Critical Success Factors for Automatic Speech Recognition in the Classroom

Steve Bennett[1], Jill Hewitt[1], Barry Mellor[2], and Caroline Lyon[1]

[1] Department of Computer Science, University of Hertfordshire, College Lane, Hatfield Al10 9AB, United Kingdom
S.J.Bennett@herts.ac.uk,
mail@SpeechSoft.co.uk.,J.A.Hewitt@herts.ac.uk,
C.M.Lyon@herts.ac.uk
[2] SpeechSoft Limited,The Carriers,Green End, Sandon, Buntingford, Herts, SG9 0RQ

Abstract. This study looked at continuous automated speech recognition (ASR) to an audience in a university lecture theatre and ran an evaluation based on a previous experiment by Ryba, McIvor, Shakir and Paez, which found that non-native speakers of English were much more favorable towards the use of ASR in class than native speakers. Our evaluation was done on a class of 29 students composed entirely of non-native speakers/ A strong indicator of the level of engagement with the technology was the linguistic ability of the user – the weaker the student's English, the more he or she tended to look at the textual output, the greater distraction experienced through poor recognition and also the greater impatience felt with slow recognition. There also seemed to be cultural differences – the Chinese students appeared to look at the textual output much more than Indian students. We conclude that the 2 axes around which successful classroom speech recognition occurs are those of accuracy and unobtrusiveness. The more accurate and unobtrusive the technology, the more successful will be the automatically transcribed lecture.

Keywords: Automatic Speech Recognition, multimodality, accuracy rate.

1 Introduction

One of the most fascinating recent studies done of speech recognition in class was Ryba,McIvor,Shakir and Paez's Liberated *Learning: Analysis of University Students' Perceptions and Experiences with Continuous Automated Speech Recognition*. [1] Using the Liberated Learning Consortium's tools, such as IBM ViaScribe™, they ran ASR over 3 lectures at Massey University NZ, and then an evaluation which contrasted specifically the opinions of native speakers as against non-native speakers of English, and discovered there a significant disparity between the favorability toward the technology of the two groups. This study not only evaluated how the students reacted to live ASR, but also what they thought of the subsequently edited captioned recordings of the event which were later made available to the students.

The Massey study was a very rigorous experiment involving taking the views of 160 students who had attended one or more of a series of three lectures. 81 of the

C. Stephanidis (Ed.): Universal Access in HCI, Part III, HCII 2007, LNCS 4556, pp. 224–233, 2007.

students were regarded as belonging to the L2 (English as a second language) group, and 79 to the native-English-speaking group (known as L1). Owing to the fact that it covered 3 separate lecturing events, we can be highly confident of the robustness of the findings.

The experiment described here was on a smaller scale but was deliberately analogous to the study of Ryba et al. We sought both to pose many of the same questions as asked in the Massey questionnaire, but also a set of other questions arising from the free-text feedback given by the students in that study.

Our study went to a class of 29 students of whom 17 were from India, 8 from China, 2 from Nigeria and 1 from Pakistan and 1 from Saudi Arabia. These numbers were gathered from self-declaration on their questionnaire. Regarding their proficiency at English language, 2 declared themselves to be lower intermediate, 11 described themselves to be intermediate, 11 described themselves as being upper-intermediate, and 3 described their level as advanced, one student did not respond to this question.

The lecture lasted 1 hour in which Dragon NaturallySpeaking version 8 speech recognition software was used, and at the end of it, a questionnaire was handed out containing 12 Likert scale questions asking their opinions on the use of ASR plus 2 more asking the students to describe their level of English and their country of origin.

Below is a comparison between the setup for the lecture series in the Massey study compared with our own:

Ryba,McIvor,Shakir & Perez	*Bennett,Hewitt,Mellor,Lyon*
1. Laptop computer	(THE SAME) In our case a Sony VAIO … CPU 1.7Mhz RAM 1MB
2. wireless microphone set	(THE SAME) in our case a Plantronics Bluetooth wireless microphone with inbuilt sound card (CS60-USB)
3.Viavoice 10™ – local voice profile	Dragon NaturallySpeaking 8 local voice profile
4.Viascribe™ display interface for automatically transcribing speech into text.	SpeechSoft SpeakView display interface software integrated with NaturallySpeaking
5. Text output to single in-class display via data projector	Slides were projected on one screen while text was projected on another
6. File storage on local hard disk	(THE SAME)
7. File transfer and editing	Done in Smirk (see text) – but N/A for this study
8. Lecture files uploaded to internal network	Done in Smirk – but N/A for this study

In our study there were were two screens being projected: one containing the recognized speech, the other the projected lecture. From which we can infer, it seems that one projector was used for both slides and text output (this is the implication of their *Figure 1* though it is not explicitly stated). In our case we used the software Smirk for the lecture rather than PowerPoint – owing to the fact that it records the sequence of slides together with the audio narration as well as pen movements written over the slide during the lecture. Smirk is a lecture capturing tool developed at the University of Hertfordshire which turns the audio recorded lecture into a sequence of jpg and mp3 files to be played back and streamed over the internet. [2]

Our trial took place in a room which had two data projectors built in – therefore the setup was presumably less onerous than that experienced by Ryba et al where the distracting quality of the setup was mentioned in the conclusions. In our case, the lecturer arrived with a laptop dedicated for speech recognition – and used the Plantronics Bluetooth microphone with it. This in earlier trials conducted by us seemed to give very good recognition, even with minimal training. In this case, the lecturer had simply run through the standard Dragon training procedure lasting no more than 30 minutes. No explicit prior rehearsal of the lecture had been attempted. In the lecture theatre itself, all that was done was to connect the laptop to the second projector, plug in the USB microphone, (checking the battery had been charged!), load (but not directly use) Dragon NaturallySpeaking, and then run SpeakView - our own text acquisition and presentation system which integrates with NaturallySpeaking.

SpeakView is an application which optionally uses PowerPoint. It has a resizable caption window which may occupy the whole display or in conjunction with PowerPoint become a caption window at the bottom of the screen which displays the recognized text. It also adds any user specified punctuation, such as full stops and new lines according to speech delays and has advanced text presentation options such as speed, color, scrolling styles and word/line spacing.[3] SpeakView can also broadcast to mobile phones and a WiFi network. In the experiment attempted here, it was put over a blank PowerPoint presentation on the one screen, while the original lecture text was projected from the other dedicated computer in the room onto a different screen.

Since our study was on a much smaller scale with fewer resources and personnel, allowing much less follow up, we did a less sophisticated questionnaire but one we believe addressed all the major points brought up in the Massey study. Here side by side are the directly related questions:

Ryba,McIvor,Shakir & Perez	*Bennett,Hewitt,Mellor,Lyon*
how much did you use the speech-text display *Not at all| Occasionally|Sometimes| Frequently| Nearly always*	I looked at the text display a lot during the lecture *Strongly disagree|Disagree|Neither agree nor disagree|Agree|Strongly agree*
The display helped me to understand the lecture Strongly Disagree|Disagree| Agree|Strongly Agree	The text display helped me to understand the lecture (options as above)
The display helped me to take notes (options as above)	The text display helped me to take notes (options as above)
I think most students can benefit from the Liberation Learning Project (options as above)	I think most students can benefit from the use of Speech Recognition in class (options as above)

In addition to these we also asked a number of questions arising from the free-text feedback in the Massey study:

Ryba,McIvor,Shakir & Perez			*Bennett,Hewitt,Mellor,Lyon*
Theme	Statements	+/-	Question (Likert Scale)
Visual	*See the words easily*	+	I was in a position in the class where I could easily see the displayed text

Distraction	*It was hard to read and served as more of a distraction*	-	The appearance of text on one of the screens all the time was distracting
Not accurate	*It would be useful if it was more accurate*	-	The speech recognition was accurate
Speed	*The speed of the system*	-	The time taken for the system to recognize the lecturer's words was too long
Colour	*Background colour should be darker (darker than lecture slides when in class) found it distracting in class*	-	The colour of the text on the screen was pleasant

The only questions we asked which did not have any real antecedence in the other study were

1. *The text display helped me translate the lecturer's speech* (Likert)
2. Level of English (Beginner, Lower Intermediate, Intermediate, Upper Intermediate, Advanced)
3. Country of origin

2 The Lecture Content

The lecture was an introduction to human factors in computing, specifically covering the ideas of Donald Norman and lasted 1 hour. There were 10 slides, and 5 occasions upon which the lecturer broke out of the slide system to draw on the tablet display of the dedicated computer. A number of the illustrations of themes contained well worn vignettes that he had delivered before while others were improvised at the time. The only real difference between this and the lecture he might have given without speech recognition was a greater deliberation in the words and the occasional accentuation of pauses. At the end of the lecture, the questionnaire was handed out at the end of the lecture to all the students seated in the class.

3 Evaluation of Results

In the section below, we will compare the results from the Massey study with our own, in the analogous categories. In the Massey study there was a great difference in results between the native speaking and the non native speaking subjects – and since all the members of the class in our experiment were non-native speaking – we therefore have decided the best comparison for our data would be with the *non-native English speaking segment* of the Massey study's student population. There were a total of 81 students belonging to the L2 non-native-English-speaking group, compared with our own 29.

Opinions regarding use of the speech text display:

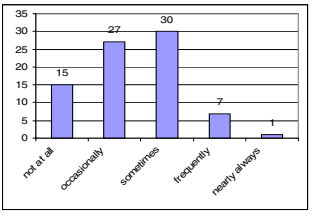

Fig. 1. Massey students (the L2 – non-native-English- speaking group) : ***how much did you use the speech-text display*** (from 80 answers) [in Ryba et al Table 1]

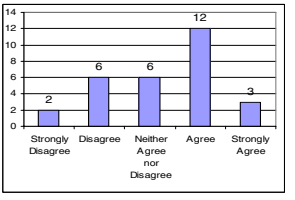

Fig. 2. Hertfordshire students: ***I looked at the text display a lot during the lecture*** (from 29 answers)

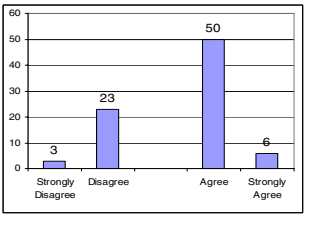

Fig. 3. Massey students (the L2 – non-native-English- speaking group): ***The display helped me understand the lecture*** (from 82 answers) [in Ryba et al Table 4]

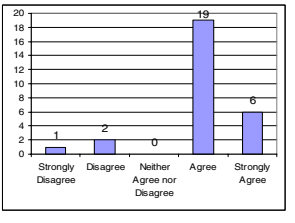

Fig. 4. Hertfordshire students: ***The text display helped me to understand the lecture*** (from 28 answers)

Opinions regarding note-taking and Speech Recognition generally

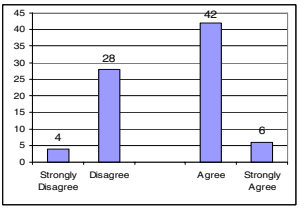

Fig. 5. Massey students (the L2 – non-native-English- speaking group): ***The display helped me to take notes*** (from 80 answers) [in Ryba et al Table 4]

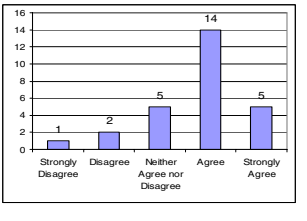

Fig. 6. Hertfordshire students: ***The text display helped me take notes*** (from 27 answers)

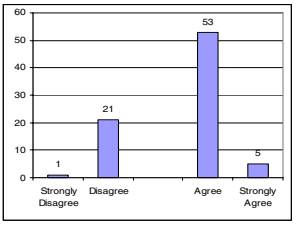

Fig. 7. Massey students (the L2 – non-native-English- speaking group): ***I think most students can benefit from the Liberation Learning Project*** (from 80 answers)

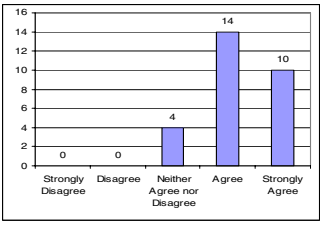

Fig. 8. Hertfordshire students: ***I think most students can benefit from the use of Speech Recognition in class*** (from 28 answers)

The greater favorability made toward our experiment (though over much smaller sample) is most likely caused by slightly better recognition and having two screens in the room and an easier setup. Caution must be expressed over these hypotheses however, because running a study over a much larger group makes it likely that:

1. Any *halo effect* is likely to be more pronounced on the smaller cohort since the relationship between the lecturer and the cohort is likely to be experienced as more immediate
2. The style of lecture is going to be different (less personal and more anonymous). One might also suspect that visibility of screens would be problematic for some of the attendees – however the treatment of this in the Massey study is mixed. While 17 students commented on the easy visuals, 31 commented about screens being hard to read and causing a distraction.
3. The smaller size of the Hertfordshire sample means less trustworthy results.

Similarly, the conclusions of the Massey study are likely to be more robust than ours, because over 3 lectures, one is more likely to obtain an average regarding lecturer performance, whereas in our own case, the positivity of some of the finding might only be a reflection of a lecture going particularly well.

Questions Arising from Free-Text Feedback in the Massey study

Regarding the potentially distracting quality of having the lecturers words outputted on screen, and levels of accuracy, the answers we received were:

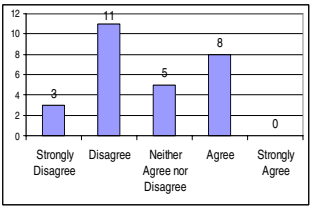

Fig. 9. Hertfordshire students: ***The appearance of text on one of the screens all the time was distracting*** (from 27 answers)

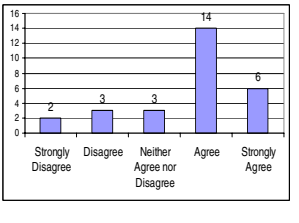

Fig. 10. Hertfordshire students: ***The speech recognition was accurate*** (from 28 answers)

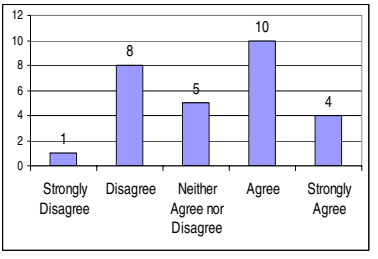

Fig. 11. Hertfordshire students: ***The time taken for the system to recognize the lecturer's words was too long*** (from 28 answers)

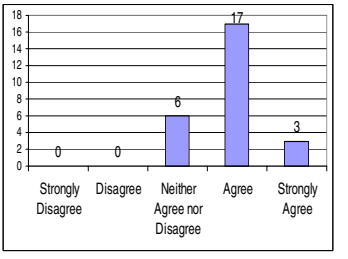

Fig. 12. Hertfordshire students: ***The color of the text on the screen was pleasant*** (from 26 answers)

4 Dependent Variables

In our analysis, there do seem to be some dependencies between the variables in this survey which we can look at through some cross-tabbing. Perhaps the most immediately striking is the difference between Indian and Chinese students. In this survey, there were 8 questionnaires received from Chinese students and 17 from Indian students. When their results are compared we find some interesting disparities:

Table 1. Difference between Indian and Chinese students

Statement	ALL	Indian	Chinese	Difference
I looked at the text display a lot during the lecture	3.28	2.88	4.13	1.25
The lecturer did not change his style to make use of the speech recognition	3	2.64	3.75	1.11
I would describe my level of English as (*Beginner, Lower-Intermediate, Intermediate, Upper Intermediate, Advanced*)	3.57	3.76	2.71	1.05

In this survey, there would seem to be a dependency between the variables relating to self estimation of language ability and country of origin. The Chinese students estimated their level to be considerably lower than that of the Indians. Perhaps this would explain why they seemed to look at the display much more than the Indian students. Equally there seems to be a big difference in the way they believe the lecturer changed his style to make use of the technology – the Indians believing he did change his style, the Chinese not. It is interesting to point out however, that the difference in attention paid to the text display did not show the same degree of difference when this measure was cross tabbed with general language ability (between the weaker English speakers and the stronger ones). Therefore, it does seem likely

that there are some cultural differences at work here as well as ones based on linguistic ability.

If do indeed try to look at the influence of language ability more exactly, and break up the cohort into two groups: those who consider themselves intermediate or lower in their language abilities (langA), and those who consider themselves upper-intermediate or advanced (langB), we find another interesting difference:

Table 2. Comparison of Students with Different Self-Declared Language Abilities

Statement	ALL	LangA	LangB	Difference
The appearance of text on one of the screens all the time was distracting	2.67	3.25	2.14	1.11
The lecturer did not change his style to make use of the speech recognition	3	3.54	2.33	1.21

(to précis: those with lower language ability *agreed the most* that the appearance of text *was* distracting, and *agreed the most* that the lecturer *had not changed his style*).

Therefore, it is those with lower language skills who found the text to be the most distracting – perhaps because they were depending on getting accurate recognition much more than those who did not require it. Similarly those with the lower language abilities most believed that the lecturer did not change his style to suit the technology. Perhaps this can be explained by the fact that since this group are struggling to truly understand the content, and therefore are less sensitive to questions of form and style than those who do not have any such difficulty with language.

5 General Observations

Once we factor out questions of accuracy, it seems the major variable controlling the perceived value of automatic speech recognition in class is the language ability of the individual student – with some element of cultural difference between Indian and Chinese students also visible. However, in general terms we can say, the weaker the student's grasp of English, the more attention they will pay to the textual output display, the greater distraction they will experience through the misrecognition of words, the greater impatience they will feel if there is a long latency between utterance and its recognition, and also the greater tolerance or even non-cognizance of variation in lecturer style owing to the presence of the technology. That is to say the students who can most benefit from this technology, will be the most demanding of it, and the most critical of its shortcomings. Conversely those who were more confident in their English language ability had very little issue with the intrusive nature of the speech recognition – in most cases it seems, they simply ignored the display.

This is quite different from the results of the Massey study where considerable negativity seems to have been registered by the native speakers. A potential explanation might be that, in our case, looking from the visual perspective of the students, the slides of the presentation as well as the lecturer were on the left of the room and the text display was on the screen to the right of the room.

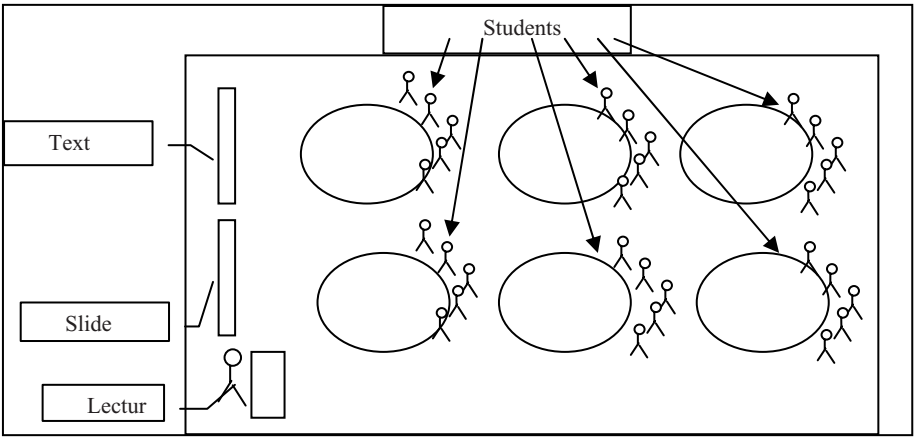

Fig. 13. Layout of the teaching accommodation

This would mean that those students who did not wish to look at the text display and be distracted by it, could just focus on the lecturer instead.

6 Conclusions vis-à-vis the Massey Study

The authors make 8 recommendations. I will go through the first four which relate to the practicalities of speech recognition – since the final 4 have a pedagogic flavor which is beyond the scope of this study.

1. Improved speech recognition – While Ryba et al advance more training as a way to achieve this, we believe the quickest way is to get a good microphone. Our own success with recognition improved enormously after using the Plantronics CS60-USB microphone. This improved our recognition rates enormously. The other intervention involves playing with the settings – both Dragon and ViaVoice have controllers which specify the tradeoff between accuracy and speed. In our experience, slowness seems to be forgiven far more easily than mis-recognition. An unavoidable adjustment does need to be made in the lecturer's style, particularly if there is some latency in the recognition. Good recognition requires a more deliberate delivery and longer pauses. The challenge is for the lecturer to integrate speaking like this with a discursive mode which still remains undeniably theirs.
2. *Improved setup procedure that is less distracting* – this may have been caused by the large cohort size in the Massey study. Our study was facilitated by the fact we used a room which had two data projectors built in. We also began at 9.00 and were able to occupy the room before the students arrived. Certainly it seems that having two built-in projection screens in one of the superior teaching rooms of the university was a major help in the success of our experiment.
3. *More interactive features to improve communication with* students – Ryba et al talk of potential learning scenarios, for instance collaborating as a class to list points for

discussion which then appear on the screen. We think this is an interesting idea, but we again wish to make sure that academics do not have to adjust their practices unnecessarily in order to make use of the technology

4. *Seating arrangements so that there is a specific area where people who wish to see the speech-text screen can chose to sit in that area.* As we mentioned above, the layout of our room meant that for most of the students, the recognized text was not foregrounded. Students could divert their attention to the text display as and when necessary.

This much smaller scale experiment seems bear out and maybe put in sharper relief the majority of the findings of the Massey study. There are however a number of areas where further research is necessary.

Firstly, studies need to be made which correlate these measures of student satisfaction against true measures of recognition accuracy. In our study, unfortunately the software crashed at the moment of saving the transcription of the lecture and so we could only compare student opinions against their *felt* perception of the recognition accuracy, not against an absolute measure of that accuracy itself. Secondly a study such as this needs to be carried out on a larger cohort and over a longer period of time, as was the case for the Massey study, to check whether the more favorable results we obtained as compared to their study might merely be the reflection of more intimate surroundings and potentially superior teaching accommodation and some luck *on the day*.

However, we can conclude that the critical success factors for successful speech recognition in class are 1) level of accuracy of transcription 2) unobtrusiveness of the technology, both at the level of setup and at the level of visualization. If these two factors are attended to, ASR is a technology which can seriously improve the learning experiences of students in the lecture theatre.

References

1. Ryba, K., McIvor, T., Shakir, M., Paez, D.: Liberated Learning: Analysis of Students Perceptions and Experiences with Continuous Automated Speech Recognition. The Electronic Journal of Instructional Science and Technology, vol. 9(1) (2006)
2. Bennett, S., Hewitt, J., Kraithman, D., Britton, C.: Making Chalk and Talk Accessible. In: Proceedings of the 2003 conference on Universal usability (Vancouver, Canada), ACM Press, New York (2003)
3. Hewitt, J., Lyon, C., Britton, M.B.: SpeakView: Live Captioning of Lectures. In: Proc. HCI International, vol. 8 (2005)