrees and with different dynamics. People 1, 2 and 5 fluently and calmly changed their face, and people 3 and 4 did it dynamically.

In step 4, the first two persons were compared with each other. They transmitted the same information. Due to the fact that the person 2 recited the memorized text, she spoke more slowly and was more focused on the correctness of the transmitted content. This resulted in her expression being less dynamic and enhanced than in the case of the person who communicated the content freely and smoothly. The movement of markers across the face was more pronounced than in the case of the person 2. The beatboxing skills presented by the person 3 allowed to observe the very strong involvement of all muscles around the mouth. In comparison to free speech or reading text, performing beatbox caused lip movement in all directions and with high dynamics. To maintain the rhythm, the person moved his head uniformly. During the singing the subject used a speech apparatus with less dynamics and did not move his head as much as in the case of beatbox. He kept his mouth wide open most of the time and he narrowed his eyes, which may mean more effort during this type of activity. The skill presented in comparison to normal speech is characterized by a more extensive use of the speech apparatus. Similar behavior was exhibited by a person 4. However, she performed a less expressive and demanding piece, which meant that the changes in the position of the markers were smaller. The pantomime performed by this person and the person 5 in particular showed that during the activity of one part of the facial muscles there is activation of other muscles around it. When combining the movement of several areas, all markers change their position significantly. These persons also presented the work of muscles in different directions, so that the individual fragments of the face work appeared in an aligned, opposite or delayed way. This reflects the great flexibility and possibilities of human facial expressions. In comparison with other persons presenting their skills in stage 3, persons dealing with pantomime have greater opportunities to modify the appearance of both the whole face and the selective separation of individual areas on it.

The largest number of errors was caused by prolonged, excessive approach and disappearance of markers placed on the eyelids. Such a situation took place most often during the third stage. Therefore, it was not possible to determine the exact number of glints and

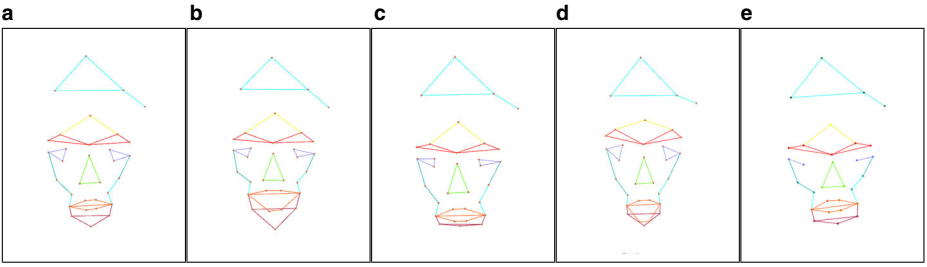

Fig. 12 Five exemplary emotional expressions of a person 3: **a**) neutral pose, **b**) feeling of joy, **c**) feeling of sadness, **d**) feeling of astonishment, **e**) feeling of anger

discontinuities for the person 4. The use of interpolation algorithm for the position of markers in the moments of their lack of visibility allowed to obtain a continuous material that effectively served the practical application in the form of computer animation. The various stages allowed for the observation of the used face areas in various activities and for the recognition of differences in the way they are activated by different people.

7 Conclusions

The multimodal speech database was extended by FMC data and made accessible to the research community. The earlier absence of this modality in speech corpora was probably caused by a relatively large workload that needs to be invested in application of face motion capture technology for this purpose. After analyzing the entire registration process, post-production and practical use of recordings, it is possible to formulate conclusions regarding further improvements leading to a wider use of motion capture systems for recording multimodal speech or facial expression databases. Especially FMC might serve as a source of ground-truth data for facial expression databases, but the process of data acquisition can be further improved. Namely, the number of cameras is an element that can have a positive impact on the quality of recorded data, because increasing this number may enhance the visibility of markers placed in different planes by more devices, positively affecting the accuracy of the mapping of their position. Additional cameras might be placed around the person being recorded. If the system allowed registration of markers in each position, the actor could move and rotate the entire head freely. The problem of mapping the movement of the eyelids can be eliminated by changing the position of the markers eye_l_1 and eye_r_1 and placing them under the eyebrows. This will help to avoid the problem of markers approaching each other and covering them by the eyebrow fold. In this way one will be able to reproduce squinting or wide eyes opening, but not blinking. It is blinking, however, the repetitive movement performed at high speed, which does not require an exact reproduction in three-dimensional face models. Facial hair cause markers to fall off the face surface. To eliminate this problem, it is possible to earlier apply a tape placed in the position where the marker is to be glued. In addition, alternative solutions in the form of a special ink that reflects infrared radiation might be used instead of sticking markers. To minimize the likelihood of the cumulation of errors, the recording sessions should be divided into the smallest possible steps. It also allows quick reconstruction, analysis of the correctness of recorded data and their effective organization.

The recordings were made at 120 frames per second. It is possible to increase the frame rate, but after analyzing the most dynamic fragments of recorded data (the pantomime presented by the person 4 and the beatbox of the person 3), it was found that the movement was correctly and fluently rendered. Increasing the frame rate is associated with a significant increase in the amount of data saved. This can adversely affect the performance of the entire system and hinder work during post-production, hence the frame rate used should be as low as possible while maintaining a smooth representation of the most dynamic movements.

The collected information about the way words are spoken can be used to create algorithms that recognize elements of speech based on the movement of the mouth. To make this work more accurate, an increased number of markers in the mouth area should be employed. The results of the conducted experiments may also be used for the creation of software that recognizes emotions based on the movement of markers on the entire face or for transcribing speech to IPA (International Phonetic Alphabet) or for synthesising speech with prosodic features reflecting speakers' emotions.

Acknowledgements Research sponsored by the Polish National Science Center, Dec. No. 2015/17/B/ST6/01874.

References

Aloff, M. (2013). Disney's snow white at 75. *Virginia Quarterly Review*, *89*(1), 238–244.

ABCP channel (2018). Benedict Cumberbatch—Behind-the-Scenes of The Hobbit: Desolation of Smaug, https://www.youtube.com/watch?v=Wu9XPEdBelY, (accessed on Nov. 2018).

Barsegyan, A. (2017). Perception neuron inertial motion capture vs optical mocap systems and the first production motion capture session experience, http://cgicoffee.com/blog/ 2017/04/first-production-motion-capture-session-report (accessed on Nov. 2018).

Calvert, T.W., Chapman, J., Patla, A. (1982). Aspects of the kinematic simulation of human movement. *IEEE Computer Graphics and Applications*, *2*(9), 41–50.

GameCrate channel (2018). Kevin Spacey talks acting in video games, https://www.youtube.com/watch?v=dnvpO-ohCRY (accessed on Nov. 2018).

Czyżewski, A., Kostek, B., Bratoszewski, P., Kotus, J., Szykulski, M. (2017). An audio-visual corpus for multimodal automatic speech recognition. *Journal of Intelligent Information System*, *49*, 167–192. https://doi.org/10.1007/s10844-016-0438-z.

Hasegawa, M., Mori, T., Shirakura, K., Watanabe, K., Yuminaka, Y. (2016). Non-contact vital sensing systems using a motion capture device: medical and healthcare applications. *Key Engineering Materials*, *698*, 171–176.

Hassanat, B. (2014). Visual passwords using automatic lip reading, international journal of sciences: basic and applied research. *IJSBAR*, *13*(1), 218–231.

Hellblade (2018). http://www.hellblade.com (accessed on Nov. 2018).

Jiang, J., Li, H., Mu, K., Tao, J., Yang, M. (2016). *Emotional head motion predicting from prosodic and linguistic features*. New York: Springer.

Kahou, S.E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, V., et al. (2016). Emonets: multimodal deep learning approaches for emotion recognition in video. *J Multimodal User Interfaces*, *10.2*, 99–111.

Le, B.H., Zhu, M., Deng, Z. (2013). Marker optimization for facial motion acquisition and deformation. *IEEE Transactions on Visualization and Computer Graphics*, *19*(11), 1859–1871.

Maxwell, D., & Ginsberg, C.M. (1984). *Graphical marionette, ACM SIGGRAPH Computer Graphics* Vol. 18. New York: ACM Press.

Media Magik Entertainment channel (2018). Avatar exclusive – behind the scenes (The Art of Performance Capture), https://www.youtube.com/watch?v=P2_vB7zx_SQ (accessed on Nov. 2018).

Menache, A. (2011). *Understanding motion capture for computer animation*, 2nd. Burlington: Morgan Kaufmann.

Minghao, Y., Jinlin, J., Jianhua, T., Kaihui, M., Hao, L. (2016). Emotional head motion predicting from prosodic and linguistic features. *Multimedia Tools and Applications*, *75*, 5125–5146. https://doi.org/10.1007/s11042-016-3405-3.

Motion Analysis (2018). Motion capture system: https://www.motionanalysis.com (accessed on Nov. 2018).

Munhall, K.G., Jones, J.A., Callan, D.E., Kuratate, T., Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: head movement improves auditory speech perception. *Psychological Science*, *15*(2), 133–7.

Nature video channel (2018). Creating Gollum, https://www.youtube.com/watch?v=w_Z7YUyCEGE (accessed on Nov. 2018).

Nogueira, P. (2011). Motion capture fundamentals; a critical and comparative analysis on real-world applications. Instituto de Telecomunicaç,ões, 18 de Novembro de 2011.

Noire, L.A. (2018). http://lanoire.wikia.com/wiki/MotionScan (accessed on Nov. 2018).

OptiTrack (2018). Motion capture system: http://www.optitrack.com (accessed on Nov. 2018).
Parent, R. (2009). *Computer animation complete*. Burlington: Morgan Kaufmann.
PlayStation channel (2018). BEYOND: Two Souls Making of – Capturing Performance, https://www.youtube.com/watch?v=5DwHjNenAmw (accessed on Nov. 2018).
Reverdy, C., Gibet, S., Larboulette, C. (2015). Optimal marker set for motion capture of dynamical facial expressions. In *MIG'15 Proceedings of the 8th ACM SIGGRAPH conference on motion in games, pp. 31–36, Paris, France, November 16–18, 2015*.
Schulz, A. (2010). Motion capture technical report, Instituto National De Matematica, Pura E Aplicada, Rio de Janeiro, May 6.
Sulovska, K., Fiserová, E., Chvosteková, M., Adámek, M. (2017). Appropriateness of gait analysis for biometrics: initial study using FDA method. *Measurement*, *105*, 1–10. https://doi.org/10.1016/j.measurement.2017.03.042.
Tracklab (2018). Motion capture system: https://www.tracklab.com.au (accessed on Nov. 2018).
Vicon (2018). Motion capture system: https://www.vicon.com/ (accessed on Nov. 2018).
Vryzas, N., Liatsou, A., Kotsakis, R., Dimoulas, C., Kalliris, G. (2018). Speech emotion recognition for performance interaction. *Journal of the Audio Engineering Society*, *66.6*, 457–467. https://doi.org/10.17743/jaes.2018.0036.
Vryzas, N., Vrysis, L., Kotsakis, R., Dimoulas, C. (2018). Speech emotion recognition adapted to multimodal semantic repositories. In *13th international workshop on semantic and social media adaptation and personalization (SMAP), Zaragoza, 2018, pp. 31–35*. https://doi.org/10.1109/SMAP.2018.8501881.
Williams, R. (2001). The Animator's Survival Kit, Faber & Faber.
WIRED channel (2018). Dawn of the Planet of the Apes: transforming human motion-capture performances into realistic Apes, https://www.youtube.com/watch?v=4NU9ikjqjC0 (accessed on Nov. 2018).
Zarins, U., & Kondrats, S. (2014). Anatomy for sculptors: understanding the human figure, EXONICUS.
Zhang, S., Zhang, S., Huang, T., Gao, W. (2018). Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia*, *20*(6), 1576–1590. https://doi.org/10.1109/TMM.2017.2766843.