AUTOMATIC EVALUATION OF DYSARTHRIC SPEECH AND TELEMEDICAL USE IN THE THERAPY

Elmar Nöth[1], Andreas Maier[1], Arnd Gebhard[1,2], Tobias Bocklet[1], Wilfried Schupp[3], Maria Schuster[4], Tino Haderlein[1]
[1]University of Erlangen-Nuremberg, Germany, [2]Siemens Healthcare, Forchheim, Germany, [3]m&i Fachklinik Herzogenaurach, Germany, [4]Klinikum der Universität München, München, Germany
Contact: Elmar Nöth, Lehrstuhl für Mustererkennung, Martensstr. 3, 91058
Erlangen, Germany, noeth@cs.fau.de, +49 9131 8527888

Abstract

After a stroke, the speech quality of the patients is often reduced. This is usually caused by a deficit of the motor abilities of the vocal tract. The result is slurred speech. In the various patients, however, very different forms can appear. In the course of therapy, evaluation of the speech quality is required to determine the success of the treatment. At the moment, this assessment is performed only perceptually. This form of assessment is subject to strong intra- and inter-individual variation. Therefore, an "objective" assessment is not guaranteed. In this study, we present a rater-independent method for evaluating speech disorders in dysarthria. We use methods of automatic speech recognition. The idea is to determine the speech intelligibility – the main outcome parameter of speech – automatically by an automatic speech recognizer. A correlation of -0.89 was obtained between the criterion "intelligibility" and the recognition rate of the automatic system, in a preliminary study. The second part of this paper deals with an additional problem with this kind of patient. Very often, the stroke leads to partial facial paresis and generally to reduced mobility. Therefore, it is desirable that therapy sessions are performed in a telemedical setup. We report on our work towards such a telemedical diagnosis and rehabilitation system which will allow sessions with a therapist and – at the same time – diagnose the patient and track the recovery process. We describe the equipment (web camera, 3-D camera, stereo microphone, and Internet connection), the patient's environment, and the working environment of the therapist. Depending on the network connection, live images of the patient can be sent to the therapist at a rate of 20 frames per second (fps) at existing LAN connections or 3 fps with a DSL 6000 connection. At the patient host, a three-dimensional face model of the patient performing a facial exercise can be generated and transferred to the therapist in real-time (LAN) or three times real-time (DSL 6000).

1 Introduction

The quality of the speech in patients after a stroke is often reduced (Urban et al. 2001). A deficit of the motor abilities of the vocal tract is usually the cause for "slurred" speech. There is a considerable variance in the speech outcome across different patients. In the course of therapy, evaluation of the speech quality is required to determine the success of the treatment (Ludlow 1994). There is no objective, validated, automated procedure for the determination of the speech intelligibility in patients with dysarthria. The perceptual assessment of the intelligibility by speech therapists is not objective and, therefore, subject to inter- and intra-individual variation. In particular, experience is a crucial factor (Paal et al. 2005). In order to obtain a more reliable assessment, patients are often evaluated by an expert committee or panel. However, this is usually performed only for clinical studies and research, because a lot of time and effort are required. In this paper, we first present the use of an automatic speech recognition system to evaluate the intelligibility. Furthermore, we use automatic prosodic features which are also extracted from the speech signal and compare them with a number of other perceptual criteria. There is some previous work on automatic evaluation of dysarthric speech. Van Nuffelen et al. (2009) describe an automatic evaluation method based on phonemic and phonological features. They reach correlations between their objective and perceptual phoneme intelligibility scores from about 0.8 for the two feature groups alone to 0.94 for a combination of the feature groups. Their system is not yet combined with 3-D information (see below) which is favorable for the detailed description of pathologic speech production nor is it provided via internet for telemedical application. For the German language, Ziegler and Zierdt (2008) developed a telemedical system for the evaluation of intelligibility which demonstrated high reliability. They use perceptive evaluation of unknown speech leading to the quantification of speech disorder in accordance to Schiavetti's claim to quantify the percentage of intelligible words of a word sequence (Schiavetti 1992). However, this telemedical system is based on manpower and therefore includes a time lag until results are given.

The second part of the paper deals with our current work to integrate these diagnosis tools into a telemedical diagnosis and rehabilitation system where the patient can perform a therapy session at home. This is especially important, if the patient has reduced mobility. Apart from analyzing the speech of the patient, we have to provide a fast real-time transmission of the speech signal to the therapist in such a setup. In order to be able to better evaluate the facial expressions of the patient (especially asymmetries due to facial paresis), we create a 3-D model of the patient's head so that the therapist can have a side view of the patient doing facial exercises. In this paper we concentrate on the transmission speed of our system. The rest of the paper is organized as follows: In Chapter 2, we describe the method for the evaluation of dysarthric speech. In Chapter 3, we describe the patients that we used for the acoustic analysis in this pilot study. The results are presented in Chapter 4. In Chapter 5, we deal with the extension of the system to a telemedical system. In Chapter 5.1., we describe the technology to acquire 3-D information. In Chapter 5.2., the patient's work place is looked at in more detail, and in Chapter 5.3., results concerning the real-time properties with different Internet connections are presented. The paper ends with an outlook and summary.

2 Evaluation of Dysarthric Speech

The speech data were recorded over the Internet with our "Program for Evaluation and Analysis of All Kinds of Speech Disorders" (PEAKS, Maier et al. 2009). PEAKS runs in any Internet browser and is based on Java technology which allows platform-independent use. The data are transmitted to our server and evaluated centrally (cf. Figure 1). PEAKS currently can only be used as an offline system and is currently not intended for telemedical use, i.e., the patient and the therapist are in the same location when the therapist accesses the PEAKS web page. Then a client is downloaded which displays the text to be spoken. The patient records his or her speech data which are locally stored by the client. After the recording, the data are securely transmitted to the server. There they are analyzed by a speech recognition system which is based on Hidden Markov Models (HMMs) and a prosody module. As training data for the speech recognition system, solely normal speakers were used. We did not include dysarthric training data in order to be able to better judge the deviation from "normal" speech. The training data were from the Verbmobil project (Wahlster 2000) and covered most regions of dialect of Germany.

The recognizer is described detail in Stemmer (2005). It was developed at the Pattern Recognition Lab (Lehrstuhl für Mustererkennung) of the University of Erlangen-Nuremberg. As features we use 11 Mel-frequency cepstrum coefficients (MFCCs) and the energy of the signal plus their first-order derivatives. The short-time analysis applies a Hamming window with a length of 16 ms, the frame rate is 10 ms. The 12 delta coefficients are computed over a context of two time frames to the left and the right side (56 ms in total). The recognition is performed with semi-continuous HMMs. The codebook contains 500 full-covariance Gaussian densities which are shared by all HMM states. We only used a unigram language model to weigh the outcome of each word model in order to put more weight on the recognition of acoustic features.

The result of the analysis is the number of correctly recognized words with respect to the reference (word correctness WC).

$$WC = (C / R) * 100\ \% \qquad (1)$$

C denotes the number of correctly recognized words, and R is the number of words in the reference.

Furthermore, automatic prosodic features which model energy, fundamental frequency, length of voiced and voiceless segments, jitter, and shimmer were investigated. The prosody module is described in detail in Zeißler et al. (2006). In our case, it takes the forced time alignment of the text to be read (not the recognized text) and the speech signal as input. Thus, the timing information and information about the underlying phoneme classes (such as long vowel) can be used by the prosody module. First, the prosody module extracts the so-called basic features from the speech signal with a frame rate of 10 ms. These are the energy, the fundamental frequency (F0), and the location of voiced and unvoiced segments in the signal. In a second step, the actual prosodic features are computed to model the prosodic properties of the speech signal. For this purpose, a fixed reference point has to be chosen for the computation of the prosodic features. We decided in favor of the end of a word because the word is a well-defined unit in word recognition. The end of a word can be provided by any standard word recognizer, and therefore this point can more easily be defined than, for example, the middle of the syllable nucleus in word accent position. For each reference point, we extract 21 prosodic features. These features model F0, energy, and duration, e.g. the maximal F0 in the current word. In addition, 16 global prosodic features for the whole utterance are calculated. They cover the mean and standard deviation for jitter and shimmer and information on voiced and unvoiced sections. The last global feature is the standard deviation of the fundamental frequency F0. In order to evaluate pathologic speech on a test level, we calculate the average, the maximum, the minimum, and the variance of the 37 turn- and word-based features for the whole text to be read. Thus, we get 148 features for the whole text. A more detailed description of the automatic speech evaluation system is given in Maier et al. (2009). As the speech evaluation system is more or less unchanged, we would like to focus on the novel combination of the 3-D camera system with the speech evaluation system (see below).

The results of the speech and prosody analysis are available shortly after recording. To compare the human and the automatic evaluation, four speech therapists with at least five years of experience rated the criteria “intelligibility”, “roughness”, and “prosody”. The ratings were performed using a five-point scale assessment for each criterion, and the average per patient was computed in order to obtain rater-independent scores. The agreement between the human and the automatic evaluation was determined as a Pearson (1896) correlation.

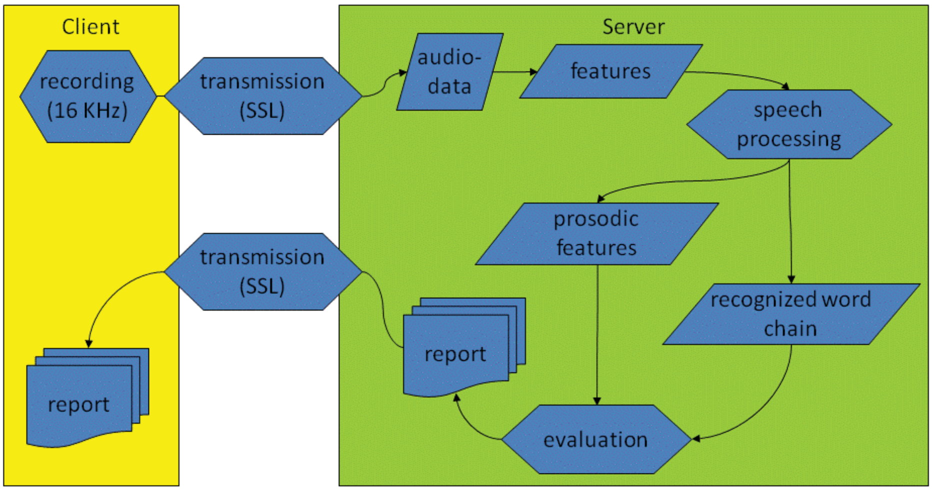

Figure 1. Diagram of the client-server architecture of PEAKS (Maier et al. 2009)

3 Patients

For this study, 28 patients with dysarthria were recorded during post-stroke rehabilitation. The patients were 39 to 76 years old. Depending on the severity of the dysarthria, the treatment can take 3 to 18 weeks, with an average duration of 5 weeks. Written informed consent was obtained from all patients participating in the study prior to the examination. Approval was received by the ethical standards committee on human experiments using human subjects at the University Clinic Erlangen.

The data consisted of reading a standard text, the German version (http://de.wikipedia.org/wiki/Die_Sonne_und_der_Wind) of “The North Wind and the Sun” (http://en.wikipedia.org/wiki/The_North_Wind_and_the_Sun). Overall, the text contains 108 words, of which 71 are disjoint. It is widely used in speech therapy in Germany. Figure 2 shows the recording setup using a standard PC at the m&i Fachklinik Herzogenaurach. The audio data were collected using lapel microphones. One is attached to the clothes of the therapist (on the right) and one was attached to the patient’s clothes (on the left). This procedure allows segmenting the speech of both easily. All analyses were performed on the patient’s audio data only.

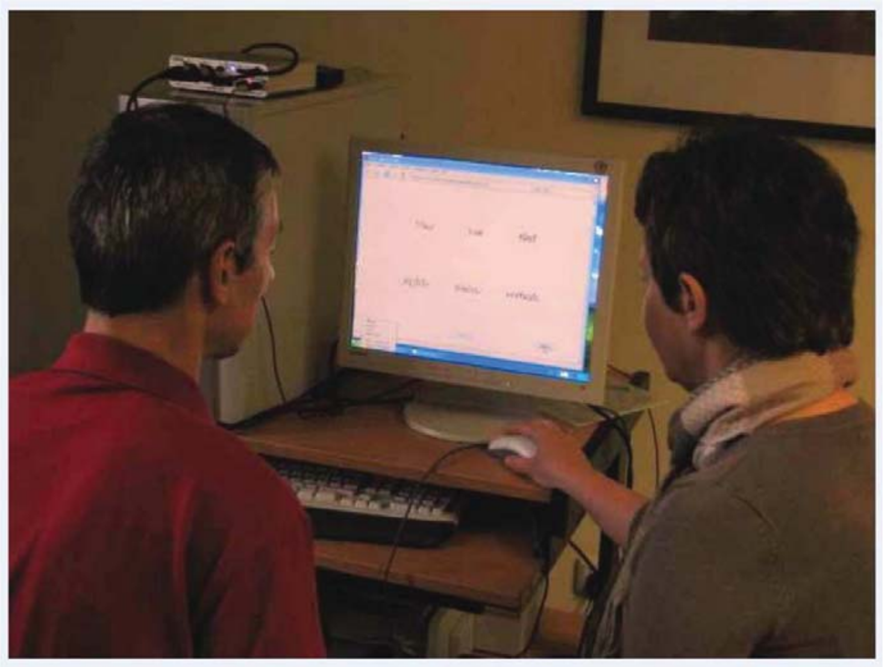

Figure 2. Recording setup

4 Automatic Evaluation of Dysarthric Speech

The perceptual evaluation of the human raters proved to be very consistent. The inter-rater correlation, i.e. the correlation of one rater and the average of the other three raters was in the range of 0.75 and 0.80. The average of all four raters was 0.78. The results of the evaluation of one rater compared to the average score of the other three raters is depicted in Figure 3. Small amounts of uniform noise are added for better visualization purposes.

The mean of all four experts was used as the reference to train an automatic system. As depicted in Figure 4, there was a significant correlation of $r = -0.84$ between the perceptual assessment of the intelligibility of the four human raters and the word correctness of the automatic speech recognition system ($p < 0.01$). The higher the word correctness, the smaller the human score should be, because '1' on the assessment scale means 'very good', and '5' means 'very bad'. Therefore, the negative correlation is expected. The high correlation is in line with previous studies (Maier et al. 2009).

The evaluations of the criterion "prosody" correlated with the ratio of the length of voiced and voiceless segments, $r = 0.82$ ($p < 0.01$). "Roughness" and the average number of voiceless segments correlated with $r = 0.81$ ($p < 0.01$). The correlation with "jitter" was only $r = 0.66$ (the computation of jitter is described in Levit et al. 2001). Also on the criteria "prosody" and "roughness", high, significant correlations between human evaluation and automatic prosodic features were found. The relationship of "prosody" and the ratio of the length of voiced and voiceless segments can be explained by the fact that both features are related to accentuation in speech. The correlation between "roughness" and the average number of voiceless segments is also plausible: High roughness may disturb the automatic fundamental frequency extraction algorithm, resulting in an erroneous classification of voiced signals as voiceless. Hence, a long voiced segment may be divided into several short voiced and voiceless segments. Eventually, this increases the number of voiced and voiceless segments. This hypothesis is supported by the observation that the number of voiced segments also correlates at r = 0.80 (p < 0.01) with “roughness”.

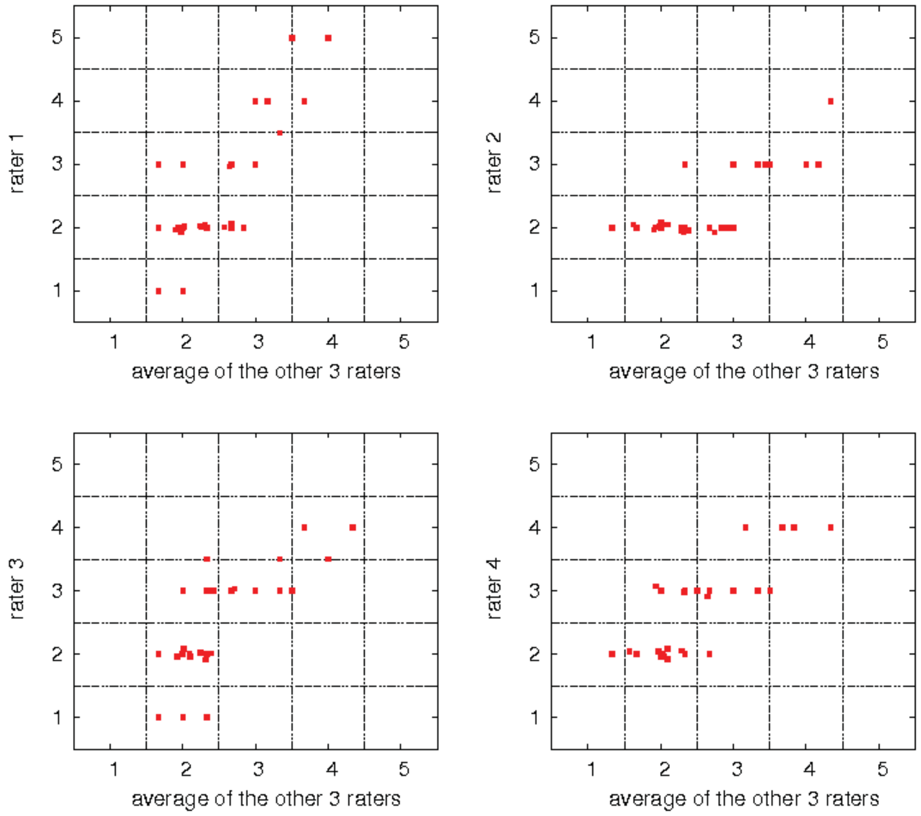

Figure 3. Correlation of one of the four experts and the average of the other three raters

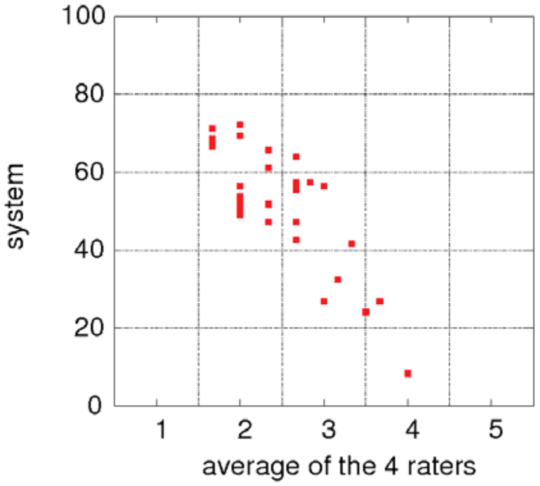

Figure 4. The correlation between the human experts and the automatic system is high (r = –0. 84) and significant (p < 0. 001)

5 Towards a Telemedical System

We have shown that an automatic diagnosis of the intelligibility can be done and this information provides a second opinion to the speech therapist. However, so far, the therapist still has to be present in person. In many cases, this poses a problem since the patients are often immobile because of a stroke that caused the dysarthria in the first place. Thus, it might be very difficult for the patient to come to the therapist's office. An alternative is that the therapist comes to the patient. This, however, is very costly because a highly trained specialist spends a significant amount of time on the road, especially in rural areas. This is why we have been working on a telemedical solution. Telemedicine is a rapidly developing application of clinical medicine where medical information is transferred through interactive audiovisual media for the purpose of consulting, and sometimes remote medical procedures or examinations. Apart from the obvious ease for the patient and the reduction of costs, a telemedical therapy might actually show better results than a face-to-face therapy. The reason is that the patient is aware of the fact that he has lost fundamental capabilities by only being able to produce slurred speech or not being able to close both eye lids. It might well be that the patient is more uninhibited to show his or her "weakness" to a therapist who sits 300 km away than to perform the exercise face-to-face with a therapist (whom he might even know personally).

In our scenario, we want the speech therapist to work in a hospital and the patient to be temporarily provided with a laptop with audiovisual equipment and fast Internet access. Since the stroke can lead to facial paresis, the visual impression is very important. Dysarthria is a complex and varying disease characterized by a lack of motor control for orofacial, sometimes also respiratory and laryngeal movements, which needs adapted therapy concepts according to the etiology and phenomenology. Quite often other motoric restrictions are combined and locomotion is limited. That's where telemedical therapy will fit perfectly: to provide medical support to dysarthric patients who cannot visit a speech therapist. The therapist then hears and sees the patient's articulation and movements and can teach the patient via internet. Of course, therapy then is limited to verbal teaching without thermic or tactile stimulation except for including a naive "co-therapist" who is guided via internet. However, in order to better judge how well a patient can perform an exercise like pursing the lips, the therapist should – at least in some parts of the therapy session – have a 3-D view of the patient. This is why we work on a telemedical system with capabilities far beyond a videophone system. The telemedical system should provide a real-time audiovisual communication of the patient with the therapist, if necessary, provide a 3-D view of the patient's face, and monitor the progress of the patient with respect to the intelligibility of his or her speech and the quality and symmetry of facial movements. In this way the system can provide help and guidance for computer-assisted practice between the therapy sessions in addition to providing a platform and second opinion for the telemedical therapy sessions. Basis of the new system is our PEAKS client-server platform, which has to be extended with respect to multimodality and real-time capabilities. Figure 5 shows the architecture of the intended telemedical system. The client system for the patient consists of a PC or laptop, Internet connection, stereo microphones, a webcam, a 3-D camera and an illumination source to better control the visual scene. The therapist has the same equipment except for the 3-D camera. In the next chapter, we have a look at the 3-D camera which is based on the time-of-flight principle.

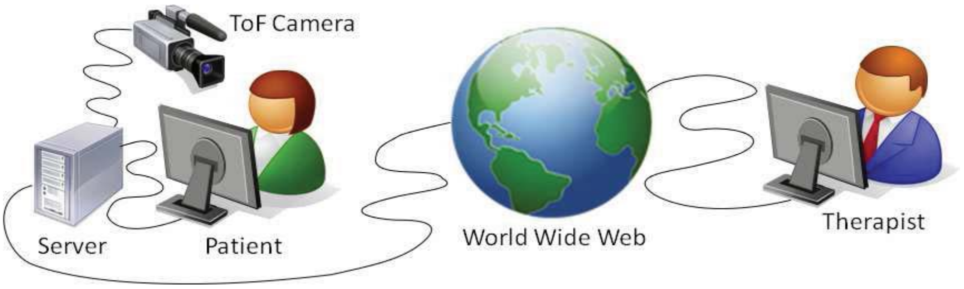

Figure 5. Schematic overview of the telemedical system (Stürmer et al. 2008)

5.1 Time-of-Flight Imaging

Time-of-flight (TOF) imaging is an emerging technology that provides a direct way to acquire 3-D surface information with a single sensor. Active light sources attached to the camera emit an incoherent cosine-modulated optical signal in the non-visible spectrum of the infrared range (850 nm). The light is reflected by the scene and enters the monocular camera where each TOF sensor element performs a correlation of the local optical signal with the electrical reference of the emitted signal (Xu et al. 1998). Based on this correlation, the phase shift *Φ* representing the propagation delay between both signals is measured. The distance *d* can then be computed straightforward,

$$d = \frac{c}{2 \cdot f_{mod}} \cdot \frac{\Phi}{2\pi} \qquad (2)$$

where f_{mod} denotes the modulation frequency, *c* the speed of light. For reasons of periodicity of the cosine-shaped modulation signal, the validity of this equation is limited to distances smaller than $c/(2 \cdot f_{mod})$. At a typical modulation frequency of 20 MHz, the non-ambiguity range is about 7.5 m. The TOF imaging technology benefits from several advantages over other 3-D surface acquisition techniques. The device is compact, portable and easy to integrate. It also provides precise metric information in the sensor coordinate system in real-time, and no calibration steps are necessary. With respect to potential applications in the security, automotive and consumer electronics industry (Kolb et al. 2009), a decrease of manufacturing costs can be expected with mass production being an all-solid-state off-the-shelf technology. Figure 6 shows the principle of a TOF camera.

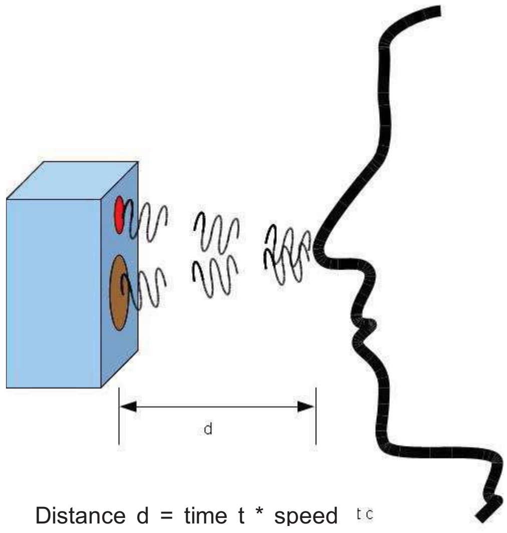

Figure 6. Operating mode of a time-of-flight camera

5.2 A Detailed Look at the Patient's Work Place

Figure 7 shows a patient during a session. The light is directed towards the patient's face to have controlled illumination. The speech signal is recorded via 2 stereo microphones at 22.05 kHz (16 bit). The webcam records a 640x480 pixel color image stream (24 bit per pixel). The TOF camera produces a 176x144 pixel depth map (16 bit depth information, 8 bit intensity). The patient either sees the face of the therapist (whenever the therapist wants to demonstrate an exercise) or a control image of himself. The patient can check whether the illumination, the position, and the distance to the recording devices are in a valid range, because this is indicated with an ellipsis around the face. For this the localization of the face both in the 2-D webcam image and the 3-D TOF image is necessary. We use the Viola-Jones Algorithm (Viola & Jones 2001). Normally, only the webcam image stream and the speech signal are transmitted to the therapist in real-time. The number of pictures per second depends on the Internet connection (see below). For certain exercises, e.g., pursing the lips, a 3-D model of the face is created. For this, the TOF depth map and the webcam are registered. After the exercise, this 3-D model is transferred to the therapist. For an exercise lasting about 3 seconds, ca. 480 kB of data have to be transferred. Figure 8 shows a test person during the exercise "showing the teeth". Here, the viewing direction and the time change from left to right are visible. In the leftmost image, the mouth is still closed; in the central image, the teeth are visible, and in the right image, the mouth is closed again. The viewing direction can be altered by the therapist using the mouse.

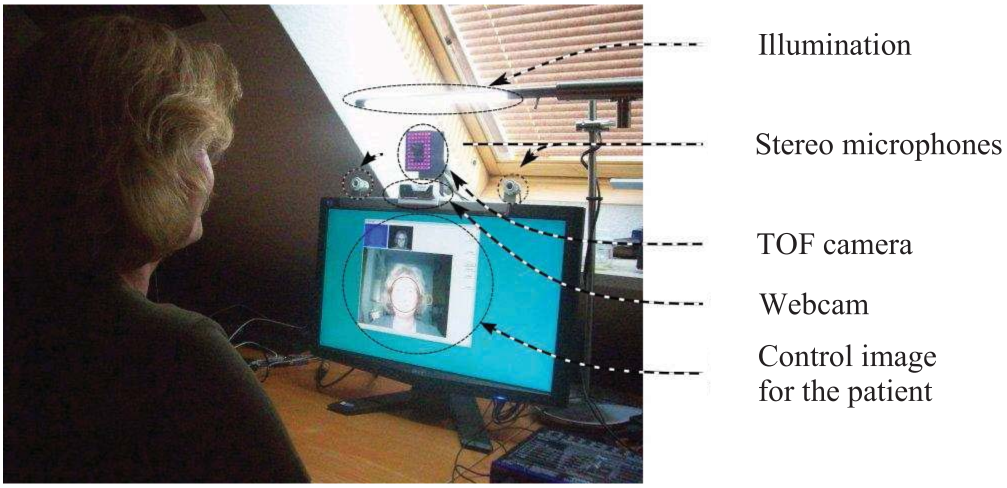

*Figure 7.*The recording station for the patient

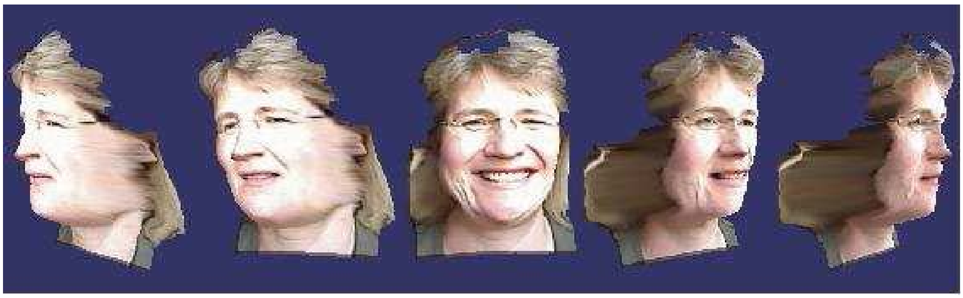

Figure 8. Face of a test person during the exercise “showing the teeth” from different viewing directions

Figure 9 shows the work place of the therapist. In the normal mode, only the 2-D image stream is shown (left image in Figure 9). During an exercise, a 3-D face model is computed which is transferred to the therapist (right image in Figure 9) who can then look at the exercise from different viewing angles.

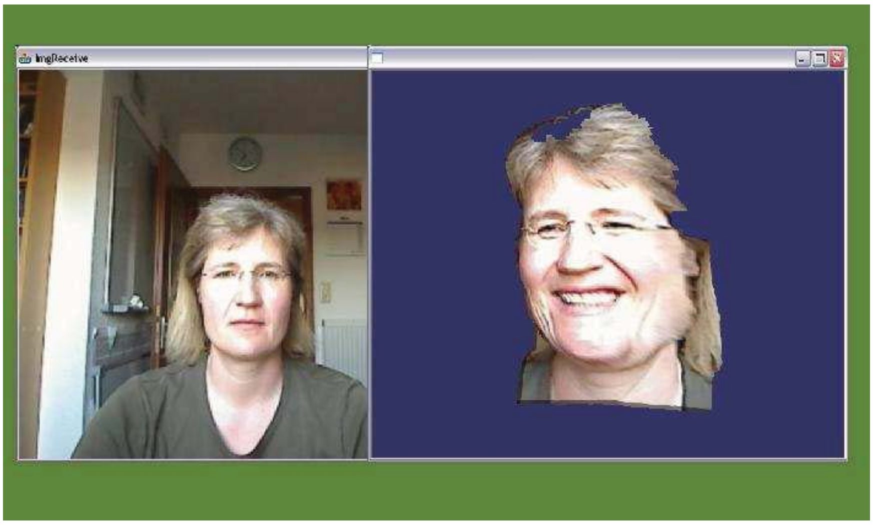

Figure 9. Screen of the therapist’s work place

5.3 Results Concerning Transmission Speed

In the last section, we have described some extensions of PEAKS towards a telemedical system. We have created a mobile workplace that can be used from the home of a patient or in a rehabilitation clinic with a remotely located therapist performing a therapy session or checking whether a patient is correctly performing exercises performed between therapy sessions. In this section, we will describe some experiments concerning the transmission speed of such a telemedical system. For this we performed some test therapy sessions where the test person first read "The North Wind and the Sun" and then performed 4 typical facial exercises which lasted about three seconds each:

Raising of the eyebrows
Closing of the eyes
Showing the teeth
Pursing of the lips

Even though we have a module that automatically classifies the asymmetry of the face during these exercises (Gebhard et al. 2001), we will only report here on the transmission speed with different Internet connections. Since we only used volunteers without facial pareses during the development of our system, we tested three different scenarios:

The work place of the patient and the therapist are in one LAN. Transmission speed is 100 MBit (scenario LAN).

The work place of the patient and the therapist are connected via Internet with a DSL 6000 connection (scenario DSL).

The work place of the patient and the therapist are connected via Internet with a UMTS stick for the patient's computer (scenario UMTS).

Using the parameters described in Chapter 5.2., we measured the number of images per second for reading the text and the transmission time for the 3-D models of the exercises. As can be seen in Table 1, the UMTS scenario is not yet applicable without much stronger compression for telemedical applications. In the DSL and the LAN scenario, the transmission speed is fast enough to perform some tests with real patients.

Table 1. Transmission speed for different Internet connections

	Live images per sec.	Duration of 3-D transmission (in sec.)
Scenario LAN	20	1
Scenario DSL	3	10
Scenario UMTS	2	24

6 Outlook and Summary

We have presented some preliminary work towards a computer-assisted telemedical diagnosis and therapy system for people with dysarthria. The system is intended for real-time telemedical therapy sessions or for computer-assisted exercise sessions. In both setups, the transmission speed is of crucial importance and it is desirable that an automatic analysis of the patient's performance can be done, both with respect to the acoustic signal and the symmetry properties of the face. In the therapy session scenario, the automatic analysis acts as a second, objective opinion, by diagnosing the speech of the patient with respect to properties like "intelligibility", "prosody", and "roughness" and assessing the symmetry properties of the face during exercises after facial paresis. The system also automatically documents the progress of the therapy. In the exercise scenario, the system has to "supervise" the patient during the exercises that he performs between two therapy sessions and summarizes these sessions for the therapist. In a pilot study, we have shown that the acoustic analysis is in high agreement with the judgment of a panel of experienced speech therapists. We have also shown that the transmission speed is acceptable for a LAN and a DSL scenario. This could even further be improved using compression of the depth data as proposed in Stürmer et al. (2008). We are currently conducting data collection in a rehabilitation clinic with dysarthric patients. This will allow us to verify these very encouraging results with respect to the speech analysis on a larger group of patients and to reevaluate and improve the classification of facial pareses described in Gebhard et al. 2001. There, only a 2-D image was available and we expect significant improvements with the additional 3-D information. During this study, we want to compare the telemedical setup with a face-to-face setup.

Furthermore, we will develop instruction dialogues so that the system can analyze exercises between therapy sessions. Finally, we want to improve the UMTS scenario, i.e., increase the compression rate until the transmission speed is acceptable and test whether the quality of the images is good enough to allow a telemedical live session.

References

Gebhard, A., Paulus, D., Suchy, B., Fucak, I., Wölfel, S. & Niemann, H. 2001. Automatische Graduierung von Gesichtsparesen. In H. Handels, A., Horsch, T. Lehmann, H.-P. Meinzer (eds.): *5. Workshop Bildverarbeitung für die Medizin.* Springer, Berlin. 352-356.

Kolb, A., Barth, E., Koch, R. & Larsen, R. 2009. Time-of-Flight Sensors in Computer Graphics. In *Eurographics 2009 - State of the Art Reports.* 119-134.

Levit, M., Huber, R., Batliner, A. & Nöth, E. 2001. Use of prosodic speech characteristics for automated detection of alcohol intoxication. In Bacchiani, M., Hirschberg, J., Litman D., & Ostendorf, M. (ed.): *Proceedings of the Workshop on Prosody and Speech Recognition 2001*, Red Bank, NJ. 103-106.

Ludlow, C. L. 1994. Motor Speech Disorders: Advances in Assessment and Treatment. *Neurology* 44, 2220-2221.

Maier, A., Haderlein, T., Eysholdt, U., Rosanowski, F., Batliner, A., Schuster, M. & Nöth, E. PEAKS – A System for the Automatic Evaluation of Voice and Speech Disorders. *Speech Communication* 51(5), 425-437.

Paal, S., Reulbach, U., Strobel-Schwarthoff, K., Nkenke, E. & Schuster, M. 2005. Beurteilung von Sprechauffälligkeiten bei Kindern mit Lippen-Kiefer-Gaumen-Spaltbildungen. *Journal of Orofacial Orthopedics* 66(4), 270–278.

Pearson, K. 1896. Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia. *Philosophical Transactions of the Royal Society of London* 187, 253–318.

Schiavetti, N. 1992. Scaling procedures for the measurement of speech intelligibility. In R.D. Kent (ed.): *Intelligibility in Speech Disorders: Theory, measurement and management*, John Benjamins, Philadelphia. 11-34.

Stemmer, G. 2005. *Modeling Variability in Speech Recognition.* Logos Verlag, Berlin.

Stürmer, M., Maier, A., Penne, J., Soutschek, S., Schaller, C., Handschu, R., Scibor, M. & Nöth, E. 2008. *3-D Tele-Medical Speech Therapy using Time-of-Flight Technology.* In *Proceedings of the 4th European Congress for Medical and Biomedical Engineering.* Antwerp, Belgium.

Urban, P. P., Wicht, S., Vukurevic, G., Fitzek, C., Fitzek, S., Stoeter, P., Massinger, C. & Hopf H. C. 2001. Dysarthria in acute ischemic stroke: Lesion topography, clinicoradiologic correlation, and etiology. *Neurology* 56:1022-1027.

Van Nuffelen, G., Middag, C., De Bodt, M. & Martens, J.-P. 2009. Speech technology-based assessment of phoneme intelligibility in dysarthria. *International Journal of Language and Communication Disorders* 44(5), 716-30.

Viola, P. & Jones, M. Robust Real-time Object Detection. In *Second International Workshop on Statistical and Computational Theories of Vision – Modeling, Learning, Computing, and Sampling.* Vancouver, 2001, no pagination.

Wahlster, W. (ed.) 2000. *Verbmobil: Foundations of Speech-to-Speech Translations.* Springer, New York, Berlin.

Xu, Z., Schwarte, R., Heinol, H., Buxbaum, B. & Ringbeck, T. Smart Pixel - Photometric Mixer Device (PMD) / New System Concept of a 3D-Imaging-on-a-Chip. In *5th International Conference on Mechatronics and Machine Vision in Practice.* 259-264.

Ziegler W. & Zierdt, A. 2008. Telediagnostic assessment of intelligibility in dysarthria: A pilot investigation of MVP-online. *Journal of Communication Disorders* 41(6), 553-557.

Zeißler, V., Adelhardt, J., Batliner, A., Frank, C., Nöth, E., Shi, R. & Niemann, H. 2006. The Prosody Module. In W. Wahlster (ed.): *SmartKom: Foundations of Multimodal Dialogue Systems*. Springer, Berlin. 139-152.

.