Speech development and therapy using the Kinect

S Frost, R J McCrindle

School of Systems Engineering, University of Reading, Whiteknights, Reading, UK

s.frost@microsoft.com, r.j.mccrindle@reading.ac.uk

www.rdg.ac.uk/sse

ABSTRACT

The use of computers and technology to treat patients with developmental problems or rehabilitation needs is an emerging field. Implementation of such treatment methods however has not traditionally been easy, requiring expensive equipment, significant programming experience and the time of trained medical professionals. The release of gaming systems with natural user interfaces has opened up new possibilities for creating home based therapy and rehabilitation systems that are more engaging, affordable and customisable to individual needs. This project leverages the high quality voice and facial recognition capabilities of the Microsoft Kinect natural user interface, and affordable hardware, to provide an interactive speech therapy application that can be used by patients in their own homes, whilst also collecting metric data for remote monitoring by medical professionals to ensure that engagement with, and appropriate progression of, treatment is occurring.

1. INTRODUCTION

In the UK, 2.5 million people, and in the US 7.5 million people, have a speech or language difficulty (RCSLT; NIDCD). Speech and language therapists assess and treat speech, language and communication problems in people of all ages to help them communicate better. Speech and language therapy is an extremely diverse field, treating problems in adults and children that may have been caused by a wide range of developmental conditions, diseases or disabilities such as stroke, learning disabilities, neurological disorders, head injury, deafness or dementia (Patient.co.uk). Individuals with developmental issues may require the assistance of trained therapists to promote correct development of speech, whilst sufferers of conditions such as stroke, and other brain traumas or diseases may require speech therapy in order to return their ability to speak properly.

As the field is diverse, methods of categorising different types of speech problems have been developed across medical practices which include for example (Stroke Association, 2011):

Aphasia which covers problems related to understanding or expressing speech and language correctly, and which is subdivided into component groups including Broca's (expressive) aphasia, Wernicke's (receptive) aphasia and global aphasia. Receptive aphasia is the inability of a person to understand language, or to make sense of the words that one is hearing and their meaning. Expressive aphasia is the inability to speak words, or determine the correct word to use when speaking, for example, jumbling sentences or mixing up similar words.

Dysarthria which covers problems forming particular sounds with the mouth during speech, due to weak muscles or the inability to properly control the muscles in one's mouth. This can often lead to others having difficulty understand a person as their speech may be slurred, imprecise or quiet. It may be developmental as a result of brain changes before or during birth or acquired as the result of a head injury or medical condition such as stroke or Parkinson's disease. Dysarthria is often present in patients who also suffer from expressive aphasia.

Dyspraxia/apraxia which refers to the inability to move the muscles required for speech in the correct sequence to form words and sounds properly. This is often shown by a patient's inability to correctly order mouth sounds in long, compound words especially when component sounds are similar, by making inconsistent mistakes when speaking, or by incorrectly varying the rhythms, stresses, and inflections of speech that are used to help express meaning.

Each of these different types of speech disorder has differing common methods of treatment and associated activities and exercises that patients perform to help combat their condition (Deane et al, 2001). These can range from flash cards, reading exercises or vocalisation tasks which are often performed at home, and then repeated

regularly with therapists to determine how a patient is progressing. Since these activities are often formulaic and well defined, they provide an appropriate target for simulation within computer-based applications and/or support for speech and language therapists. However, to correctly develop and apply software that is useful to therapists and beneficial to patients the appropriate methods, reference models and tools must be understood so that the software produced can be most effectively developed and exercises selected to match individual patients' therapy needs (Glykas and Chytas, 2004; Schröder et al, 2007).

2. SUPPLEMENTING SPEECH THERAPY THROUGH TECHNOLOGY

The last decade has seen a growth in the number of computer-based systems and specialized speech recognition software to supplement treatment of patients who have speech disorders stemming from developmental problems, or conditions such as post traumatic brain injury or stroke. AUDIX, for example, is a knowledge-based multimedia system that provides auditory discrimination exercises via a mix of visual stimuli (words, pictures) and audio (voice) clips to adult aphasic patients to help them discriminate between minimal pair sets of words (Grawemeyer et al. 2000). Another multimedia system designed for aphasic patients is the IMITATE system which has a pool of treatment stimuli consisting of audio-visual recordings of 2636 words and 405 phrases of varying complexity spoken by six individuals differing in gender, age and race (Lee et al. 2009). A Chochrane review of speech and language therapy for aphasia following stroke has been carried out by Kelly, Brady and Enderby (2010) and Dean et al (2001) have undertaken a comparison of speech and language therapy techniques for dysarthria in Parkinson's disease.

Saz et al. describe an interactive tool and speech recognition system to facilitate language development skills for children and young adults aged 11-21 with neuromuscular disorders such as dysarthria (Saz et al. 2009), and Umanski et al (2010) report a voice-based rhythm game targeted at children with speech motor disorders such as apraxia and dysarthria. Schipor (2014) have developed a framework, PhonEM, and a system, Logomon, designed to help correct children's inability to pronounce words due to dyslalia taking into account their emotional states during the computer based speech therapy sessions. Other systems such as SPREAD (Speech and Phoneme Recognition as an Educational Aid for the Deaf and Hearing Impaired) (Sadural and Carreon, 2012) and a CSLU vocabulary tutor (Kirschning and Toledo, 2004) are computer assisted programmes to aid developmental speech therapy for hearing impaired users.

An increasing range of interactive software packages are becoming available on the market for speech therapy covering common therapy activities and custom therapy pathways including Speech Sounds on Cue (Propeller), StepByStep (Steps Consulting Ltd) and SWORD (University of Sheffield). Some speech therapy software is highly specialized, and only licensable for commercial use meaning that the only access a patient has to the technology is through their therapist or medical institution, however other software packages and games, are becoming increasingly available for purchase (Bungalow Software) including those available for download to tablets and smartphones (Virtual Speech Centre).

As well as development for mobile platforms, recent developments in home and consumer technology such as the Nintendo Wii (Nintendo), PlayStation Move (Sony) and Microsoft Kinect (Microsoft a,b) have brought gaming and interactive systems into many homes. These hardware and software systems can provide novel and unique interactive opportunities, and run on high quality, yet affordable hardware. At present, many of the games and interactive activities available as mainstream entertainment packages are being used informally by relatives or carers of patients with speech problems. Such adoption does however highlight potential issues with this approach, and the software it uses. This is due to existing software and hardware platforms being developed primarily for use by people with no disability, and hence the input modes, phrases and interaction patterns may not be entirely suitable for patients with speech therapy needs, potentially causing negative reinforcement, incorrect pronunciation or repetition. Whilst there has been some specialised development of the Kinect for speech therapy applications (Mravak et al. 2013; Lanz et al. 2013) the field is still at an embryonic stage with the number of games developed to support speech therapy being far less than those developed to support motor rehabilitation.

The aim of this project has been to develop, using the Microsoft Kinect natural user interface, an engaging interactive therapy application prototype that can be deployed using cheap accessible hardware into patients' homes in order to supplement existing speech therapy plans. The developed application provides a number of activities, presented in an interactive and engaging manner, designed to replicate the functions of common therapy methods. While a patient is undertaking these activities, data can be collected and stored remotely about their performance, choices or other appropriate metrics. Provision of a local interface allows a therapist or carer to steer the session and to monitor a patients responses whilst they are undertaking the activities. Metric data from the session is available through a second application which visualises the data for medical professionals to review.

3. REQUIREMENTS

Based on reviews of the literature, product evaluations and discussions with speech therapists and language development experts a number of key considerations were identified as being important, in addition to incorporation of appropriate therapy exercises, for a Kinect based speech and language therapy system:

3.1.1 User Engagement: whilst there are many software packages available whose interactive pathways and activity style have been purposefully designed to help with speech problems some appear to have been designed with only the medical outcome in mind, providing a bland and uninspiring interface to the patient. This can cause the patient to disengage from the application, or begin neglecting its use, subsequently not gaining the full benefit of the software. Meetings with the therapists confirmed that one of the most challenging tasks for patients undergoing therapy for speech issues was maintaining engagement and interest in the exercises whether traditional or software based. Existing applications are often fundamentally useful, but can sometimes appear as outdated and visually or interactively unappealing when used for extended periods of time. This was especially true for younger patients, but was also identified as a problem for adults.

3.1.2 Suitability and ease of use in the home: provision of speech therapy in the home environment is useful in terms of providing continuity of therapy between face-to-face sessions and in terms of engagement. However, some current software has complex support requirements, and may be difficult to use for those without sufficient computer skills. The complex nature of the software can also require therapists to be present to operate the software, or the equipment used, not only to ensure the patient can access the software, but to ensure the correct course of treatment takes place.

3.1.3 Affordability: some of the current speech therapy packages are expensive and may require multiple home user licenses, and professional or clinician licenses which can become prohibitively expensive.

3.1.4 Customization: existing applications are often without the ability to freely and easily customize or tailor the exercises being undertaken to a particular patient or their condition. This may mean that the usefulness of these packages is diminished for border case patients or those with atypical symptoms.

3.1.5 Feedback: existing applications also often lack the ability to play back sample sounds and provide feedback as to the accuracy of the speech recognition or sound processing. Some applications incorporate videos to demonstrate proper mouth shape, but none appear to measure mouth shape as part of the feedback.

3.1.6 *Metrics:* data collection and patient progression measures are present in some applications in a very simple form, but often require the purchase of more expensive professional versions of the application. In the best cases timings and exercises are measured but there tends to be a lack of medically useful information recorded.

3.1.7 Integration: whilst existing software products offer appropriate features, they tend to focus on one aspect of the therapy process. Few currently available packages integrate speech exercises, metrics visualisation and storage, and real-time customisation in one integrated package.

These considerations stemming from the literature and product reviews and discussions with therapists demonstrate the need for a low cost application that is capable of interpreting speech, and providing feedback on the quality and completeness of spoken words.

4. OVERVIEW OF THE SPEECH DEVELOPEMNT AND THERAPY SYSTEM

4.1 Microsoft Kinect

The Microsoft Kinect is a natural user interface device that incorporates both video skeletal tracking and voice recognition technologies. It is able to interface with both the Microsoft Xbox games console and with personal computer devices. Version 1 of the sensor enables 20 joints and 40 facial points to be tracked, whilst the recently launched Kinect sensor version 2 enables 25 joints plus open/closed hand states to be recognised as well as additional facial expressions.

The voice recognition subsystem of the Kinect interface is a quad microphone array with built in low level pre-processing, designed to optimise the system for speech recognition. Microsoft provides an API which allows open use of a number of voice recognition libraries, including a specific library for the Kinect and the standard Microsoft Windows dictation library. This allows the Kinect voice recognition system to work in a number of environments for different purposes, depending on the library selected.

4.2 System Overview

The system developed is comprised of multiple components organised into four separate applications with different therapy related functions, running on distinct devices, and communicating over a local TCP/IP network. The 'Client Application' is the games-based programme that delivers the speech therapy and with which the patient interacts; the 'Data Application' is the data storage application that records and stores the metric data from each game session for a particular patient; the 'Supervision Device' is the interface a therapist uses on a tablet or smart phone to control the therapy exercises delivered to the patient; and the 'Data Application Front End' visualises the data stored in the database in order to display the metrics requested by the therapist. An overview of the system architecture is shown in Figures 1 and 2.



Figure 1. Distribution and network diagram.



Figure 2. System architecture diagram.

4.3 Interactive Element

A prototype application has been developed to help with a patients practice regime, targeting users who as part of their therapy must work through training sets of words such as those shown in Figure 3, in an effort to correctly pronounce each element of the word. Traditionally this is a repetitive task, which can be arduous to complete. To combat this, the prototype system presents the words in the form a game, in which the user collects points by pronouncing each segment, and then the entire word. When a certain number of points have been collected, they can be submitted to a leader board, or used to access games or activities. This is intended to give the user incentive to complete each word set.

The carnival flash card game shown in Figure 4 is a GUI (Graphical User Interface) component that presents users with words, broken into phonetic chunks as part of a fairground environment. These light up as the computer detects correctly spoken phonemes to show the user how they are progressing with pronunciation of the word. A power bar is also displayed on screen to show how correct an attempt is at any time. If the phoneme is correct, the application lights up a bar on the block, leading to the prize, in much the same way as a "strong man" machine works at a fairground when it is hit with a mallet. This is intended to provide the user with feedback and capture their attention as they work to complete the game.



Figure 3. Training set of words.

Figure 4. Interactive Game Interface.

4.4 *Metric Gathering*

In order to ensure that the application is providing correct and effective support to the patient, it is capable of collecting and reporting data to the speech therapist about each session the user undertakes. Currently, a large amount of metric data is collected relating to the progression of the user through each part of each word, with regards to the time they take, their accuracy of pronunciation and the percentage of words in the training set completed. It is intended that the data visualised will be further customised to provide precise and targeted data relevant to individual patients or plotted against identified norms, however actual data sets have still to be finalised in conjunction with the medical professionals and extensive user testing has yet to be completed.

As data is collected it is reported to a central server, where it is analyzed and moved into a database which keeps track of each user's session details. This database is intended to be private, and only accessible by doctors and therapists who will use it to record their level of engagement with their therapy programme, track the progress of a patient, and from the result determine the most appropriate course of therapy.



Figure 5. Example Metric Graph.

A small client program collects the data, and allows it to be browsed and graphed, or used to generate reports depending on the type of data as shown in Figure 5.

4.5 Mobile Device Supervision

The interactive environment also provides an interface, developed as part of the project, which is available to view on a smartphone and/or tablet device (Figure 5). The application delivered as part of the prototype allows a therapist (or carer) to connect to the session, and see past and current words and upcoming words and segments, or if they wish add custom words and segments while the application is running, without interrupting the ongoing interaction. This inclusion allows close supervision by a carer or therapist to steer the course of

treatment, for example by focussing on particular words, preventing repeat words, or skipping words which present particular difficulties.



Figure 6. Supervisor Device Application.

4.6 Facially Assisted Recognition

One of the major hurdles for people suffering from speech related ailments is often their inability to perform sufficiently distinct mouth shapes to pronounce a word correctly. Speech therapy exercises often require a patient to practice these mouth shapes based on paper images like those shown in Figure 7. The Microsoft Kinect skeletal and facial tracking features provides the ability to recognize the facial structure of someone who is in close proximity to the sensor and overlay a map of triangular shapes on the screen representing the major features of the face as shown in Figure 8. These features can then be automatically compared to the correct mouth shapes either with or without words being uttered.



Figure 7. Sample Speech and Language Therapy Exercises.



Figure 8. Facial Recognition Map.

This mapping is also used to assist the recognition engine in correctly determining broken speech. The mouth shape given by the facial recognition is analyzed and the average height and width of the set of triangles representing the lips measured. By comparing this data to the suggested recognition semantic elements, the application can determine if one of several basic mouth shapes is being employed, and infer if it is compatible with the suggested recognition. This provides an extra confidence metric both to ensure the correct operation of the interactive game, and as an additional logged metric.

4.7 Deployment

The client interactive application has been developed in C# using the Windows Presentation Foundation. It is intended for use with a Microsoft Kinect for Windows device, and as such makes use of the full C# API. The application is fully unthreaded, and capable of managing a multi Kinect aware environment.

The application exposes a Windows Communication Foundation RESTful SOAP service that allows a mobile device to connect and manage the ongoing session from within the same wireless or wired network. This service is intended for use through the developed Windows Phone application, but could be invoked by any Web capable device.

The application has been developed to run under Microsoft Windows 7 as this provides the widest base of hardware requirements possible for the system, to ensure low cost deployment into patients' homes.

5. TESTING

The prototype application has been tested against a partial list of possible phonemes, with full debugging information displayed on the screen and recorded into the test results manually. The setup for this test, at a standard desk with the development hardware can be seen in Figure 9.



Figures 9. Client Application Prototype Testing.

One key test undertaken was the system's ability to recognise spoken phonemes, using both recognition methods (audio and video). This test was undertaken using a set of the most common phonemes in the test word vocabulary. Each phoneme was repeated 10 times, and the output from both recognisers recorded. The correlation between the audio recognition, and whether or not the camera identified the mouth shape as being correct are shown in Figure 10. These results are currently being further analysed and the models adjusted.



Figure 10. Audio and video recognition accuracy.

The application was further tested to ensure that the therapist's device commands were successfully carried out by the game engine. This was setup using a private WiFi network at the development system, with the network

settings typically emulating those that would be found in most homes with modern broadband connections; the connection can be seen in Figure 11.



Figure 11. Testing of the supervisory interface.

6. CONCLUSIONS AND FURTHER WORK

Overall this project has succeeded in developing a targeted proof of concept application utilising modern consumer technology for assistance with the treatment of speech disorders in the home. This is extending the scope of the Kinect for speech rehabilitation purposes. The project prototype has also demonstrated the concept of remote therapy through being able to collect session metrics for real-time monitoring of the session and/or post session analysis. Whilst there is still much further development and testing to take place and adjustment of the system to further increase accuracy, the system has provides a framework on which subsequent developments can take place. These include:

- Remote Customization: development of a modular system that treats each therapy mode as a game level
 will provide the ability for therapists to remotely load new modes and activities. This will allow extended
 periods between one-to-one treatment visits in the home whilst still providing access to tailored therapy
 for patients.
- Visual & Audio Logging: the existing prototype makes use of the high quality voice recognition features of the Kinect. These, and the high quality camera could be used, with the patients consent, to take audio logs of their speech and visual logs of their mouth movement to assist therapists to accurately monitoring therapy sessions remotely. These logs could be stored and uploaded through the existing architecture, and even furnished with real time data overlays.
- Data Warehousing & Analysis: re-development of the systems metric database to use a data warehousing technology such as Microsoft CUBE (Microsoft c) would provide the ability to track trends and common patterns in patient therapy response data. This data could then be used to help future patients, and customize the interactive software to further directly address patient needs.
- Other Medical Targets: due to the modular nature of the application, and the use of the Kinect sensor, the prototype application could theoretically be used to provide treatment for a more diverse range of ailments including physically limiting problems. For example muscle damage which requires exercise sessions which could be monitored and analyzed for correctness by the Kinect skeleton tracker, and relayed to relevant medical staff, or personal trainers.
- Further Discussions, Development and Tests: following on from this initial proof of concept we are intending to interact further with speech and language therapists, to develop more games and to conduct further trials with regards to the accuracy of the system and its use as a therapy aid. We are also investigating how much more accurate the Kinect version 2 sensor is than the first generation sensor that was used for this development.

Acknowledgements: We would like to thank the speech therapists and linguistics experts for their valuable comments during the development of the system.

7. REFERENCES

- Bungalow Software, Speech and Language therapy software, [Online], Accessed July 1st 2014, Available http://www.bungalowsoftware.com/products_by_category.htm
- Deane, KH, Whurr, R, Playford, ED, Ben-Shlomo, Y, and Clarke, CE, (2001), A comparison of speech and language therapy techniques for dysarthria in Parkinson's disease, *Cochrane Database Syst Rev.* 2001 (2): CD002814.
- Glykas, M, and Chytas, P, (2004), Technology assisted speech and language therapy, Int. Journal of Medical Informatics **73**, 529–541
- Grawemeyer, B, Cox, R, and Lum, C, (2000), AUDIX: A knowledge-based system for speech-therapeutic auditory discrimination exercises. *Studies in Health Technology and Informatics*, **77**, 568–572.
- Kelly, H, Brady, MC, and Enderby, P, (2010), Speech and language therapy for aphasia following stroke (Review), The Cochrane Collaboration, *The Cochrane Library*, **7**, John Wiley & Sons.
- Kirschning, I, and Toledo, T, (2004), Language training for hearing impaired children with CSLU vocabulary tutor, *WSEAS Transactions on Information Science and Applications*, **1**(1), pp. 20-25.
- Lanz, C, Olgay, BS, Denzler, J and Gross, H-M, (2013), Automated classification of therapeutic face exercises using the Kinect, In Proceedings 8th Int. Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP 2013), Barcelona, Spain, pp. 556-565.
- Lee, J, Fowler, R, Rodney, D, Cherney, L, and Small, SL, (2010), IMITATE: An intensive computer-based treatment for aphasia based on action observation and imitation, *Aphasiology*, **24**(4), pp. 449-465
- Microsoft a, Kinect For Windows, [Online]. Accessed July 1st 2014, Available: http://www.microsoft.com/en-us/kinectforwindows/.
- Microsoft b, Kinect for Xbox, [Online] Accessed July 1st 2014, Available http://www.xbox.com/en-GB/Kinect
- Microsoft c, Introduction to Cubes, [Online] Accessed July 1st 2014, Available http://technet.microsoft.com/enus/library/aa216365(v=sql.80).aspx
- Mravak, S, Nadramija, D, Pulić, E, and Njegovan-Zvonarević, T, (2013), Autonomous patients training and monitoring system for stroke patients, Central Eurporean Conference on Information and Intelligent Systems, Varazdin, Croatia, September 18-20, 2013 pp. 188-192.
- NIDCD, National Institute on Deafness and Other Communication Disorders, [Online] Accessed July 1st 2014, Available http://www.nidcd.nih.gov/health/statistics/vsl/Pages/stats.aspx
- Nintendo, Nintendo Wii, [Online] Accessed July 1st 2014, Available http://www.nintendo.co.uk/Wii/Wii-94559.html
- Patient.co.uk, Speech Therapy An Introduction, [Online] Accessed July 1st 2014, Available http://www.patient.co.uk/doctor/Speech-Therapy-An-Introduction.htm
- Propeller, Speech Sounds on Cue, [Online], Accessed Juy 1st 2014, Available http://www.propeller.net/categorys/86.htm
- RCSLT, Royal College of Speech and Language Therapists, The All Party Parliamentary Group on Speech and Language Difficulties, [Online] Accessed July 1st 2014, Available http://www.rcslt.org/about/parliamentary_work/appg_sld_history
- Sadural, ST, and Carreon, MT, (2012), Appreciating speech through gaming, *In Proc.* 9th Intl Conf. Disability, Virtual Reality and Associated Technologies, Laval France, 10-12 Sept 2012, pp 141-148.
- Saz, O, Yin, S-C, Lleida, E, Rose, R, Vaquero, C, and Rodríguez, WR, (2009), Tools and technologies for computer-aided speech and language therapy, *Journal Speech Communication*, **51**(10), pp. 948-967.
- Schipor OA, (2014), Improving computer assisted speech therapy through speech based emotion recognition, In *Proc.* 10th International Scientific Conference eLearning and Software for Education, Bucharest, April 24-25, 2014.
- Schröder, C, Schupp, W, Seewald, B, and Haase, I, (2007), Computer-aided therapy in aphasia therapy: evaluation of assignment criteria, *Int. Journal of Rehabilitation Research*, **30** (4) pp. 289-295
- Sony, PlayStationMove, Motion Controller [Online] Accessed July 1st 2014, Available http://uk.playstation.com/psmove/
- Stroke Association (2011), Communication problems after stroke, Stroke Association, June 2011.
- Steps Consulting Limited, StepByStep, [Online]. Accessed July 1st 2014, Available: http://www.aphasia-software.com/Portals/0/assets/docs/StepByStep45-SoftwareGuide.pdf

- Umanski, D, Kogovšek, D, Ozbič, M, and Schiller, NO, (2010), Development of a voice-based rhythm game for training speech motor skills of children with speech disorders, *In Proc.* 8th Intl Conf. Disability, Virtual Reality and Associated Technologies, Vina del Mar/Valpariso, Chile, 31 Aug- 2 Sept, 2010, pp 255-262.
- University of Sheffield, SWORD, Computerised therapy for acquired apraxia of speech, [Online], Accessed 1st July 2014, Available

http://shg.sheffield.ac.uk/casestudies/computerisedtherapyforacquiredapraxiaofspeec.aspx.

Virtual Speech Centre, Speech Therapy Mobile Apps Development, [Online], Accessed July 1st 2014, Available http://www.virtualspeechcenter.com/MobileApps.aspx