Pilot Experiments on Children's Voice Recording

Sawit Kasuriya and Alistair D N Edwards
Human-Computer Interaction Group,
Department of Computer Science, University of York,
Heslington, York, UK, YO10 5DD
Tel: +44 (0)1904 432775, Fax: +44 (0)1904 432767
sawitk@cs.york.ac.uk and alistair@cs.york.ac.uk

ABSTRACT
Automatic speech recognition is being used increasingly in a variety of applications. There is great potential for its use in educational applications for children. However, the accuracy of recognition of child speech is very low. There are probably a number of reasons for this, but one is the difficulty in collecting high-quality recordings of children to be used in the building of speech models. If a better interface can be provided between the child and the recording equipment then it may be possible to collect better samples. Interfaces have been designed to be tested to that end, using alternative interface paradigms: push-to-talk and a limited time recording with and without a progress bar. These alternatives will be compared by collecting speech samples and measuring their quality.

Categories and Subject Descriptors
D.2.2 [**Software Engineering**]: Design Tools and Techniques – *user interfaces*; H.1.2 [**Information Systems**]: User/Machine Systems – *human factors*

General Terms
Human Factors, Experimentation.

Keywords
Voice recording, speech interface for children.

1. INTRODUCTION
The Speech recognition technology has been investigated for decades and is used in an increasing number of applications. Most of them, however, have been developed for adult users. Speech recognition also has great potential for children's applications in education, and entertainment. A speech-input program avoids all of the problems inherent in using the conventional keyboard and mouse. Children may find it motivating to be able to talk to a responsive computer. Most importantly for our purposes, a speech interface can be used as a tool to investigate children's language skills at a pre-literate age.

ICMI-MLMI'09 Workshop on Child, Computer and Interaction, November 5, 2009, Cambridge, MA, USA.

Regardless of all the attractions of speech recognition, there is a problem in that speech recognition rates are very low. For instance in [1] and [2] the best accuracy attained is of the order of 77.30%. There are many reasons for this, related to the need to build a model of speech to be used by the recognition engine [3, 4]. That is to say that before a speech recognizer can be built, a large set of samples of speech must be collected as the basis of a speech model. Problems in collecting samples and building the model include the fact that children's speech organs are in a constant state of development, so that there is little stability, and hence the models (and subsequent recognition engines) are very age-dependent. Collecting the samples is also difficult because it is necessary to get people (children in this case) to speak prescribed words precisely and on cue. Samples which are incomplete, include noise or extraneous words will all degrade the accuracy of the recognition engine based on the model.

The objective of the work described here is to investigate whether by making a more accessible interface between the child and the recorder, it is possible to collect better samples. Such an interface will also have application in any software in which children's speech is recorded or recognized.

This paper represents work in progress. The design of three interfaces is presented. The pilot experiments of those interfaces are implemented and evaluated on five young children. This work is part of a larger project which has the objective of developing speech-based tools for screening for dyslexia in pre-literate children.

2. RECORDING METHODS
Recording technology has gone through a number of generations in recent years, but even modern devices have basic controls reminiscent of mechanical tape recorders: *play, record, fast-forward,* etc. These controls are simple and familiar to adults, but they may be over-complex for children. In an educational context, the child needs to be able to concentrate on the educational task and not the controls. The basic requirement is for the child to speak while the software is in record mode – with a minimum of intervention.

In this our experiments, two approaches are used to investigate which one is the more suitable: *push-to-talk* and *limited time recording*.

2.1 Push-To-Talk
This method resembles that used on walkie-talkies and similar devices, whereby the speaker presses a button and holds it for the duration of their utterance. The recorder will be activated to start recording when the button is pressed and recording will finish

when the button is released. In a software interface the child will press and release a button on the screen by way of the mouse button.

Push-to-talk leaves the child in control – which may be both an advantage and a problem. The child needs to speak only when they are ready, when they have mentally prepared a response and releasing the button should be a clear signal that they have finished speaking. A possible disadvantage is that the child may hold the button for as long as they like, and may do so for much longer than necessary.

2.2 Limited Time Recording

As children might have no experience with voice recording, it may be better to give control to the recorder to automatically start and stop itself. In this approach, the child is prompted to give a response within a prescribed time during which the device is in record mode. There are two variations on this approach. In both the child is prompted to say the required word and the recorder remains in record mode for long enough to capture the speech. In one version the child is shown a progress bar, whereby they must have finished speaking before the bar reaches its right-hand extremity (see Figure 1). In the other version, no indication of time is given. By testing both versions it will be possible to see whether cleaner, more complete recordings are obtained with the progress bar, or whether this puts the child under undue pressure. For this experiment, we set a time for recording at around 1.9 seconds since it should be enough for a three-syllable word.

Figure1. Sample of a progress bar

3. EXPERIMENTAL TASKS

Our work is aimed at children aged 5 to 9, in the first stages of formal education, the stage known as Key Stage 1 in the UK national curriculum [5]. At that age, pupils should be able to read digits, letters and simple words. The tasks, therefore, will be to recognize and speak words of three classes: the digits (0 to 9), the letters (A to Z) and 44 simple words selected from the isolated word lists in [6]. This gives a total of 98 words. It would be unrealistic to expect a child to read all of them in one session, so we divided all those words into four subsets of 3 digits, 7 letters and 11 simple words, making 21 words in each session. We expect each child should spend less than 10 minutes in each session. However, we will also need to collect data for each of the three recording methods.

3.1 Participants

Five young children (g01-g05) participated in this pilot experiment. All of them are girls and their ages are shown in Table 1. They all have been familiar and experienced with using computer at their home and school.

Table 1. Participants' ages (years)

Participant	Age
g01	5
g02	8
g03	8
g04	6
g05	7

3.2 Experimental Procedures

The children in this study were in an early stage of literacy, during which children may not be able to read some written words and therefore pictures were used as prompts instead of written words. Some words, however, are difficult to represent unambiguously in a picture, such as, 'sea', 'wall' and 'road'. Children may struggle to name those pictures. Therefore, the children were encouraged to try to guess and say any word prompted by the picture.

The experiment was divided into three phases: *pre-learning*, *learning* and *post-learning*, as shown in Figure 2. Before starting these three phases, the child was briefed as to what response was expected, how they could record the voices during experiments, how to use buttons in the software and also the steps of experiments. For limited time recording, the child was particularly encouraged to utter a word as soon as he/she perceived the picture. By contrast, in the push-to-talk interface, the child was free to take time thinking about a picture before pushing a button to make the recording. This also implied that the child was permitted to be silent if he/she had no idea about the picture and could not make any guess at the word.

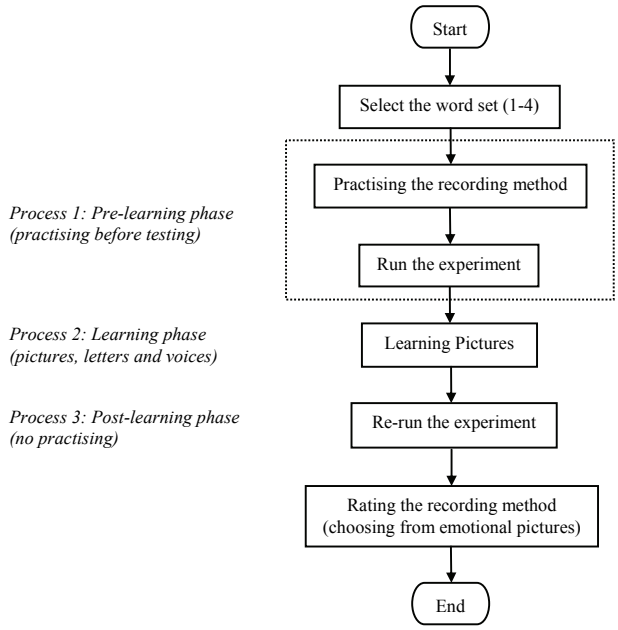

Figure2. Experimental procedure in each recording method

As shown in Figure 2, each experimental session starts with practice. The practice set consists of eight pictures representing two digits, two letters and four words. This process helps children to feel comfortable and relax before doing the experiment. Children are allowed to practise the recording method until they fully understand the interface. The next process, the *pre-learning phase*, starts immediately after the practice. Children have to name twenty-one pictures in this experiment. As the primary objective of this study is not testing literacy skills of children, we then provide the learning program to help children to learn and memorise all the pictures used in experiments. This step is the *learning phase*. The child will learn particular pictures, which they have already seen in the pre-learning phase, through displaying the written word and playing its pronunciation. After finishing this learning phase, it is hoped that the child will not have any difficulties naming particular pictures in the post-learning phase since the same pictures are used in both the learning and experimental phases.

At the end of experimental procedure, the child is asked to rate their feeling with the recording interface by choosing one of three emotional 'smiley' pictures (happy, neutral and sad) [7].

Figure 3. 'Smiley' faces used to represent *happy*, *neutral* and *sad* respectively to the child participants

All three experimental interfaces were tested following the procedure shown in Figure 2. The order of presentation of the interfaces was the same for each child:

- limited time recording without a progress bar (M1),
- limited time recording with a progress bar (M2) and
- push-to-talk (M3).

4. RESULTS AND MEASUREMENTS

To compare the recording approaches in our experiments, we need some measurements to indicate the speech quality, children's satisfaction and compatibility with a speech recognition system.

The time taken to complete each study will be an indication of the efficiency of the interface. The times of each of the recordings can also be compared with the minimum time for each utterance, giving an indication of the quality (i.e. how much extraneous recordings have been made). However, the most important factor of voice recording is its quality. Therefore we will be mainly concerned about measuring the recording quality of each interface and therefore time measurement is not presented in this paper.

There are many aspects to consider the results of these experiments. In this paper, however, we concentrate on four topics:

- learning process,
- recording methods,
- recording quality and
- children's satisfaction

Firstly, we compare the learning results between pre- and post-learning. Secondly, the three proposed recording approaches are compared by counting the number of good recording utterances. Analysing the errors of recording and its classification will be described later. Lastly children's satisfaction of recording interfaces will be discussed.

4.1 Results of Learning Process

As the participants of this study were in pre-literate age, their knowledge is limited with large individual differences. Some children have difficulties in naming a picture rather than having problems with the recording interface despite the pre-learning phase. This phase was designed to help children to remember pictures and to teach them how to name the pictures. We expected that after the pre-learning phase, the child will be able to name all the pictures. Nevertheless, some children still did not remember picture or know what word to say. Two measurements were proposed to evaluate the self-learning process, classified as *wrong word* (WW) and *missed word* (MW). Wrong word implies a child named a picture with a word other than that expected, while missed word means nothing was recorded in time since the child kept silent. The results of self-learning process are shown in Table 2.

Table 2. The number of wrong words (WW) and missed words (MW) in pre- and post-learning phases

ID	Set	Method	Pre-learning		Post-learning		Improvement	
			WW	MW	WW	MW	Total	%
g01	1	M1	0	4	0	4	0	0
	2	M2	1	4	0	0	5	100
	3	M3	1	0	0	1	1	100
g02	2	M1	0	1	0	0	1	100
	3	M2	4	0	0	0	4	100
	4	M3	4	0	0	0	4	100
g03	2	M1	0	1	0	1	0	0
	3	M2	3	0	1	0	3	100
	1	M3	1	0	0	0	1	100
g04	1	M1	2	12	3	1	12	85.71
	2	M2	2	5	1	1	5	71.43
	3	M3	0	1	1	0	1	100
g05	2	M1	4	1	0	0	5	100
	3	M2	3	0	1	0	2	66.67
	4	M3	4	0	1	0	3	75.00
Total			29	29	8	8	47	81.03

Comparing the results of pre- and post-learning in Table 2, it is clear that the number of wrong words (WW) and missed words

(MW) decreased after self-learning phase. This implies that children have learned and remembered some of the words. However in some cases children seemed to forget the pictures as they scored 0% in the post-learning phase. This occurred in the limited time recording without a progress bar (M1) condition. This was also the first condition presented to the children and so may be because children usually felt nervous when doing experiment at the first round. This requires further investigation.

4.2 Comparison between recording methods

To evaluate and compare the results of each recording method, we will need to consider the acceptable recorded utterances as the key measurement. It is because the good recording always gives better accuracy of speech recognition. Therefore the method which provides more acceptable-quality utterances, will be most suitable for application to recognition.

Table 3. The percentage of good recording in pre- and post-learning process during experiment

ID	Method	Pre-learning	Post-learning	Improving
g01	M1	52.38	71.43	19.05
	M2	66.67	85.71	19.05
	M3	47.62	71.43	23.81
g02	M1	80.95	95.24	14.29
	M2	95.24	100.00	4.76
	M3	38.10	23.81	-14.29
g03	M1	85.71	95.24	9.52
	M2	90.48	100.00	9.52
	M3	85.71	100.00	14.29
g04	M1	33.33	80.95	47.62
	M2	76.19	95.24	19.05
	M3	0.00	4.76	4.76
g05	M1	95.24	95.24	0.00
	M2	95.24	90.48	-4.76
	M3	9.52	66.67	57.14
Mean	M1	69.52	91.67	**18.10**
	M2	**84.76**	**94.29**	9.52
	M3	36.19	53.33	17.14

In this pilot phase a subjective measure of speech quality has been applied. The comparison of recording quality for each method is shown in Table 3. Limited time recording with a progress bar (M2) was the most successful approach with an acceptance rate of nearly 85% for pre-learning and almost 95% for post-learning. On the other hand, push-to-talk (M3) approach was the worst interface with an acceptance rate, post-learning of about 50%. Almost all participants performed poorly when using push-to-talk method. For example, g04 and g05 did not succeed in using push-to-talk to record their voice at all during the pre-learning process. However, the number of acceptable utterances in limited time recording without a progress bar (M1) and push-to-talk (M3) was dramatically increased in the post-learning phase in all cases except g02.

One would expect that the results in the post-learning phase would be an improvement on the pre-learning phase. However, this was not the case in the instance of push-to-talk (M3) for g02. The number of acceptable utterances was markedly decreased in the post-learning phase. This appeared to be because she misunderstood the operation of the push-to-talk button, and rather treated it as a click-and-go button. In other words, instead or pressing the button, holding it while speaking and then releasing it, she pressed and released and then spoke.

4.3 Recording Qualities

There are many ways to measure the quality of recording. This experiment was arranged to prioritise an objective measure, encompassing the quality of the speech signal in terms of its recorded amplitude, its completeness (i.e. is any of the utterance clipped off?) and its accuracy (did the child say the required word?). Therefore we formulated seven categories to classify the quality of recording, which are acceptable quality, no recording and five errors of recording. Since children have a limited time to record, their utterances are easily clipped off. The recording errors are divided into five categories depending on the missing area of a recorded utterance.

(1) Category 1 (C1): Missing nearly the whole utterance. Only a short section of the beginning of the utterance is recorded. The missing recording in this category is classified as the worst recording.
(2) Category 2 (C2): Missing the beginning of the utterance. This rarely happens in limited time recording as the recorder will automatically start after the button is pressed. In practice it only occurs in an interface that allows children to have full control.
(3) Category 3 (C3): Missing almost half of the utterance. This missing error usually occurred in the limited time recording when children spend most of the time thinking about a picture.
(4) Category 4 (C4): Missing the end of the utterance. The difference between this missing error and category C3 is only a small part at the end of the utterance is clipped. Generally, the utterance in this category is almost complete but there is no silence at the end of the recording. It seems that recording might not be finished. In some cases this error may be acceptable as the recorded voice is understandable. However, we still do not know how this error would affect the accuracy of speech recognition.
(5) Category 5 (C5): A combination of C2 and C4. For the push-to-talk method, children may press the button before speaking and release it before finish speaking. As its result, both beginning and end part of the utterance will be clipped off.

As seen in the previous section the results of the post-learning phase were better than in the pre-learning phase, so only errors of recording in post-learning will be considered in this section. All qualities of recorded utterances of each recording method collected from five children are summarised in Table 4.

Table 4. Summary of recording errors in post-learning

Method	Acceptable		Category of recording error						
	n	%	None	C1	C2	C3	C4	C5	Total
M1	92	88	6	1	0	4	2	0	13
M2	99	94	1	0	0	2	3	0	6
M3	56	53	5	9	6	1	21	7	49
Total	247	78	12	10	6	7	26	7	68

For the limited time recording method without a progress bar, nearly 90% of utterances were classed as good recordings. With a progress bar the figure is nearly 95%. Approximately one half of recorded utterances were good quality when using the push-to-talk recording method. Obviously, the results of push-to-talk were the worst recording method in this experiment. Missing the end of utterances (C4), was the most frequent error in push-to-talk. This type of error was near to 50% of all errors in push-to-talk recording.

4.4 Children's Satisfaction

Not only was quantitative data analysed in this study, but the children's subjective satisfaction was also collected. A simple question was asked to assess each interface using the 'smiley' symbols in Figure 3. The results of the children's subjective satisfaction are shown in Table 5. Some children felt happy with more than one interface. In these cases a supplementary question was asked as to which interface they liked *best*. Their choices are marked with another 'happy-smiley' face in Table 5. These results are scored, whereby a happy smiley scores 1, neutral 0 and unhappy -1 (Figure 2), with a bonus +1 for the favourite.

Table 5. Children's subjective satisfaction

Participant	Recording interface		
	M1	M2	M3
g01	☺☺	☺	☹
g02	☺	😐	😐
g03	😐	😐	☺
g04	☺	☺☺	☺
g05	☺	😐	☺☺
Score	5	3	3

From the results in Table 5, it is apparent that most of children preferred limited time recording without a progress bar (M1). Only one child scored one of the interfaces negatively, and that was for the push-to-talk method (M3), but there was no difference overall between M2 and M3; evidently feelings were mixed regarding M3.

5. CONCLUSION

Speech input technology has great potential for use with children. However, there are problems in getting children to interact with this technology which are not the case for adults. In this study we made one attempt to see whether the child-computer interface can be designed in such a way as to overcome some of these unusual barriers. Three recording methods were proposed to experiment on young children in order to study children's behaviour and to analyse the quality of recordings.

The qualities of recorded utterances were clearly improved after the children had some practice. Push-to-talk (M3) was the worst interface in terms of errors in recording, when children have full control during recording. Limited time recording with a progress bar (M1) showed the best results, while limited time recording without a progress bar (M2) was subjectively the favourite.

There are a number of limitations to this study. First and most obvious is the small number of participants. Children are also likely to be more influenced by novelty and attention. Thus, as noted above, there may have been an order effect which influenced their results in the tests – both in the quality of recordings (i.e. that more MW and WW errors occurred in the first, M1 condition) and in the subjective ratings (that is, on the first presentation, the child had nothing to compare with).

Nevertheless, this pilot study has given us a starting point from which we can carry out a larger-scale and more rigorous investigation. In the long term we expect this to contribute to both better speech models for recognition of children's speech and for the application of speech recognition in educational applications.

6. ACKNOWLEDGMENTS

We would like to thank free clipart resources in following URL: http://office.microsoft.com/en-gb/clipart/default.aspx and http://www.clipart4schools.com.

7. REFERENCES

[1] Giuliani, D., and Gerosa M. 2003. Investigating recognition of children's speech. *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'03*, Volume 3.

[2] D'Arcy, SM, L.P. Wong, and M.J. Russell. 2004. Recognition of read and spontaneous children's speech using two new corpora. *Proc.*, *ICSLP*.

[3] Potamianos, A. and Narayanan, S. 2003. Robust recognition of children's speech. *IEEE Transactions on Speech and Audio processing*, Volume 11, 603-616.

[4] Elenius, D. and Blomberg, M. 2004. Comparing speech recognition for adults and children. *Proc., FONETIK*.

[5] http://curriculum.qca.org.uk/key-stages-1-and-2/index.aspx.

[6] Batliner, A., et al. The PF-STAR Children's Speech Corpus. *Ninth European Conference on Speech Communication and Technology*, ISCA.

[7] Davies, J. and I. Brember. 1994. The Reliability and Validity of the 'Smiley' Scale. British Educational Research Journal, 20(4), 447-454