Evaluation of the prototype mobile phone app Pugh: a 3D cartoon character designed to help deaf children to speech read

D M Shaw¹, M Patera¹, E Paparidou¹, R Wolff²

¹School of Computing, Science and Engineering, University of Salford, MediaCityUK, Salford Quays, Greater Manchester, UK
²German Aerospace Center, Simulation and Software Technology, Lilienthalplatz 7, Brunswick, GERMANY

d.m.shaw@salford.ac.uk, m.patera@salford.ac.uk, xelen.pap@gmail.com, robin.wolff@dlr.de

¹www.salford.ac.uk, ²www.dlr.de/sc

ABSTRACT

Pugh, a 3D cartoon character, is a prototype smartphone application developed at the University of Salford. Its purpose is to provide speech-reading exercises for hard of hearing and deaf children. This paper discusses the design of the application, the test process and acknowledges that the technological limitations of the platform and the character’s non-human characteristics provide some interesting challenges. A preliminary test was conducted to evaluate speech perception and lipreading from Pugh. The findings proved that Pugh is not an accurate speaker. Further development of the lip movement and facial expressions is required in order to achieve accuracy.

1. INTRODUCTION

Pugh, a 3D cartoon character, features in a mobile phone application for iPhone/iPod/iPad that aims to provide speech-reading exercises for hard of hearing and deaf children in the United Kingdom. The application is a prototype that we developed at the University of Salford, UK, with the view to expand the ways in which technology can be used to facilitate the learning process of speech perception. At the beginning of 2012 there were more than 570,000 applications available for Apple products alone, therefore the idea to develop this prototype educational software as an iPhone app seemed appropriate. This type of platform, which provides visual, auditory and vibro-tactile sensory feedback, offers the opportunity to develop educational material that could be inclusive to people with disabilities and reach a wider audience.

The application prompts users to type in text and Pugh will speak the words, with sound or without (Fig. 1). The speed at which he speaks can be adjusted to facilitate lip reading. The users can also change the perspective view. Our main research questions are whether the current prototype of Pugh could effectively imitate the qualities of a human speaker and whether the application could be used in speech-reading exercises? Due to its existing non-human design Pugh can only replicate the most prominent lip movements. Phoneticians have classified the features of speech into three main categories, depending upon the emphasis that is given to the event (Harley, 2000):

1. The articulatory phonetics, which study how speech is produced
2. The auditory/perceptual phonetics, which study how sound is perceived and
3. The acoustic phonetics, which focus on acoustic and physical characteristics of sound.

Although all three categories are important in order to provide an accurate evaluation of Pugh, this study gives more emphasis on the articulatory phonetics of the speaker (Pugh), and the auditory/perceptual phonetics of the listener (the user of the application). Even though the present programming of Pugh has some limitations, as the iPod touch graphics did not allow for the addition of enough skeleton bones in the 3D geometry mesh, it seemed crucial at this stage of the development to examine and to identify the accuracies and inaccuracies of the avatar’s lip movements.

This paper describes the development of the application and the design of the speech-reading test, it presents and discusses the findings of the preliminary evaluation and proposes a plan for future work.
Figure 1. A screenshot of Pugh as he appears in the iPhone application.

2. THE CONCEPT OF PUGH’S DESIGN

Many experiments and investigations have been conducted to evaluate the effectiveness of audiovisual techniques for developing deaf children’s speech-reading ability. Computer-based instruction is emerging as a prevalent method to train and develop vocabulary knowledge for both native and second-language learners (Druin & Hendler, 2000) and individuals with special needs (Adamo-Villani & Wright, 2007; Adamo-Villani, 2006; Barker, 2003). Withrow (1965) supports that the use of 2D motion pictures can extend the speech-reading and auditory stimuli of a deaf child and reinforce the ability to produce speech. Saleh (1965) has proved that the use of a 2D image and 2D motion pictures can increase up to 50% the comprehension of the environment and the speech perception of a child. Although many experiments have proven that the proper use of 2D images and animation can enhance the speech-reading ability of deaf children, the use of the 3D visuals has not yet been thoroughly investigated. Massaro (2006; 2003) by creating Baldi, a realistic 3D human head, proved that a 3D avatar can be used for speech perception. Baldi is a computer animated tutor that uses accurate visual and audible speech for language and vocabulary learning (Animated Speech Corporation, 2006) which has also been released as an iphone application (Psyentific Mind Inc. 2012). The Baldi head has a human structure therefore the mapping of accurate facial and lip movements was easier to achieve using motion capture data from a real person. Although Baldi has been developed with all ages in mind, a realistic floating head might not be very appealing to young children. The rationale for developing a cartoon character such as Pugh was to cater for this younger audience.

Pugh is a cartoon type fictional character, designed to appeal to children between the ages of 4-10 years old. Pugh’s face and body are not human, which presents further complications and difficulties in making it an accurate speaker. The challenge of this study is to develop a non-human character, which could be as accurate a speaker as Baldi. The idea of children engaging and interacting with a ‘friendly’ character is a concept that Pugh has been modelled on. Therefore the creation of a believable character was important for this application to engage with children. Pugh has been designed with eyes on storks so that he can add expression to the words that he is saying. Believability does not require human form and this is suggested by characters such as Bambi. It is shown even more strongly by the character of the Flying Carpet in the Disney animated film Aladdin (1992). Moreover, the application allows the user to interact with Pugh which adds an important aspect to the learning process. Goldin-Meadow and Singer (2003) discuss how deaf children learn more by using their fingers than their mouths and ears. This notion of the deaf learning through touch inspired the interactions programmed in this application. The mobile phone and tablet platform was chosen because it provides this type of interactions. Applications that feature talking cartoon animals have been developed by Out Fit 7 Ltd. (2012; 2010) but these have been mainly created for entertainment and not for educational purposes hence the characters are not accurate speakers yet they are engaging and fun to interact with. The objective of this project is to create a character that combines both.
3. THE DEVELOPMENT OF THE APPLICATION

3.1 The 3D Model of Pugh

A polygonal 3D geometry mesh was built using Autodesk Maya based on an initial clay model of Pugh. Care was taken to keep the number of polygons as low as possible, as the application’s target platform only offered low graphics performance. The mouth area was designed to have a higher density of vertices in order to reduce artefacts when animating the lips. A skeleton rig was created for the lips area while a basic skeleton was added for the arms, fingers and legs. The lips had bones in the corners, four at the top and four at the bottom. We additionally placed bones for the cheeks, chin and under the nose for reference and to animate facial expressions (Fig. 2).

![Figure 2. Screenshots of Pugh’s 3D model, shaded and wireframe.](image)

3.2 Motion Tracking and Animation

To capture the lip movements of a human speaker we utilised an optical marker-based motion tracking system by Vicon with six cameras (Fig. 3). The markers were placed at the same positions as the bone control points in the skeleton rig of the 3D model. Each marker was a reflective sphere with 5mm radius. With this marker size, the tracking system offered an accuracy of approximately 2-5mm. We captured several sequences of reading out a list of words, each emphasising one of the 44 phonemes of the English phonetic alphabet chart. Additionally, we captured video footage as reference material for evaluating the mouth movement of the 3D model and for animating the tongue.

When the prototype application was developed, the iPhone 3GS had just been introduced but the iPhone 2 was still widely spread; hence a requirement was that the app would run on devices starting from iPhone 2. This generation, however, only supported the OpenGL ES 1.1 graphics standard, allowing a maximum number of nine bones per mesh. Including the mouth, eye brows, arms, legs and fingers, Pugh’s skeleton had more than 100 bones. Although the mesh could have been separated into smaller patches, each controlled by nine bones maximum, this would have created unrealistic hard edges at the seams of the mesh patches when animating the mesh. As a consequence, we decided not to implement a skeleton based animation so we used the traditional morphing method instead, where the geometry is animated by interpolating vertex positions between meshes that represent certain poses of Pugh while articulating. An advantage of this method is that the poses could be tuned on a finer granularity than what would have been possible with the skeleton method. For example, skin and wrinkles could be sculpted in more detail and independently, where neighbouring bones might have influenced each other. The disadvantage is that subtle movements from motion tracking could get lost due to the linear interpolation between selected poses. Hence, a large number of intermediate poses were necessary to guarantee enough detail, which in turn increased memory consumption on the mobile device.
Figure 3. Arrangement of reflective markers on a subject.

For the prototype application, we created poses for nine visemes (visual phonemes). These included /a/, /e/, /i/, /o/, /m/, /sh/ and /l/. All occurring phonemes were mapped to one of these visemes based on similarity, for example phonemes ‘ae’, ‘er’ and ‘hh’ would use the viseme /a/, or phonemes ‘m’, ‘b’ and ‘p’ would use the viseme /m/. We also created a few additional poses used in animation sequences when Pugh is idle and not speaking in order to make him more life-like and show random movements, such as looking around, yawning or scratching himself. Our system allowed blending of several morph targets within one animation sequence so that ‘speaking’ poses could be mixed with idle poses to achieve a natural impression.

3.3 Simulation Environment

The simulation environment of the application was implemented using the Oolong Engine. Oolong bundles a set of C++ libraries helping to develop and port games to iOS based mobile devices. It provides a run-time framework with access to the graphics system, math functions, file format utilities and touch screen. On initialisation, the application loads the geometry and texture files for the displayed scene. It then allocates the required memory and prepares the morph targets by registering a number of vertex buffer objects. The simulation loop continuously reads user input and triggers the respective actions. It also advances the simulation time, switches between poses and controls the blending weight for animation.

The speech audio output was generated using the Flite library, a run-time speech synthesis engine developed at Carnegie Mellon University (Black & Lenzo, 2009). It analyses a sentence and creates a set of keys representing the phonemes in the words of the sentence. For example, the word ‘hello’ produces the phoneme keys ‘pau’, ‘hh’, ‘ah’, ‘l’, ‘ow’ and ‘pau’. Based on this set of phonemes, an audio file is synthesised and stored on the file system. This audio file is then played back when Pugh speaks and audio output is enabled by the user. The set of phonemes is also used to select the respective visemes for lip animation. Synchronisation of the audio output and lip animation was achieved by controlling the blending time between visemes to match the individual phoneme timings extracted from additional phoneme information provided by Flite.

The user interface was implemented using the standard tools of Xcode for iOS development. The application shows a splash screen when loading and provides a separate help page with instructions. The main screen shows the 3D graphics window in full-screen with Pugh centred standing on a planet-like surface. A text bubble above Pugh and control buttons at the right and bottom of the screen are displayed as overlay icons, allowing the user to enter text, show/hide the text bubble, replay speech output, and to adjust the speed of the animation. The perspective and zoom can be controlled via the typical two-finger gestures. Tapping on him would trigger one of three short animation and sound sequences (e.g. a response to a tickle).

4. PRELIMINARY EVALUATION AND FINDINGS

4.1 Design of the Speech-reading Test

Campbell and Mohammed (2010) survey of 52 lipreading tests, as well as other tests that examine the lipreading skill from a clinical and neuropsychological perspective (Mohammed et al. 2006) have been consulted in order to design a subjective experiment for the preliminary evaluation of the application. Taking into account Pugh’s initial limitations it was deemed rational to design a short test that would examine the lip accuracies and inaccuracies only for the easy consonants and vowels. The test followed some basic
guidelines as proposed by Markides (1980), thus it consisted of five visible consonants (p, b, m, f, v) and five vowels (a, e, i, o, b) in VCV, CVC, CV, VC words, 15 high frequency words (primary school level) and three sentences.

4.2 Participants

Originally the test was designed to be performed by deaf pupils of secondary or further education UK schools which teach oral communication but because Pugh is a prototype and hence unfamiliar to many users, it was considered important to use adult experienced speech readers for this initial test. Based on Bernstein et al. (1998) work, deaf are considered the best speech readers. To create an accurate test the participants had to present a homogeneity in their characteristics: a) They had to be over the age of 12 years old in order to have a cognitive knowledge of the vocabulary and some years of experience with speechreading b) To have some level of hearing loss (profound or severe) in order for sound to be tested as well c) Oral communication to be their primary means of communication d) English to be their mother tongue or English to be their dominant language e) know how to speechread or be relevant to purpose of the test (phoneticians, speech therapists, linguists, etc.

An online test was created in order to reach more participants. This contained 37 videos captured from an iPhone’s screen showing the character speaking the predefined words. From the six participants who performed the test two were deaf. Their responses are discussed below.

4.2 Findings

The secondary research on phonetics, speech-reading and existing learning resources for hearing impaired children, in conjunction with the findings of the preliminary subjective test, demonstrated that, at the moment, Pugh is not an accurate speaker. Due to the low number of participants, a statistical analysis could not be achieved by combing all the above information and by analysing the results of the tests, there is strong evidence that speechreading from the existing version of Pugh is very difficult. We applied the recommendations as published by Bernstein et al. (2010), Dodd (1980), Markides (1980), Jeffers and Barley (1980) for scoring and analysing the responses. For the first three parts of the test (VCV, CVC, CV, VC) a sufficient proportion of the speechreading of the consonants was correct. The responses giving a consonant of the same viseme class cannot be considered as a mistake. The inaccuracies caused by the limitations in the technology to animate Pugh’s lips more accurately, are also confirmed by the fourth part of the test (vowels) as seen in Table 1. Pugh cannot articulate the vowel ‘o’; thus the inaccurate responses of the participants cannot be considered as a mistake of speech perception. The consonant ‘p’ was chosen to be the stable variance, because Pugh can articulate this phoneme accurately. The vowel ‘a’ was perceived correctly by all the participants.

Table 1. Participants’ responses to Part 4 of the online test (vowels)

<table>
<thead>
<tr>
<th>Part 4</th>
<th>Syllables</th>
<th>Subject 1</th>
<th>Subject 2</th>
<th>Subject 3</th>
<th>Subject 4</th>
<th>Subject 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowels</td>
<td>CVC syllables C = P</td>
<td>Normal hearing</td>
<td>Normal hearing</td>
<td>Normal hearing</td>
<td>Deaf/Severe level of hearing loss</td>
<td>Deaf/Severe level of hearing loss</td>
</tr>
<tr>
<td>The participants had to lip-read the Vowel in the middle of the CVC syllable.</td>
<td>PAP</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>PEP</td>
<td>E</td>
<td>I</td>
<td>A</td>
<td>I</td>
<td>E</td>
</tr>
<tr>
<td></td>
<td>POP</td>
<td>A</td>
<td>A</td>
<td>I</td>
<td>E</td>
<td>O</td>
</tr>
<tr>
<td></td>
<td>PIP</td>
<td>A</td>
<td>I</td>
<td>A</td>
<td>U</td>
<td>I</td>
</tr>
</tbody>
</table>

Furthermore, the incorrect responses for the fifth part and the sixth part of the test (words and sentences), confirm that the animation of the lips and tongue require further development. However, there was a great proportion correct of the bilabial phonemes and the labiodental phoneme ‘f’, which was perceived correctly by all the participants. The word ‘baby’ was perceived correct by all of the participants. Researchers have stressed speechreading perception could be more accurate when the observer is aware of the theme. The fact that words were presented without any indication has caused confusion to the speech perception. It would have been more suitable if some indication was provided for each word. This could be achieved by
presenting all the words, and the sentences, inside a story (e.g. "Mary and John got married and had a baby"). Another way of offering a hint about the words is by using categories (e.g. list of vegetables, list of colours, list of numbers and so on).

This investigation helped us to understand and identify the weaknesses of the animated character and how we could make Pugh an accurate speaker by improving its lip, facial and body gestures. The test gave us a clear indication of which vowels, consonants and words Pugh can or cannot articulate well. Also, the avatar’s idle position seemed to confuse the users as Pugh is closing his mouth too quickly after completing a word, which visually adds a bilabial closure at the end, causing the assumption that some words ended with m, b or p. In addition, the facial characteristics of Pugh, the eyes, eyebrows, cheeks, which are not programmed to match human expressions, result in an expressionless face which requires the observer to use only the lip movements for speech perception. Additionally, at the moment Pugh cannot include the suprasegmental characteristics of language such as intonation and stress, therefore it cannot for example pose a question. As a result, a great amount of information is lost as the observer obtains many data while reading from the whole face (and torso).

Even though the online test was designed based on previous studies and proposed guidelines we were not able to recruit a satisfactory number of deaf participants to increase the validity of the test. Furthermore, in order to reach more participants the test was distributed online and as a result the homogeneity of the group could not be ensured. Another limitation of the test design was that sound was not tested. The absence of sound can decrease the performance of the deaf participants; studies have shown that an audiovisual stimulus obtains greater amount of success in speech perception than audio or visual stimuli alone. Moreover, the videos of Pugh speaking were captured from the front angle thus the participants could not change the perspective (e.g. side view or ¾ angle) to gain a better view of the lip movements. All the participants commented that Pugh should have more ‘human like’ characteristics. They found it slightly misleading that the character had an expressionless face and that his eyes, eyebrows and hands did not move properly. They would also prefer if Pugh could provide some information using finger-spelling; this feature is discussed below in our future work section. From the lessons learned through this preliminary evaluation we have acknowledged the test procedure’s limitations and what it needs to be done to improve it to evaluate the next versions of this application.

5. CONCLUSIONS AND FUTURE WORK

The technological limitations and the character’s non-human facial characteristics provide some interesting challenges for the future development of this application. The results have reinforced a need for further development of the lip movement and the facial expressions. Pugh has the potential of becoming an educational, entertaining and engaging application for deaf children but in order to achieve this, the contribution and participation of specialist speech therapists in deafness, audiologists, clinical psychologists, teachers of oral communication for deaf or hard of hearing, throughout the development stages is very essential for the success of this application. We have also approached the producers of a children's network channel (CITV) to discuss the potential employment of Pugh, once accurate, as a signing character for children’s programmes on the network.

This investigation and secondary research on phonetics highlighted the multidimensionality and complexity of speech perception, and the importance of not only accurate lip/teeth/tongue movements but also facial expressions, torso movements and hand/fingers gestures. It is evident that the cartoon character’s design has to be reviewed and altered to accommodate these features and animations. As technology continuously evolves, Pugh’s accuracy in lip movement could be improved by creating a more detailed skeleton rig to achieve an emulation of a human mouth’s motions. The tongue, for example, could be mapped more accurately to enhance lip speaking. Part of our future work plan is the addition of cued speech. Cued speech will create a more believable character that has gesticulations that coincide with lip speaking. Deaf children combine strings of gestures to convey ideas, or propositions, similar to the way in which speakers combine strings of words to make statements. Speakers tend to produce only a single iconic gesture for each major idea they say (Mayberry & Nicolaidis, 2000). Another valuable component is the potential to present multiple sources of information, such as text, sound and images in parallel (Chun & Plass, 1996). The intent is to maintain the cartoonish look of the character and introduce hand gestures, expressions and accurate lip speaking.
6. REFERENCES


