3D Facial Video Retrieval and Management for Decision Support in Speech and Language Therapy

Conference Paper · June 2017

DOI: 10.1145/3078971.3078984

5 authors, including:

Isabel Cristina Ramos Peixoto Guimarães
Escola Superior de Saude do Alcoitão
115 PUBLICATIONS **762** CITATIONS
SEE PROFILE

Margarida Grilo
Escola Superior de Saude do Alcoitão
13 PUBLICATIONS **22** CITATIONS
SEE PROFILE

Sofia Cavaco
Universidade NOVA de Lisboa
55 PUBLICATIONS **206** CITATIONS
SEE PROFILE

João Magalhães
Universidade NOVA de Lisboa
125 PUBLICATIONS **716** CITATIONS
SEE PROFILE

3D Facial Video Retrieval and Management for Decision Support in Speech and Language Therapy

Ricardo Carrapiço[1], Isabel Guimarães[2], Margarida Grilo[2], Sofia Cavaco[1] and João Magalhães[1]
[1]NOVA LINCS, Computer Science Dept., Universidade NOVA Lisboa
[2]Escola Superior de Saúde do Alcoitão
Lisbon, Portugal
r.carrapico@campus.fct.unl.pt,iguimaraes@essa.pt,margarida.grilo@essa.pt
scavaco@fct.unl.pt,jm.magalhaes@fct.unl.pt

ABSTRACT

3D video is introducing great changes in many health related areas. The realism of such information provides health professionals with strong evidence analysis tools to facilitate clinical decision processes. Speech and language therapy aims to help subjects in correcting several disorders. The assessment of the patient by the speech and language therapist (SLT), requires several visual and audio analysis procedures that can interfere with the patient's production of speech. In this context, the main contribution of this paper is a 3D video system to improve health information management processes in speech and language therapy. The 3D video retrieval and management system supports multimodal health records and provides the SLTs with tools to support their work in many ways: (i) it allows SLTs to easily maintain a database of patients' orofacial and speech exercises; (ii) supports three-dimensional orofacial measurement and analysis in a non-intrusive way; and (iii) search patient speech-exercises by similar facial characteristics, using facial image analysis techniques. The second contribution is a dataset with 3D videos of patients performing orofacial speech exercises. The whole system was evaluated successfully in a user study involving 22 SLTs. The user study illustrated the importance of the retrieval by similar orofacial speech exercise.

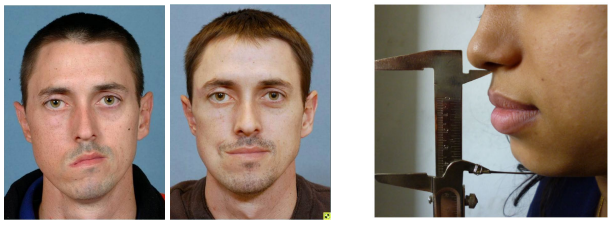
KEYWORDS

Speech and Language Therapy, Facial Anthropometry, Decision Tool, 3D Video Retrieval

ACM Reference format:
Ricardo Carrapiço[1], Isabel Guimarães[2], Margarida Grilo[2], Sofia Cavaco[1] and João Magalhães[1]. 2017. 3D Facial Video Retrieval and Management for Decision Support in Speech and Language Therapy. In *Proceedings of ICMR '17, Bucharest, Romania, June 06-09, 2017,* 8 pages.
DOI: http://dx.doi.org/10.1145/3078971.3078984

1 INTRODUCTION

Speech is a form of human communication that starts to develop during infancy. While young children typically produce many speech errors, as the child becomes older, many speech production errors are corrected without the need of a specialist's intervention. However, in many cases speech and language therapy is required for the treatment of communication related problems.



(a) (b)

Figure 1: (a) Facial paralysis is a disorder that requires specific exercises. (b) Close facial measurements with calipers (figure from [1]).

To assess the patients conditions, SLTs use several speech and language exercises. The assessment of the patient is mainly done by observation: the SLT decides if the patient is correctly producing speech sounds or not mostly based on normative data [10]. In addition to the exercises, SLTs also take measurements of the patient's face, in particular when assessing oral-motor disorders (figure 1a). The tools used by the SLT for assessment or during speech and language therapy can be quite intrusive which creates discomfort to the patient. As an example, figure 1b shows a caliper being used to take facial measurements. This is especially critical when applied to children, who may be impatient and may not understand the benefits of its use. For this reason, speech and language therapy could benefit from an improved acquisition and visualization of the therapy exercises information [11].

The work of an SLT is diverse and multidimensional, nevertheless interactive technologies and 3D video systems can be useful and helpful on several SLT's tasks. Serious games are an example of computational tools that have already introduced many successful changes [13, 14, 16]. Other tools are focused on health data to help both the SLT and the patient by providing the needed information for the SLT and by being almost transparent to the patient [11, 18–20]. The SLT relies in both speech and visual evidence to make the best clinical decision with respect to the best possible therapy [17]. By providing new sources of information or further improving the current ones, computer vision can help improving the quality and capability of speech and language disorders assessment.

This paper proposes a system that provides SLTs with state-of-the-art 3D video managament and retrieval methods to improve

ICMR '17, Bucharest, Romania

DOI: http://dx.doi.org/10.1145/3078971.3078984

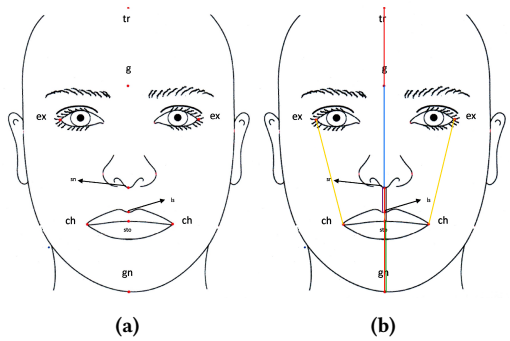


Figure 2: Facial anthropometry in speech and language therapy. (Figure from [3].) (a) Eight facial anthropometric points. (b) Seven anthropometric measurements.

health information management processes. Existing systems for speech and language therapy [11, 18] do not offer the same functionality as we propose in this paper. There are already speech repositories [11] but they do not integrate a 3D viewer. And the only tool that supports 3D [18] is focused on telemedicine and is not a video repository. Our 3D video retrieval and management system supports multimodal health records and provides the SLTs with tools to support their work in many ways: (i) it allows SLTs to easily maintain a database of their patients speech exercises; (ii) it supports three-dimensional orofacial measurement and analysis in a non-intrusive way; and (iii) it offers the possibility of searching patient speech-exercises by similar facial characteristics, using facial image analysis techniques.

The system is described in section 3. Section 4 describes a dataset of 3D videos with patients performing orofacial speech exercises. Section 5 reports the user study evaluation results.

2 RELATED WORK

Most of electronic health records (EHR) management systems are too generic and do not offer search or annotation functionalities for rich multimedia content. Current work in this direction only offers visualization of 3D synthetic models [20] or medical imaging [2]. Moreover, information retrieval methods with a ranking by relevance algorithm, are only now making way to such systems [19]. In our system, we propose two key advances over the state-of-the-art: the facial anthropometry of real patient's 3D digitized model, and the search for similar orofacial exercises.

Facial anthropometry consists of the measurements of the human face either in rest or in activity. In the particular case of speech and language therapy, facial anthropometry is related to seven specific anthropometric measurements between eight points of the face that are considered relevant (figure 2) [3]. Knowing these measurements, and their distance variations through time, allows the SLT to better assess the patients and measure their progress. There has been work in computational 3D facial anthropometry, including inference methods on 2D images [8], 3D face modeling with laser scans [15], inference on 3D images [21] and recent advances in the detection of facial landmarks [23] which already offer very good accuracy. Our system uses a semi-automated method, so as not to break the

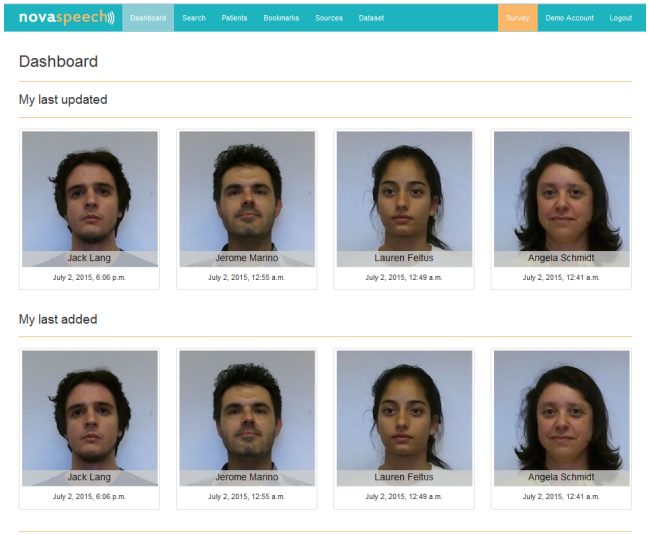


Figure 3: NOVASpeech is a system for storing, searching and visualizing video recordings of orofacial speech exercises.

SLT trust on the system – we "suggest" some key landmarks and provide the SLT with the tools to mark further landmarks or adjust existing ones.

Recent work in 3D shape retrieval [9, 22] relies on too strict measures to be used in speech and language therapy. The advances observed in the ImageCLEF Visceral track [12] concerning 3D image retrieval in the health domain already offer important metrics that can relax the similarity computation required by our problem. Closer to our setting is the work of Chen et al. [6] that aims at analysing 3D facial data to infer facial expressions. However, due to the specific facial characteristics we are looking for in our setting, we opted for not using 3D face features and instead explore multiple evidence from metadata fields, multiple visual descriptors and a wise diversity strategy that is focused on the SLT information needs.

3 THE NOVASPEECH FRAMEWORK

NOVASpeech is a web tool[1] that allows SLTs to easily maintain a database of their patients recorded therapy exercises (figure 3[2]). The tool is focused on having useful and easy to use features that will improve the SLT experience when evaluating patients. Apart from being used by individual SLTs, the tool can be used, for example, at health institutions, where the SLT works, as it allows multiple private user accounts. With each SLT having his/her own account, the patients information is private to each SLT (unless the SLT explicity gives access to his/her patients' data to his/her colleagues).

The exercises are captured while the patient performs the therapy exercises, without any interruptions, using only a Kinect camera and a computer (section 3.1). By being on the web, the SLT can access the patients information from anywhere, including home and without having to go through a software installation process.

[1] http://3dspeech.novasearch.org/
[2] All names and information shown in this paper are illustrative and do not correspond to real patients.

It suffices to have a web browser. Being a support tool for speech and language therapy, one important part of this experience is the **browsing and analysis** of the therapy exercises, which helps the SLT on assessing the patient's speech sounds or oral-motor disorder and on evaluating the patient's progression. For this reason, NOVASpeech offers different functionalities to improve this experience (section 3.2). To help the SLT on assessing different aspects of the disorder, the tool supports different types of data, which include audio, video, visible images and infrared images, and 3D models of the patients.

One of the major innovations of our work is the **visualization and annotation of the patient's exercises in 3D**, allowing a type of control and examination that was not possible before: the SLT can now rotate and move the patient's 3D model to observe from different perspectives (figure 4). This feature can be used for several purposes by the SLT and it is of particular interest for facial anthropometry, as it allows to take relevant anthropometric measurements without the need to use a caliper directly in the patient's face (more details in section 3.3). The second main functionality is the possibility of **searching** through the SLT's patients. NOVASpeech uses computer vision techniques to improve the SLT experience when using this fucntionality (section 3.4).

3.1 Patient data management

Generating and adding new information to NOVASpeech is one fundamental task that needs to be addressed. *How to record the data* and *how to add the recordings to the tool* are important questions that the SLT, not being a specialist in these matters, should not be worried about. To address these questions, we developed a tool called NOVASpeech Kinect Recorder that captures and saves the information generated by the Kinect Sensor. At the recording stage the files do not contain any metadata, like the patients name or type of exercise, but these can be inserted when the data is added to NOVASpeech.

NOVASpeech organizes the data according to the requirements from the SLTs. Each SLT has an account that holds all his/her patients. The patient entities have exercises, which are the recordings made by the SLT during the therapy session. Each recording consists of a sequence of frames (visible and infrared). These entities are all related in a way that leads to the SLT who inserted the data in the system.

SLTs have full access over their patients data. They can add, edit and delete the patients and exercises. This way SLTs have a tight control over what they have in NOVASpeech and can maintain a more dynamic experience by, for example, editing the notes of a patient to reflect his/her performance or motivation upon adding a new exercise to the therapy session.

The patient entity holds personal information (name, gender, age and photo). For privacy concerns, the patient's health information is stored as a separate entity and holds data related to oral habits, like onychophagia or chewing objects, nasal obstruction, orthodontic treatment, facial trauma and SLTs' free-text notes.

A speech exercise is a central data entity in our system. This entity, stores the type of exercise (rest or dynamic), the video frames and the audio file. A dynamic speech exercise can be an oral diadochokinesis training, (repetition of sounds like 'pa' or 'ma' to assess articulatory rate in speech production) or a labial training with stretching or protrusion of the lips. Like with the patient entity, each exercise also has a text-free notes field. Table 1 shows a summary of the fields for each data entity.

Patient details	Patient condition	S&L Exercise
Name	Oral habits	Rest
Gender	Nasal obstruction	Oral diadochokinesis
Age	Orthodontic treatment	Labial training
Photo	Facial trauma	Notes
	Others	

Table 1: Summary of data fields

3.2 Data navigation and examination

One main feature of NOVASpeech is the navigation and analysis of the therapy exercises. With each exercise being associated with a patient, the SLT first has to select the patient before having access to the exercise list (*Patients* option in the top menu of figure 3).

Upon selecting a patient, we are redirected to his/her page, which shows all the information that the system has about the patient, plus his/her exercise list. To easily navigate through this list, we added two features: a timeline representation of the exercises, and the possibility of applying filters to the list and timeline (figure 5). The timeline allows the SLT to inspect the sequence of exercises that the patient performed.

By selecting an exercise, the user is redirected to the exercise page (figure 6). Similar to the patient page, the exercise page has the patient information, so the SLT can maintain context of the patient's condition, and also the exercise data. The SLT is then offered with a visualization of the exercise audio signal and the corresponding video frames. With NOVASpeech the user can see the current frame plus the two preceding and following frames (figure 6.a). This way it is easier for the SLT to examine the differences on the patient face as the frames are in a timeline. It is also possible to navigate through this sequence manually, controlling it with the scrollbar. Every frame can be seen in more detail by clicking on it, opening the frame page.

The exercise is also composed of an audio track. To control the track we added an audio player with waveform support, so the SLT can visually identify the parts where the patient has emitted a sound (figure 6.c). The player allows selecting a region and playing only that portion of the audio. The audio player is connected to the sequence of frames in a way that playing the audio controls the progress of the sequence. There is also a slider that the SLT can use to control the sequence play speed, by making both the audio play and image transition slower or faster.

Finally, inspecting the frame allows the SLT to see the image in more detail. The frame page also has emphasis on frame navigation, by having a frame selector that allows a frame-by-frame displaying (figure 6.b). This way the SLT can see the differences between frames more easily. The SLT can also see the infrared image that the Kinect sensor captured, mapped onto a visible color space and choose to interact with the 3D model, as will be explained in the following section.

Figure 4: NOVASpeech 3D visualizer with a model being rotated.

Figure 5: Timeline representation of the exercises and the corresponding exercise filters.

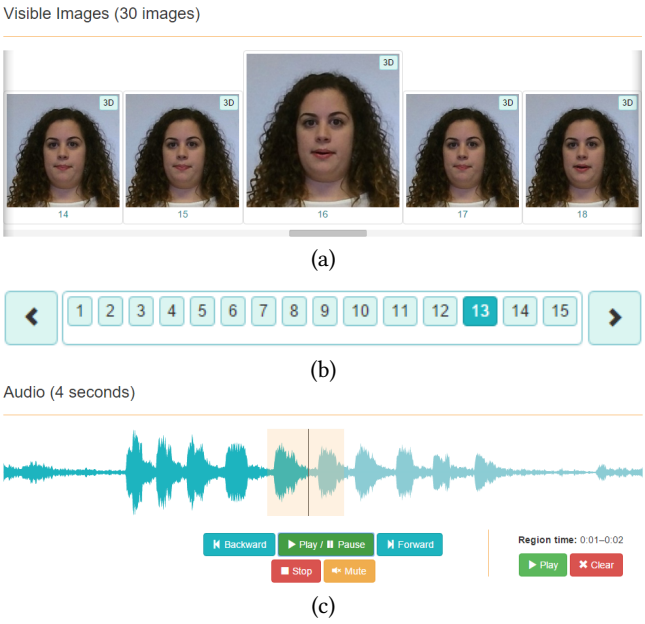


Figure 6: Exercise visualization: (a) video frame sequence (b) frame selector, and (c) audio player with a region selected.

3.3 3D visualization and facial anthropometry

The NOVASpeech 3D visualizer uses the 3D point cloud of the video frame to form the 3D model of the patient. Exhibiting the patient's 3D model is a unique feature of NOVASpeech, as it allows non-intrusive facial anthropometry. The model can be moved and rotated, as seen in figure 4, similarly to an object on a scene with the origin being the Kinect camera position. It should be noted that the model shows a partial view of the patient face. This is because the camera is in front of the patient, thus not capturing side views of the patient's face. However, it is sufficient for speech and language therapy because the face is complete.

Seeing the patient from different perspectives is important and is something that before, was only possible by photographing the patient multiple times, requiring the patient to be still. This means that it cannot be done during a speech sound exercise, as asking the patient to stop or make a particular position can be unnatural and difficult, specially for patients with facial paralysis, where they cannot accurately control part of their face. Being able to examine a 3D model with only one capture and during an exercise opens new opportunities. One of such opportunities is, for example, rotating the model to inspect the gap between a patient lips, while he/she is producing speech sounds, without any sort of intrusion of the patients space.

Another important aspect of the 3D visualizer is the facial anthropometry. Double-clicking on the model marks a point and a total of two points can be marked at the same time, as exemplified in figure 7. Marking a point in three dimensions can be confusing. The approach we took is that the point has to exist in the point cloud (the model) for it to be marked. After selecting two points we obtain the three dimensional euclidean distance between them. This facial measurement is then associated to the patient exercise and kept as a record for later search or filtering. The annotation of points completely eliminates the need of a vernier caliper, as it allows measuring the patients face at any moment during the speech sound exercises, without any sort of contact with the patient.

3.4 Search for similar exercises

NOVASpeech's search mechanism allows SLTs to find exercises (restricted to the SLT's patients) that have similar facial anthropometrics. In addition, it allows choosing the facial anthropometry search type, by searching only the face's lower or middle third for visible images and the whole face for both visible and infrared images (figure 8). This functionality is particularly useful to the SLTs who can now more easily support their clinical decisions on evidence from past patient data.

To be searchable, an image needs to have annotations of where the eyes are. NOVASpeech provides an initial automatic detection of the eyes and the SLT can then adjust the position of the eyes on the query image. NOVASpeech's search mechanism compares the input image characteristics with the remaining images in the database, and then forms a rank based on the results. Due to the

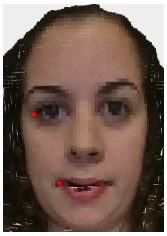


Figure 7: NOVASpeech 3D visualizer with marked points for facial anthropometry

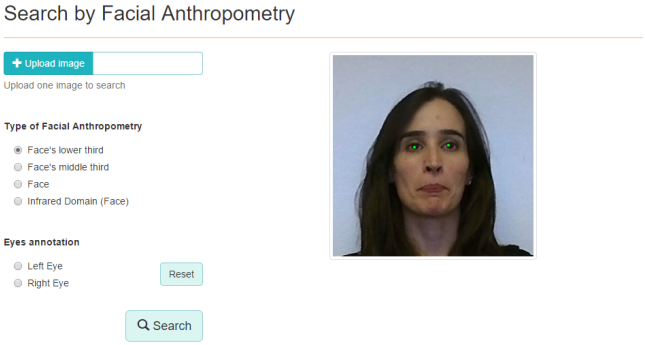


Figure 8: Facial anthropometry search

nature of the data organization in our setting, it makes more sense to rank exercises instead of single images. For this reason we opted to merge the images in their respective exercises. The rank fusion process is applied before the merging into exercises, and the result diversity is applied afterwards.

Ranking of face images. To robustly capture the different facial traits of the patient, we rely on several different image descriptors: Gabor filters, Local Binary Patterns, Color and Edge Directivity Descriptor [4] and Fuzzy Color and Texture Histogram [5].

For each type of image descriptor, we take the face image descriptor vector and compute the distance between the query and the indexed images. This creates one rank for each type of feature descriptor for all frames of all exercises that we have in our index.

Rank fusion. When there are several ranks that offer different views of the same data, we need to combine them into one single rank to present the search results to the end user. To accomplish this rank fusion we use reciprocal rank fusion (RRF) [7]. RRF is based on the fact that while highly-ranked images are important, lower-ranked images should not be penalized hardly. RRF assumes there is a set of images D to be ranked and a set of ranks R, each one corresponding to an visual feature. The score of each image is calculated by the following expression:

$$RRFScore(d \in D) = \sum_{r \in R} \frac{1}{k + pos_r(d)} ,$$

where $pos_r(d)$ is the position of image d in rank r, and k can be adjusted to diminish the difference between highly-ranked and lower-ranked images. As suggested by the Cormack et al. [7], we used $k = 60$.

Results diversity. Now that we have a single rank of all the images in the system, we need to compute the actual rank of the orofacial exercises. To generate the rank, it is important to obtain some diversity in the results so we do not end up with several results from the same person.

To increase the results diversity, we applied a logarithmic penalty to the score of each result in the rank. This approach was inspired by the inverse document frequency used in information retrieval and is as follows:

$$Score(r \in Rank) = log\left(\frac{1}{\#occurrences(r)}\right) + 1 .$$

The expression was applied progressively through the rank instead of statically, i.e. instead of applying the same value to all the instances of that person. More specifically, the $\#occurrences(r)$ increases by one (starting at one) each time we see the occurrence of a person .

In summary, **rank fusion** enable us to combine the different features in a single rank and using **result diversity** made the rank less monotonous by allowing other patients to rank up (instead of returning only exercises from the patient in the query).

4 3D VIDEO OROFACIAL SPEECH DATA

Given the pioneering nature of this research, we created a new 3D depth video dataset for Speech and Language Therapy[3]. We recorded a total of 31 volunteers, 25 females and 6 males with an average age of $\approx$ 25 years. Most volunteers were graduate students. The volunteers were asked to do five different speech sound exercises while seated in front of a camera. These exercises were chosen as being good indicators of patient progress.

The whole process was designed and curated by two SLTs - the sequence of exercises and the recording instructions were provided by them. They were also present during the capture. The exercises are: (1) stretching the lips, (2) protrusion of the lips and repeatedly saying the sounds (3) "pa", (4)"ba" and (5) "ma". The patients were also recorded while in rest, doing a neutral face. Each one of the volunteers was recorded six times for a total of 186 recordings. Table 2 illustrates each exercise with an image of a 3D model from the dataset.

A Kinect 2.0 camera was used to capture 3D video, visible and infrared images, audio and field depth. The recordings were done with an application implemented in the scope of this work.

5 EVALUATION

In order to validate NOVASpeech, we performed a user study with volunteer student and professional SLTs. There were 22 volunteers with ages between 21 and 39 years who participated in the study. Their average age was $\approx$ 27 years. The study started with a short presentation of NOVASpeech, after which the volunteers tested and explored the tool. Afterwards they answered a survey regarding the experience.

The survey included some questions on personal data and work experience (such as the age, years of experience, and usage of computer tools for speech and language therapy) and questions about

[3]Available at: http://3dspeech.novasearch.org/

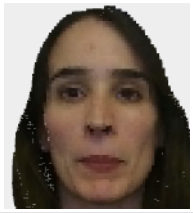
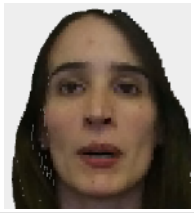
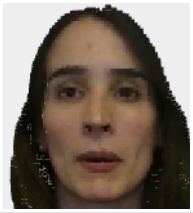
Exercise type					
In Rest	Stretching	Protusion	"pa"	"ba"	"ma"
					

Table 2: Example of the speech sound exercises in the dataset, illustrated with 3D images.

Figure 9: Example of the 3D models present in the NOVASpeech dataset.

NOVASpeech. The latter included questions about: (i) the visualization of speech and orofacial exercises, (ii) 3D facial anthropometry, (iii) search and (iv) general usability.

The survey showed that all volunteers were computer literate and had all the needed background to evaluate the tool. Their work experience was divided between no experience for students (36.4%), 1-5 years (27.2%) and 6-10 years of experience (36.4%). Also, all volunteers used computer tools to assist them during speech and language therapy sessions or to evaluate the speech sound productions of patients. They used computer games (85.7%), audio recording tools to record the speech exercises (71.4%) and tools to keep the patients data (66.7%).

5.1 Orofacial and speech exercise visualization

The user study included a set of questions to assess NOVASpeech exercise visualization features. These include the visualization of the exercise image sequence and inspection of frames (figure 6.a and b), the control of the playback speed, and the waveform visualization and audio player (figure 6.c).

NOVASpeech navigation and visualization of speech exercises had great acceptance. All volunteers considered useful to observe and analyze the orofacial dynamics across multiple simultaneous frames, as shown in figure 6.a. However, one volunteer considered that the images were too small for the type of observation he/she needs to perform.

An important feature to observe the orofacial exercises in detail is the control of the playback speed. Figure 10 shows that all volunteers considered slow motion useful for at least one purpose, but mostly for inspecting orofacial movements. Most participants considered that fast forward is also useful for analyzing the orofacial fatigue while producing speech sounds over a long period of time.

The audio player is an important component of NOVASpeech. Both its features, namely the waveform visualization and play of an audio region, also had great acceptance. All volunteers considered them beneficial. Moreover, one SLT mentioned that it could be useful to show the patient a normal rate of talk in the case of him/her speaking slower or faster than normal.

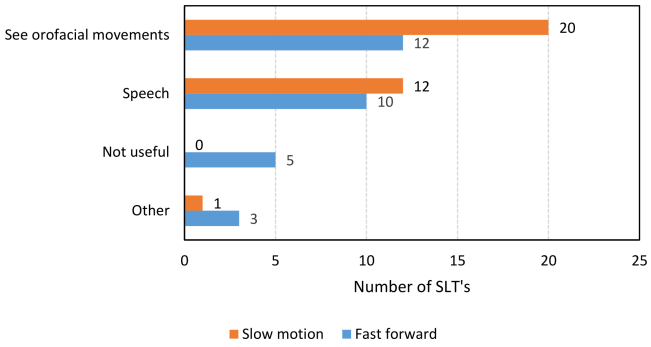


Figure 10: Orofacial speech exercise inspection with slow motion/fast forward

5.2 3D facial anthropometry assessment

As mentioned above, one of the main innovations of NOVASpeech is the set of features offered to navigate and inspect the 3D models of the patients, including the possibility of using the 3D models for facial anthropometry. The survey included questions to evaluate the usefulness of these features. The volunteers could rate the features with a 5-point Likert scale.

Figure 11 shows the results for the 3D visualization of exercises in rest and with motion (like stretching the lips, or speech exercises), where 1 corresponds to *not useful* and 5 corresponds to *very useful*. The results show that the 3D visualization is very useful for dynamic exercises, with 68.2% of the participants rating it at the maximum and no participant rating it below 3 (neutral). While the participants

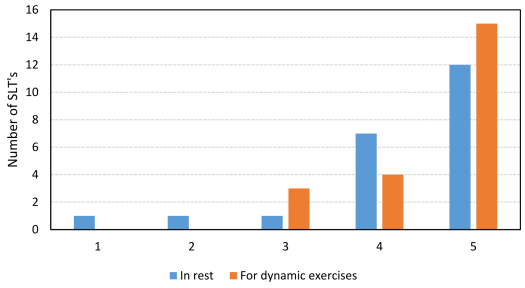


Figure 11: Utility of using the 3D visualizer

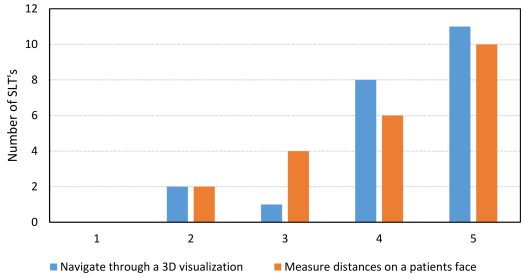


Figure 12: Usability of the 3D visualization and anthropometry

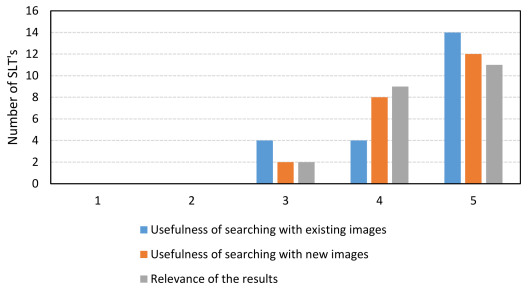


Figure 13: Usefulness of search and relevance of results

considered that the 3D visualization to examine the patient in rest was less useful, the classification was still high enough to allows us to conclude that it has its usefulness.

Most volunteers (86.3%) rated the navigation through the 3D models above 4 (figure 12), where 1 corresponds to the *very difficult* and 5 corresponds to the *very easy*. Also, most volunteers (95.5%) found the possibility of going through the 3D frames one by one very useful. This type of navigation allows the SLTs to see nearby frames almost instantly without losing context of the exercise.

Whilst most users rated measuring the distances for facial anthropometry above 4, for some users this feature was considered slightly less easy to use than navigating the 3D models (figure 12). While observing the volunteers exploring these features, we noticed that they had some initial difficulty manipulating the 3D frames, namely for rotating and zooming the frames. Nonetheless, it is also important to note that no participant thought that either task was very difficult to use. From these results and observations from the participant SLTs, we were able to conclude that while there is room

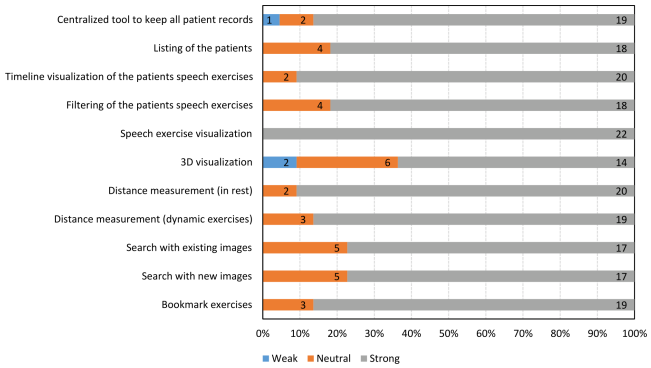


Figure 14: Usability evaluation of NOVASpeech

for improvement in the manipulation of the 3D frames, the SLTs were eager to use the system.

5.3 Search of similar exercises

Another key feature of NOVASpeech is the search of exercises by similarity to a query image. We asked the SLTs to rate the usefulness of searching with both imported images and images already in the dataset. Figure 13 shows the results obtained. With no votes under 3 in a scale of 1 to 5 (where 1 is *less useful* and 5 is *more useful*), it is clear that the SLTs consider both types of search as being similarly important and relevant.

We were also able to conclude that the search algorithm is able to retrieve relevant samples. Figure 13 shows that all volunteers rated the relevance of the retrieved samples above 3 (where 1 is *less relevant* and 5 is *more relevant*) and with an average rate of ≈ 4.41. When asked if they would use the search functionality, all participant gave a positive answer. These results allowed us to be certain of the quality of this feature as well as the utility it has in speech therapy.

5.4 Usability assessment

The survey included questions to evaluate NOVASpeech as a whole and assess its strong, neutral and weak points. Figure 14 shows which features the volunteers find that are the strongest and weakest characteristics of NOVASpeech. Speech exercise visualization is clearly the main functionality of NOVASpeech. General repository features for searching, managing, bookmarking, and filtering were all rated high by the volunteers. The results of the survey show that the tool is accessible to the end users as $\approx 82\%$ of the volunteers rated it 4 or 5 on a scale of 1 to 5 (with 5 being the easiest). However, we should not ignore the fact that two subjects voted 2 and other two subjects voted 3. It is important to make more tests and improve the areas where NOVASpeech is less accessible.

Among all the features, the 3D visualization was the one with less positive opinions. Observation showed us that controlling a 3D model is still an unusual experience, and requires some training. The volunteers were not used to that type of controls, leading them to feel that it is not as easy to use as the other features (section 5.2). Still, while the 3D visualization was rated as a strong feature by less participants than the other features, more than half of the participants (63.6%) rated it high. Also, it is interesting to note that

the answers to this question contrast with the answers about the 3D anthropometric features, which was rated as a strong feature by 90.9% of the participants. The answers to these questions allowed us to conclude that all the implemented features provide important functionalities to the SLTs.

The last question was about the SLTs' interest in sharing patients' data with colleagues to help them in their decision process. Almost all the volunteers (91%) were interested in this functionality, which allowed us to understand that sharing data between SLTs is important.

6 CONCLUSIONS

We proposed NOVASpeech, a 3D video management and retrieval system directed to assist the SLTs. The system supports multimodal health records and provides several different features that were designed to contribute to speech and language therapy as a whole.

Apart from allowing SLTs to maintain a database of their patients' recorded therapy exercises, the tool provides visualization and navigation features that help SLTs on inspecting and analysing their patients' data. These features allow SLTs to more easily assess the patients' disorders and monitor their progress.

The major novelty of NOVASpeech is the visualization and annotation of the 3D models of patients' faces. To the best of our knowledge, this is the first tool for speech and language therapy that offers the possibility of visualizing and navigating 3D data, allowing a type of examination that was not possible before. Moreover, NOVASpeech allows to take facial measurements for facial anthropometry in a non-intrusive way. This is an innovation that has been received with enthusiasm by SLTs who experimented the tool.

Another significant feature offered by NOVASpeech is searching similar exercises. Given a query image, NOVASpeech returns data from recorded exercises that have similar facial anthropometrics. This feature allows SLTs to support their clinical decisions based on similar cases.

In order to assess the offered features, we run a user study with 22 volunteers. NOVASpeech was very well received by the participant SLTs, who showed interest in acquiring the tool. The results showed that most features were well designed and easy to use.

Finally, another significant contribution of our work is a 3D video dataset of orofacial speech exercises used for assessing the patients: (1) rest position, (2) stretching the lips, (3) protrusion of the lips, and repeatedly saying the sounds (4)"pa", (5) "ba" and (6) "ma". Overall, the set of features implemented by NOVASpeech and the characteristics of the dataset, introduce several new advances and materials that are unique in relation to the state-of-the-art.

ACKNOWLEDGEMENTS

This work has been partially funded by the CMU|Portugal research project BioVisualSpeech, reference CMUP-ERI/TIC/0033/2014 and by the project NOVA LINCS Ref. UID/CEC/04516/2013. We would like to thank Cátia Pedroso who was essential in guiding the recordings sessions and to all the volunteers at ESSA who participated in the 3D video recordings.

REFERENCES

[1] B. Basnet, P. Parajuli, R. Singh, P. Suwal, P. Shrestha, and D. Baral. An anthropometric study to evaluate the correlation between the occlusal vertical dimension and length of the thumb. volume 7, page 33—39. Dove Medical Press Limited, 2015.

[2] A. A. Bui, W. Hsu, C. Arnold, S. El-Saden, D. R. Aberle, and R. K. Taira. Imaging-based observational databases for clinical problem solving: the role of informatics. *Journal of the American Medical Informatics Association*, 20(6):1053–1058, 2013.

[3] D. Cattoni. O uso do paquímetro na avaliação da morfologia orofacial. *Revista Soc. Bras. Fonoaudiologia*, 11(1):52–58, 2006.

[4] S. a. Chatzichristofis and Y. S. Boutalis. CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5008 LNCS, pages 312–322, 2008.

[5] S. a. Chatzichristofis and Y. S. Boutalis. FCTH: Fuzzy Color and texture histogram a low level feature for accurate image retrieval. *WIAMIS 2008 - Proceedings of the 9th International Workshop on Image Analysis for Multimedia Interactive Services*, pages 191–196, 2008.

[6] J. Chen, Y. Ariki, and T. Takiguchi. Robust facial expressions recognition using 3d average face and ameliorated adaboost. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 661–664. ACM, 2013.

[7] G. V. Cormack, C. L. a. Clarke, and S. Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09*, page 758, 2009.

[8] D. DeCarlo, D. Metaxas, and M. Stone. An anthropometric face model using variational techniques. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 67–74. ACM, 1998.

[9] T. Furuya and R. Ohbuchi. Diffusion-on-manifold aggregation of local features for shape-based 3d model retrieval. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 171–178. ACM, 2015.

[10] I. Guimarães, C. Birrento, C. Figueiredo, and C. Flores. *Manual do Teste de Articulação Verbal*. Oficina Didáctica, 2014.

[11] Intelligent Video Sollutions. Video Recording for Speech Therapy Training.

[12] O. A. Jiménez-del Toro, A. Hanbury, G. Langs, A. Foncubierta-Rodríguez, and H. Müller. Overview of the visceral retrieval benchmark 2015. In *Multimodal Retrieval in the Medical Domain*, pages 115–123. Springer, 2015.

[13] M. Lopes, J. a. Magalhães, and S. Cavaco. A voice-controlled serious game for the sustained vowel exercise. In *Proceedings of the 13th International Conference on Advances in Computer Entertainment Technology*, ACE2016, pages 32:1–32:6, New York, NY, USA, 2016. ACM.

[14] A. Mourão and J. a. Magalhães. Competitive affective gaming: Winning with a smile. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 83–92, New York, NY, USA, 2013. ACM.

[15] K. F. O'Grady and M. Antonyshyn. Facial asymmetry: three-dimensional analysis using laser surface scanning. *Plastic and reconstructive surgery*, 104(4):928–937, 1999.

[16] M. A. Rahman, D. Hossain, A. M. Qamar, F. U. Rehman, A. H. Toonsi, M. Ahmed, A. El Saddik, and S. Basalamah. A low-cost serious game therapy environment with inverse kinematic feedback for children having physical disability. In *Proceedings of International Conference on Multimedia Retrieval*, page 529. ACM, 2014.

[17] R. R. Riely and A. Smith. Speech movements do not scale by orofacial structure size. *Journal of Applied Physiology*, 94(6):2119–2126, 2003.

[18] M. Stürmer, A. Maier, J. Penne, S. Soutschek, C. Schaller, R. Handschu, M. Scibor, and E. Nöth. 3D Tele-Medical Speech Therapy using Time-of-Flight Technology. pages 1500–1503. 2009.

[19] A. R. Tate, N. Beloff, B. Al-Radwan, J. Wickson, S. Puri, T. Williams, T. Van Staa, and A. Bleach. Exploiting the potential of large databases of electronic health records for research using rapid search algorithms and an intuitive query interface. *Journal of the American Medical Informatics Association*, 21(2):292–298, 2014.

[20] B. Temkin, E. Acosta, P. Hatfield, E. Onal, and A. Tong. Web-based three-dimensional virtual body structures: W3d-vbs. *Journal of the American Medical Informatics Association*, 9(5):425–436, 2002.

[21] S. M. Weinberg, S. Naidoo, D. P. Govier, R. A. Martin, A. A. Kane, and M. L. Marazita. Anthropometric precision and accuracy of digital three-dimensional photogrammetry: comparing the genex and 3dmd imaging systems with one another and with direct anthropometry. *Journal of Craniofacial Surgery*, 17(3):477–483, 2006.

[22] J. Xie, F. Zhu, G. Dai, and Y. Fang. Progressive shape-distribution-encoder for 3d shape retrieval. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, pages 1167–1170. ACM, 2015.

[23] Z. Zhang, W. Zhang, J. Liu, and X. Tang. Facial landmark localization based on hierarchical pose regression with cascaded random ferns. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 561–564. ACM, 2013.