Improving the Accessibility of the Traditional Lecture: An Automated Tool for Supporting Transcription

Miltiades Papadopoulos
Teesside University
United Kingdom
M.Papadopoulos@tees.ac.uk

Elaine Pearson
Teesside University
United Kingdom
E.Pearson@tees.ac.uk

The lecture is in its multiple forms the most commonly used method for transferring information in the University curriculum, yet there are serious questions regarding its effectiveness and accessibility in relation to disabled students and those for whom English is not their first language. Although there has been substantial progress that has been made in the area of Automatic Speech Recognition, current systems are still below the level required for accurate transcription of lectures. The Semantic and Syntactic Transcription Analysing Tool is a step forward in the production of meaningful post-lecture materials, with minimal investment in time and effort by academic staff. This paper reports on the results of a study to assess the validity of SSTAT.

Accessibility, Automatic Speech Recognition, Learning Technology, Evaluation.

1. INTRODUCTION

A number of widening participation initiatives have increased the diversity of learners in today's Higher Education. Students with disabilities are attending mainstream postsecondary education programmes in increasing numbers (HESA, 2010) while 13.4% of the total HE population are non-native English speakers (UKCISA, 2011). The traditional lecture, which remains the most dominant method of teaching despite growing criticism of its efficiency (Smith et al., 2006), isolates some students. Those with hearing disabilities find it hard to follow speech and are dependent on intermediaries. Students studying in a foreign language and those whose note taking skills are limited, perhaps through physical disabilities, find lectures hard to follow, understand and recall.

Transcriptions using Automatic Speech Recognition (ASR) technology can be used to increase the flexibility of lectures (Wald and Bain, 2004), thereby promoting inclusive learning and providing equal access to instruction for all learners. Despite the substantial progress that has been made in the area, the performance of ASR systems in real lecture situations is still below the required levels. Research into the usability of speech recognition transcription has determined that an accuracy of 90% is required, a rate that a significant majority of students finds acceptable (Stuckless, 1999; Hede, 2002).

The fundamental aim of this work is to explore innovative ways of improving the editing process of automatically produced lecture transcripts for the production of usable post-lecture material. We propose a mechanism that minimises the evaluation process of erroneous transcription. The object is support for staff to produce meaningful materials that meet the needs of students and reducing the accessibility barriers posed by the traditional lecture format. The paper reports on the results of a three-fold study designed to assess the potential of the proposed mechanism in the production of transcripts that are useful to students, place the minimum of effort on lecturers and enhance the capabilities of standard ASR tools.

2. RELATED WORK

The concept of ASR has captured the attention of researchers in the field of educational technology. Considerable work has being carried out attempting to harness its potential benefits in the instructional environment (Kheir and Way, 2007; Wald and Bain, 2008). The Villanova University Speech Transcriber (VUST) system was designed to improve the accessibility of computer science lectures through the use of computer-assisted real-time

© The Authors. Published by BISL.

transcription. The system was evaluated for recognition accuracy and perceived accessibility and was tested in controlled environments with pre-prepared lecture materials. The overall transcription accuracy of the trained system in the classroom setting was 85% (Kheir and Way, 2007), which is 5% lower than the accuracy rate claimed by Hede (2002) and Stuckless (1999) as being required to achieve meaningful transcripts. The study concluded that in order for the system's accuracy rates to be satisfactory, extensive training by the academics delivering the lecture is necessary.

A study in 2009 (Papadopoulos and Pearson, 2009) measured the transcription accuracy of trained and untrained ASR software in real lecture situations. The results demonstrated an accuracy rate of 68% for the untrained systems. A training process of approximately 100 minutes improved the accuracy by only 4.3%. Training the system mainly improved the recognition of words that are frequent in general English. There was no significant increase in the recognition specialist words and domain-specific terminology. The study concluded that the effort and workload for the editing process, in order for transcripts to be usable, would be unacceptably high.

Despite their efforts, researchers have not managed to develop a break-through solution for accurate lecture transcriptions. There still exists a significant gap between the desirable and actual transcription accuracy level. Extensive human post-editing would still be required for the production of truly usable lecture transcripts.

3. A PRAGMATIC APPROACH TO ACCURATE TRANSCRIPTION

Accepting that neither untrained nor trained systems are suitable for the production of acceptable post lecture materials, we considered a different approach by bringing together research from the Natural Language Processing (NLP) and Human Computer Interaction (HCI) domains to achieve the goal of providing usable lecture transcripts to students with a range of disabilities. We propose an alternative method that adopts computer-aided speech transcription, as opposed to fully automated transcription, together with effective user interfaces. The resultant mechanism simplifies and improves the efficiency of the editing process and targets the re-training process of the speech recognition software. In simple terms, the Semantic and Syntactic Transcription Analysis Tool (SSTAT) follows a similar pattern to other systems that have been utilised in the lecture environment. By combining a transcript restructuring mechanism that assists academics in the editing process and a method for efficient re-targeting of mistranscriptions, which improves the transcription efficiency of the ASR software is improved. The system's architecture comprises of different components, which correspond to the tool's functional activities (Figure 1):

- Recording Unit: provides means for recording the lecture presentation
- Speech Recognition Software: allows for automatic transcription of audio recordings.
- Transcript Reconstruction Unit: handles syntactic and semantic knowledge. The mechanism is able to identify erroneous transcription and categorise them according to their error type.
- Retargeting Feature: identifies the most common mistranscriptions and produces a text file, which is utilised for the retraining process of the speech recognition engine.
- User Interface: allows categorised errors identified in the transcripts to be demonstrated in a user-friendly format.

SSTAT utilises Nuance NaturallySpeaking for the speech recognition component, one of the general-purpose speech-to-text applications. NaturallySpeaking can achieve impressive accuracy rates by trained speakers in controlled

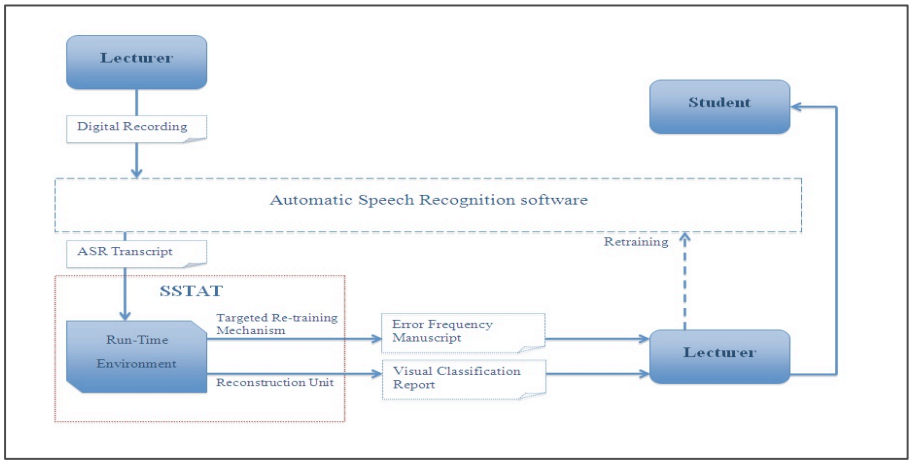

Figure 1. *Basic architecture of the system*

environments (Bennett et al., 2003). Users are required to carry out a brief initial training procedure to allow the software to get used to their voice, speech pace and accent. The preliminary training process involves dictation of social and subject-specific pre-prepared scripts.

Lectures are recorded and the audio files are processed by the ASR software. The original transcripts are then passed onto the Transcript Reconstructing Unit for analysis. Mistranscriptions are distinguished with a visual error categorisation method and presented to the academics as a *Visual Classification Report*. The process facilitates the automatic removal of lexical errors, leaving only those that need examination and judgment by the academics. Syntactic and semantic errors are highlighted in the text and can be easily pinpointed.

The Retargeting Feature identifies semantic errors in the original transcript and their occurrence and produces the *Error Frequency Manuscript*. This acts as the basis for the retraining process of the ASR system by the academic. Mistranscriptions identified are recorded by the academic to re-train the software utilising the 'Add a single word' feature in NaturallySpeaking. A detailed discussion of the technical considerations and design choices that were made during the development of SSTAT can be found in Papadopoulos and Pearson (2011).

4. EVALUATION OF SSTAT

To evaluate the potential of SSTAT in the production of meaningful transcripts in a timely manner, a large-scale study was devised. The study was divided in three phases. The first phase examined the reduction in time required to produce accurate material and the analysis accuracy of SSTAT in identifying erroneous transcription. The second phase assessed the improvement in the performance of the ASR system over a series of lectures utilising the re-training feature. The third phase measured the level of acceptability for students of edited transcripts.

4.1 Phase One: Time Saving and Accuracy

The first phase was designed to evaluate the success of SSTAT in the lecture environment as an error detection and editing tool for academics. The main objectives were to assess the performance of the prototype and to measure the reduction in editing time to produce accurate transcripts and its accuracy in error identification.

4.1.1 Methodology
The experiment was carried out in the context of a real classroom situation. Eighteen lectures were recorded the duration of which varied between 25 to 60 minutes. The lectures were part of Teesside University's undergraduate and postgraduate courses in Computing. The recordings were collected in large amphitheatre-style classrooms and lecturers used head-mounted wireless microphones. Once the recordings were collected, they were handed to the research team for transcription and analysis. Textual transcripts were obtained using the Nuance NaturallySpeaking toolkit. To determine the reduction time and evaluation accuracy, the editing process was divided into two tasks and the same procedure was repeated for each lecture recording.

Task 1 – Lecture recordings were submitted to the trained ASR system and the original transcripts were manually edited by the research team. Edited mistranscriptions included errors that altered the intended meaning of sentences and the most obvious syntactic and lexical inconsistencies.

Task 2 – The original transcripts were submitted to SSTAT for processing. The analysed transcripts were edited based on the errors identified.

The time required to complete each task was calculated and compared to the results of the preceding task. The accuracy of the analysis was calculated by identifying the transcription errors in the original transcripts and comparing them to the mistranscriptions identified by SSTAT.

4.1.2 Editing Time Reduction Results
The recordings were obtained from 4 different modules; Research Methods, Multimedia Development, Mobile Technologies and Creative Web Applications. Analysis of the results confirmed that the editing time for the analysed transcripts was significantly lower than that of the originals by an average of 40%. The first three modules followed a structured presentation procedure based on PowerPoint slides and contained a large amount of subject-specific language. Creative Web Applications was a non-technical module based on discussion between the lecturer and students on innovative design techniques over several websites and multimedia applications that were presented.

Reduction in the editing time for the transcripts of the first three modules was far greater (42%) than that of the final module (33.6%). Although reduction in editing time for transcripts derived from free-style types of lectures is significant, the prototype is more effective when dealing with structured lecture presentations. Table 1 demonstrates the results.

Table 1: *Reduction in editing time by module*

Lecture Transcript	Decrease (%)
Research Methods	*42.0%*
Multimedia Development	*42.3%*
Mobile Technologies	*41.1%*
Creative Web Applications	*33.6%*

4.1.3 Analysis Accuracy Results
Results for the accuracy of the analysis demonstrated that SSTAT was able to achieve a mean accuracy of approximately 83%. The prototype was particularly efficient in recognition of lexical inconsistencies (Lex/E), which include false starts, hesitations and repeated words. The overall accuracy for lexical errors was calculated to 88%. Repeated words are automatically categorised and highlighted as lexical inconsistencies by the *Reporting Function*, therefore identification of this category reached an approximate value of 100%. Identification of hesitations and false starts is achieved by matching every word in the transcript against the *Lexical Inconsistencies Database*, which is an extensive list of common disfluencies and filler words such as 'um' and 'erm', as well as discourse markers such as 'you know', 'I mean', and 'sort of'. A number of disfluencies were incorrectly transcribed as normal English words, such as 'um', which in many cases being transcribed as 'I', or 'erm' as 'air'. This resulted in many words being categorised as syntactic or semantic errors, rather than lexical inconsistencies.

Accurate identification of semantic (Sem/E) and syntactic (Syn/E) errors reached 78% and 83% respectively (Table 2). Most subject-specific words were identified in the transcripts because vocabulary banks contained most of the technical language, included in each transcript. Semantic mistranscriptions not identified by the tool, in most cases did not affect the meaning of sentences. Syntactic errors (Syn/E) that were not identified decreased transcripts' readability, however according to editors' comments they did not seem to lead to an inaccurate semantic interpretation of the sentence. Despite the extensive list of syntactic rules included in the *Prolog Grammar Rules* module, there were numerous syntactic errors that were not identified. This was due to lecturers' spontaneous speech that in many cases does not follow written language's formal syntactic rules.

Table 2: *Analysis accuracy for each error type*

Lecture Transcript	Sem/E	Syn/E	Lex/E
Research Methods	*76%*	*81%*	*86%*
Multimedia Development	*81%*	*85%*	*88%*
Mobile Technologies	*75%*	*83%*	*90%*
Creative Web Applications	*79%*	*83%*	*89%*

4.2 Phase Two: Effectiveness of the Targeted Re-training Feature

This phase consisted of a case study to assess the efficiency of SSTAT in reducing the required effort over a series of lectures on the same subject, by the same lecturer. The main aim was to examine the level of improvement in the transcription accuracy of the speech recognition software, utilising the Error Frequency Manuscript.

4.2.1 Methodology
Three series of lectures were recorded, the duration of which varied between 25 to 55 minutes. Two of the modules were part of Teesside University's undergraduate curriculum in Computer Science, while the third series was based on a first year undergraduate module in Multimedia. The recording process followed the procedures described in the previous experiment. Recordings were collected in real classroom situations, using head-mounted wireless microphones. Nuance NaturallySpeaking was the primary speech recognition engine.

The experiment was divided into four tasks and followed an iterative process to examine whether the accuracy of the transcripts improves over time through the use of SSTAT. The procedure was repeated for each lecture and the results of the experiment were averaged.

Task 1 – Lecture recordings were collected and submitted to the trained ASR system (Nuance NaturallySpeaking). The accuracy of the transcripts was calculated without any further processing.

Task 2 – NaturallySpeaking outputs were submitted to SSTAT for processing. The lexical inconsistencies were identified and automatically removed. The semantic and syntactic errors identified were colour-coded for easy identification and output to the 'Visual Classification Report' file.

Task 3 – The ASR software was retrained based on the 'Error Frequency Manuscript' and the lecture recording was re-submitted to the retrained system. The accuracy of the transcripts was calculated without any further processing.

Task 4 – The transcription files were analysed again by SSTAT and edited based on the errors identified. The consequent recording from the lecture series was submitted to NaturallySpeaking and the process was repeated.

4.2.2 Results
A series of five lectures from three different modules were recorded. Two undergraduate-level modules, Multimedia Design and Web Authoring, and one postgraduate level module titled Research Methods. They all included a plethora of technical words and subject specific language. Academics delivering the presentations utilised PowerPoint slides and followed a standard lecture process.

Initially, all five recordings of each module were submitted to the trained speech engine and the original transcripts' accuracy was measured for comparison purposes in the latter stages of this experiment. The mean accuracy rate was calculated to approximately 75%. Automatic removal of lexical inconsistencies, after SSTAT's analysis, raised it to 79.6%. The transcripts would still require substantial editing efforts to reach an acceptable level of quality, due to numerous topic-specific mistranscriptions that could affect the intended meaning of the text.

Subsequently, Nuance NaturallySpeaking was trained based on the Error Frequency Manuscript, which was produced during the second task and the most common mistranscriptions were added to the engine's vocabulary. The recordings were then resubmitted to the ASR system to examine whether targeted re-training improved the accuracy of the system. The results (Table 3) demonstrated a 4% increase between the mean accuracy of the original transcripts and that of the transcripts after re-training. The greatest increase was revealed in the second presentation for Multimedia Development (7.1%), while the lowest one was found in Web Authoring's third transcript (1.9%).

Table 3: *Accuracy rates between tasks*

Lecture Transcript	Original System	Retrained System
Research Methods 1	*77.3%*	*81.4%*
Research Methods 2	*76.7%*	*81.5%*
Research Methods 3	*76.3%*	*80.1%*
Research Methods 4	*79.2%*	*85.0%*
Research Methods 5	*80.7%*	*85.9%*
Multimedia Development 1	*70.2%*	*77.3%*
Multimedia Development 2	*74.9%*	*78.1%*
Multimedia Development 3	*75.6%*	*80.8%*
Multimedia Development 4	*75.4%*	*81.1%*
Multimedia Development 5	*75.9%*	*80.7%*
Web Authoring 1	*71.2%*	*74.2%*
Web Authoring 2	*73.7%*	*74.9%*
Web Authoring 3	*74.1%*	*76.0%*
Web Authoring 4	*75.8%*	*78.2%*
Web Authoring 5	*75.6%*	*78.6%*
Average (error category)	***75.5%***	***79.57%***

The experiment also aimed to investigate the efficiency of the proposed method over a series of lectures on the same subject. Each module consisted of five lectures and so the experiment measured the improvement in the accuracy of the speech engine after four passes through the system's cycle (Table 4). There was an overall increase of 4.5% between the first and fifth transcript. The greatest increase was in the Multimedia Development module (5.7%). MD was the most technically sound module, with the transcripts containing a great amount of subject-specific words. The re-training process improved the efficiency of the speech system in accurately transcribing terminology words that were mistranscribed in the previous iterations.

Table 4: *Accuracy rates after four passes through the system's cycle*

Lecture	First Transcr.	Final Transcr.	Increase (%)
Research Methods	*77.3%*	*80.7%*	*3.4%*
Multimedia Development	*70.2%*	*75.9%*	*5.7%*
Web Authoring	*71.2%*	*75.6%*	*4.4%*

The proposed training mechanism improved the accuracy by 4% on the same transcript, while it required four passes through the cycle to achieve an analogous improvement for dissimilar recordings of the same module. Each lecture presentation is focused on a different topic of the same subject and therefore contains additional technical language. The efficiency of the targeted re-training mechanism lies in its ability to train the system's language models on subject-specific language. As a result new terminology words would require additional training on these specific words, in order for NaturallySpeaking to be able to transcribe them correctly.

4.3 Phase Three: Transcript Usability and Usefulness

The purpose of this study was to determine the level of accuracy required to render transcripts usable by students and to verify SSTAT's ability to generate meaningful transcripts. The experiment aims to investigate how partially correct transcripts affect learners' usability perception, in an attempt to suggest an appropriate level of accuracy in which to aspire to. More specifically, this study examined:

- Usability – Quality Hypothesis: Learners' perceived usability of transcripts will increase with improved transcript quality. It was expected that perceived usability increases as the quality of transcripts improves. Transcripts would become usable when the accuracy level is approximately 90%.
- Usefulness – Quality Hypothesis: Transcripts' usefulness to learners is influenced by the quality of transcripts. Usefulness should increase as the quality of transcripts improves. It was expected that transcripts would become useful at an accuracy level of approximately 90%.

Although the term quality encompasses a wide range of attributes, such as whether the transcript

represents a context-preserved material, maintains the meaning it is intended to convey and retains a adequate level of readability, for the purposes of this experiment the term quality refers to the accuracy level of the transcript as the intention is to determine a 'low figure' at which transcripts become accepted and usable by students.

4.3.1 Measures
The purpose of this study was to assess how the usefulness of automatically generated transcripts is affected by transcript quality and how it affects students' perceived usability. Usefulness and usability are closely related terms (Landauer, 1996), however it is important to define them in order to justify the measures and instruments used and verify the hypotheses formulated in the previous section. Usability is a key term in computing and HCI research and refers to the quality of a system with respect to ease of learning, ease of use and user satisfaction (Rosson and Carroll, 2002). Usefulness which is less commonly used in the literature, describes the extend to which a system performs the functions it was designed to perform (Grudin, 1992). Based on these definitions two types of data were collected to compare the effect of transcripts' accuracy to students:

- Students' perception data will reveal evidence related to learners' experience when utilising the produced transcripts. Transcripts with increased readability that allow them to effortlessly pinpoint important information constitute transcripts with increased usability.
- Students' 'Context-Specific Questions' task to assess the relation between accuracy level and usefulness of transcripts. Useful transcripts should allow users to understand the content of the presentation and to capture the knowledge that is transferred during the lecture.

Students' perception data was assessed using a series of indicators derived from the distributed questionnaires. A questionnaire examined users' perceived usability of transcripts at a particular level of accuracy. Questions required users to indicate the degree of their agreement to various statements. Indicators include:

Perception of transcription quality, which included questions about the acceptability of accuracy, the degree to which misspellings affect their comprehension of the lecture and their overall experience utilising the transcript.

Perceived acceptance of transcripts, where learners indicated the degree of agreement to the statement: "I would rather have this transcript than not have any transcripts at all".

Usefulness entails a number of qualitative characteristics such as how well transcripts capture the meaning of what has been spoken or the extent to which they allows users to comprehend the information transfer in the class. To assess the usefulness of the transcripts, six sets of four context-specific questions were prepared. Three sets required context-specific critical thinking, while the rest required students to include topic-specific terminology in their responses. Each participant was required to answer a different set of questions from each category, for each part of the experiment. Combinations of the sets of questions were randomly distributed to avoid repetition and order effects. It was expected that editing would need to be performed, in order for students to be able to answer the terminology questions correctly.

The evaluation instruments used were structured questionnaires. Results were analysed to determine students' experiences utilising the transcripts, rather than to identify the causes for their answers. Thus, questionnaires followed a typical five-level Likert scale format, using continuous variables ranging from "Strongly Disagree" (1) to "Strongly Agree" (5).

4.3.2 Transcript Quality
The accuracy of transcripts was the independent variable for this experiment. We assessed the effect of transcripts' quality at three different levels:

- Original transcripts – Transcripts produced by the trained ASR engine. Transcripts' mean accuracy was approximately 78%.
- Analysed transcripts – Transcripts produced by SSTAT with removed lexical inconsistencies and improved formatting. The mean accuracy was calculated to 82%.
- Revised transcripts – Analysed transcripts that have undergone semantic and syntactic post-editing, based on SSTAT's report. The mean accuracy of the final transcripts was calculated to 88.5%.

Accuracy for each level was computed by counting correctly transcribed words and comparing them to the total number of words on each recording. To verify the total number of words, a manual verbatim transcript of each file was generated. Incorrectly transcribed words included mispronounced, substituted, added and omitted words.

4.3.3 Methodology
The study took place in the context of a classroom situation, which consisted of a 20-minute lecture presentation. Evaluations were conducted with Computing (IS) and English Literature (EL) students to test possible variations in perceived usability of transcripts across different academic disciplines. The IS group attended a presentation

on research methods, while EL students a lecture on history of English drama. Recordings were collected using wireless head-mounted microphones. Academics were required to undertake initial training of NaturallySpeaking. The process of the experiment was divided into three tasks:

Task 1 – Lectures were recorded and transcribed using the ASR system. The original transcription output (Transcript A) was saved to a word document and was distributed to the participants. Participants were required not to take manual notes during the lecture and were instructed to review the resulting transcript and complete a structured questionnaire to assess their understanding of the lecture and a survey on their views of the transcript's quality, usability and readability.

Task 2 – The original transcript was then processed by SSTAT and lexical inconsistencies were removed. The text was divided into paragraphs manually to improve formatting. The revised transcript (Transcript B) was sent to participants electronically, three days after the original lecture, at which point they needed to complete a second questionnaire. The aim of this part of the experiment was to examine the extent to which removal of lexical inconsistencies (false starts, hesitations and repeated words) improves the overall quality and usability of the transcript.

Task 3 – The transcript was then edited manually, based on the errors reported by SSTAT and emailed to students (Transcript C) five days after the completed questionnaires had been received by the research team. Similarly, participants were instructed to fill in a final questionnaire and return it to the instructor.

4.3.4 Participants
Twenty-six participants were involved in the experiment. Fourteen of them were Computing students, while twelve were English Literature students. The sample included students with a range of disabilities, which affected their note taking skills and students for whom English is not their first language. In addition, the sample included native English students with no known disabilities, who considered lecture transcripts beneficial for improved access to lectures' content.

Table 5: *Experiment Demographics*

Demographics (N=26)		**Sample**	**Group**	
			IS	**EL**
Gender	*Male*	*15*	*11*	*4*
	Female	*11*	*3*	*8*
Academic Level	*U/G*	*17*	*5*	*12*
	P/G	*9*	*9*	*-*
Language	*English*	*18*	*8*	*10*
	Overseas	*8*	*6*	*2*
Accessibility Requirements	*Deaf/Hard of hearing*	*3*	*2*	*1*
	Dyslexic	*7*	*4*	*3*
	Mobility Impairments	*1*	*1*	*-*
	None	*15*	*7*	*8*

The aim of this experiment was to assess students' perceived usability of transcripts at three levels of accuracy. Human post-editing was necessary for the purposes of this study, which was performed by members of the research team.

4.3.5 Usability Results
A repeated-measure Analysis of Variance (ANOVA) was used, as it constitutes an appropriate statistical test for within-subjects design with various confounding effects and multiple-level independent variables (Howell, 1999). Tests were conducted using a significance level of $\alpha = 0.05$ as the size of the null hypothesis' rejection region. To check which of the transcripts had the greatest impact on perceived usability, additional paired sample t-tests were conducted to make comparisons between the three conditions. Beside the tests for statistical significance, descriptive statistics were also used for some of the students' performance indicators. To assess the perception of the transcription quality two questions were included in the questionnaire:

- After reviewing the full transcript, I feel its accuracy is sufficient.
- Too many misspellings in the text make the transcript hard to understand.

Results for the first statement for the Computing group (Table 6) indicate a statistically significant difference in perceived transcription accuracy of transcripts; $F_{2,26} = 16.695$, $p < 0.001$. This was also confirmed for the second question; $F_{2,26} = 13.473$, $p < 0.001$. The t-tests demonstrated that participants were more aware of transcription errors in transcripts A and B.

For the EL group, the results (Table 7) of the first statement demonstrated a statistically significant difference in their perception of accuracy for the three transcripts; $F_{2,22} = 18.292$, $p < 0.001$. The conducted t-tests showed that perceived accuracy of transcripts was not improved between transcripts A and B ($t_{11} = -2.708$, $p = 0.20$), however a significant improvement was revealed for the third transcript ($t_{11} = -3.546$, $p = 0.005$). Equivalent results were demonstrated for the second statement; $F_{2,22} = 15.681$, $p < 0.001$. The t-tests highlighted a statistically significant improvement for the third transcript ($t_{11} = 4.841$, $p = 0.001$).

Table 6: *Perceived transcription quality for the IS group*

Statement	Test	A	B	C
Accuracy is sufficient	*ANOVA*	$F_{2,26} = 16.695, p < 0.001$		
	means	*2.57*	*3.14*	*3.86*
	t-tests	$t_{13} = -2.511, p = 0.026$		
			$t_{13} = -3.680, p = 0.003$	
Misspellings make transcript hard to understand	*ANOVA*	$F_{2,26} = 13.473, p < 0.001$		
	means	*3.36*	*3.21*	*2.21*
	t-tests	$t_{13} = 0.806, p = 0.435$		
			$t_{13} = 4.266, p = 0.001$	

Table 7: *Perceived transcription quality for the EL group*

Statement	Test	A	B	C
Accuracy is sufficient	*ANOVA*	$F_{2,22} = 18.292, p < 0.001$		
	means	*1.50*	*2.50*	*3.83*
	t-tests	$t_{11} = -2.708, p = 0.20$		
			$t_{11} = -3.546, p = 0.005$	
Misspellings make transcript hard to understand	*ANOVA*	$F_{2,22} = 15.681, p < 0.001$		
	means	*3.92*	*3.58*	*2.42*
	t-tests	$t_{11} = 1.301, p = 0.220$		
			$t_{11} = 4.841, p = 0.001$	

The tests revealed that the quality of transcripts affects students' experience for perception of usability. According to the results, students valued the accuracy level of transcript C (88.5%) and should be concluded that an accuracy level equal or greater to the third transcript's quality is required.

Perceived acceptance was assessed by the statement in the questionnaire: "I would rather have this transcripts than not have transcripts at all". Simple descriptive statistics were used for the analysis of the results. Regarding the Computing group, 57% disagreed on the statement for the first transcript, while only 21% of participants indicated that they would rather have a transcript at this level of quality. Transcript B's perceived acceptance was not significantly greater. Overall, acceptance was increased by approximately 8% and was calculated to 29%. Syntactic and semantic correction, improved the positive views towards transcript C dramatically. Students valued the final editing with 79% indicating that they would rather have transcript C than no transcripts at all. Figure 2 demonstrates the results.

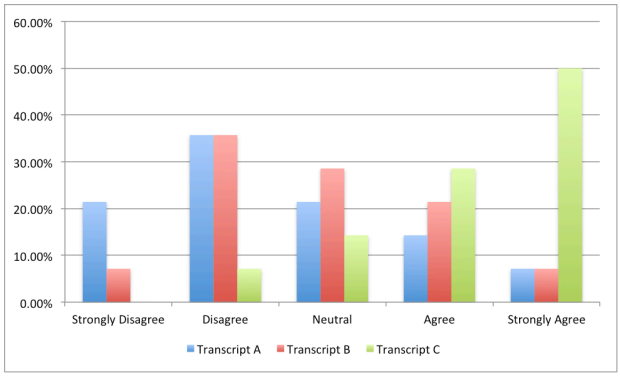

Figure 2: *Perceived acceptance for the IS group*

The results for the English Literature group revealed similar results (Figure 3). Perceived acceptance of the first transcript was predominantly low, while removal of lexical inconsistencies and improved formatting did not seem to radically change students' perception. Opinions were dramatically altered in the last transcript, where students seemed to value the syntactic and semantic corrections.

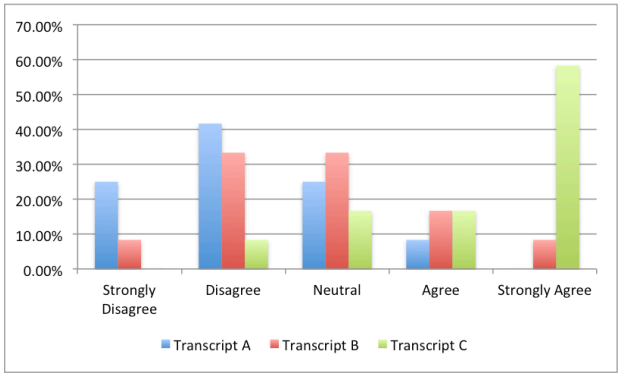

Figure 3: *Perceived acceptance for the EL group*

Comparing the results of the two groups, it was revealed that despite the fact that the original transcript's accuracy for the History of English drama lecture was noticeably higher that that of the IS group's presentation, only 8% of EL students seemed to value transcript A, compared to 21% for the IS population. This might be due to the fact that computing students seem to appreciate technological advances in the educational

environment more than students in less technical academic domains.

4.3.6 Usefulness Results

The aim of this part of the experiment was to examine whether the level of accuracy of the transcripts had an effect on students' understanding of the lecture presentation. Participants' answers were reviewed and ranked into three categories; incorrect, incomplete and correct responses. Each three-question test had a maximum value of 3 points with 1 point awarded for correct responses, no points awarded for incorrect answers, while each incomplete response was worth 0.5 points. Typically, half-complete answers included partially correct responses caused by speech recognition errors. A Friedman's test was conducted to test the transcripts' usefulness across the three conditions. Post-hoc analysis with Wilcoxon Signed-Rank tests with a Bonferroni correction applied was conducted to examine the changes of scores from one condition to the other.

There was a significant difference in correct responses for the context-specific terminology questions between each of the three transcripts for the Computing group; $\chi^2(2)= 23.132$, $p < 0.001$. Post-hoc analysis with Wilcoxon Signed-Rank tests was conducted with a Bonferroni correction applied, resulting in a significance level set at $p < 0.025$. There were no significant differences between transcripts A and B ($Z = -1.592$, $p = 0.111$), however an overall increase in correct responses was observed between transcripts B and C ($Z = -3.335$, $p = 0.001$). Following an identical process for the critical thinking questions, the results demonstrated a significant difference in the perceived understanding of the lecture depending on the transcript that was used; $\chi^2(2)= 22.167$, $p < 0.001$. A post-hoc analysis revealed that despite the fact that there were no significant differences between transcripts A and B ($Z = -1.732$, $p = 0.083$), there was a statistically significant increase in correct responses between B and C ($Z = -3.267$, $p = 0.001$). Regarding the EL group, there was a significant difference in correct answers for the terminology quiz; $\chi^2(2)= 12.667$, $p = 0.002$. A post-hoc analysis showed that similarly to the Computing group, there were no significant differences between transcripts A and B ($Z = -0.707$, $p = 0.480$), however there was a significant increase between B and C ($Z = -2.714$, $p = 0.007$). Similar results were observed for the critical thinking questions; $\chi^2(2)= 12.062$, $p = 0.002$. Comparing the transcripts, no statistical difference between transcripts A and B was observed ($Z = -0.264$, $p = 0.792$), while there was an increase between B and C ($Z = -2.546$, $p = 0.011$).

The mean rank for correct responses for the critical thinking and terminology questions for the first two transcripts was almost equal for the English literature students, while there is a much greater difference for the two tasks for the computing students. This is due to the fact that terminology for the two subjects is different. Computing includes technical language, which cannot always be transcribed by ASR software without an extensive subject-specific vocabulary and is usually harder for students to remember. On the other hand, subject-specific language for English literature does not include technical terms and is, therefore, easier for current systems to transcribe. Once the semantic and syntactic corrections had been performed, the number of correct answers was practically equal for the two tasks for both groups.

5. DISCUSSION & FUTURE WORK

SSTAT demonstrated a significant potential as an automated method for supporting the editing process of lecture transcripts. The Reporting Function allowed academics to reduce the editing time by approximately 41%, while the Text Analysis module achieved an accuracy of 83% at recognising transcription errors, which can be further improved by enriching the syntactic, semantic and lexical inconsistencies rules of the prototype. The Retargeting Feature improved the accuracy of the speech engine by an average of 4.3% after four passes through the system's cycle. The figure might not seem impressive, however the results are encouraging. A mean speaking rate of 72 words per minute, translates to 4,320 words per lecture, assuming that each lecture lasts for approximately 60 minutes. A 4.3% improvement results in 186 fewer transcription errors. Considering the fact that most errors constitute semantic mistranscriptions affecting the intended meaning of the spoken utterance, a comparable accuracy increase could improve the overall usability of the transcripts significantly.

Analysis of students' perception data revealed that perceived quality and acceptance correlate with the accuracy of transcription. The 'Context-Specific Questions' task demonstrated that performance was not greatly affected using the first two transcripts, while exposure to the third transcript resulted in a significant increase in users' performance. Inferential and descriptive statistics confirmed the hypotheses and highlighted transcripts C's usefulness as an alternative to traditional note taking. Considering the fact that the mean accuracy rate of the third transcript for both groups was 88.5%, it may be safe to assume that an accuracy level of 88.5% or greater can be considered sufficient for the production of usable post-lecture material.

Building on the evaluation results, there are several aspects of SSTAT that merit further investigation. The Reporting Function relies heavily on visual codes and presents the identified inaccuracies in a colour-coded manner. Mistranscribed words in the text are highlighted according to their error type so that they can be easily interpreted. However, there are accessibility problems associated with colour-coding techniques and consequently with SSTAT. A program that requires users to distinguish between identical shapes of different colours could pose problems to people with vision impairments. Further work includes a comprehensive investigation of alternative approaches of conveying information. Effective colour scales, which can provide a range of colours varying in hue, saturation and brightness or additional tools, enabling users to adapt colour scales according to their needs, will be examined. Future work will also involve a systematic usability evaluation of the interface with academics to identify possible usability issues and design defects. Finally, we are currently exploring possible ways to include an automatic correction process for the identified mistranscriptions.

6. REFERENCES

Bennett, S., Hewitt, J., Kraithman, D. and Britton, C. (2003) Making Chalk and Talk Accessible. In: *Proceedings of the 2003 Conference on Universal Usability (CUU 2003).* Vancouver, British Columbia, Canada. 119-125.

Grudin, J. (1992) Utility and Usability: Research Issues and Development Contexts. *Interacting with Computers,* 4(2). 209-217.

Hede, A. (2002) Student Reaction to Speech Recognition Technology in Lectures. In: *Proceeding of the Australian Society for Educational Technology Conference (ASET 2002).* Melbourne, Australia. Available at: http://www.ascilite.org.au/aset-archives/confs/2002/hede-a.html (Accessed: 11/03/2012)

HESA (2010) Student Record Data 2009/2010. Available at: http://www.hesa.ac.uk/ (Accessed: 11/03/2012)

Howell, D. (1999) *Fundamental Statistics for the Behavioral Sciences* (4th edn). Pacific Grove, CA: Duxbury Press.

Kheir, R. and Way, T. (2007) Inclusion of Deaf Students in Computer Science Classes Using Real-Time Speech Transcription. In: *Proceedings of the 12th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education (ITiCSE 2007).* Dundee, Scotland. 261-265.

Landauer, T. (1996) *The Trouble with Computers: Usefulness, Usability and Productivity.* Cambridge, MA: MIT Press.

Papadopoulos, M. and Pearson, E. (2009) An Analysing Tool to Facilitate the Evaluation Process of Automatic Lecture Transcriptions. In: *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (E-Learn 2009).* Vancouver, British Columbia, Canada. 2189-2198.

Papadopoulos, M. and Pearson, E. (2011) A System to Support Accurate Transcription of Information Systems Lectures for Disabled Students. In: *Proceeding of the 22nd Australasian Conference on Information Systems (ACIS 2011)*. Available at: http://aisel.aisnet.org/acis2011/35/ (Accessed: 11/03/2012)

Rosson, M. and Carroll, J. (2002) *Usability Engineering: Scenario-Based Development of Human-Computer Interaction.* San Francisco, CA: Morgan Kaufmann.

Smith, A., Ling, P. and Hill, D. (2006) The Adoption of Multiple Modes of Delivery in Australian Universities. *Journal of University Teaching and Learning Practice,* 3(2). Available at: http://jutlp.uow.edu.au/2006_v03_i02/smith008.html (Accessed: 11/03/2012)

Stuckless, R. (1999) *Recognition Means More Than Just Getting the Words Right*. Speech Technology. Available at: http://www.speechtechmag.com/Articles/Editorial/Feature/Recognition-Means-More-Than-Just-Getting-the-Words-Right--29567.aspx (Accessed: 11/03/2012)

UKCISA (2011) International Student Data. Available at: http://www.ukcisa.org.uk/about/statistics_he.php (Accessed: 03/03/2012)

Wald, M. and Bain, K. (2004) Using Automatic Speech Recognition to Assist Communication and Learning. In: *Proceedings of the 11th International Conference on Human-Computer Interaction.* Las Vegas, Nevada, USA. Available at: http://eprints.soton.ac.uk/id/eprint/260730 (Accessed: 11/03/2012)

Wald, M. and Bain, K. (2008) Universal Access to Communication and Learning: The Role of Automatic Speech Recognition. *Universal Access in the Information Society*, 6(4). 435-447.