A System to Support Accurate Transcription of Information Systems Lectures for Disabled Students

Miltiades Papadopoulos
Teesside University, M.Papadopoulos@tees.ac.uk

Elaine Pearson
Teesside University, E.Pearson@tees.ac.uk

Follow this and additional works at: http://aisel.aisnet.org/acis2011

Recommended Citation
http://aisel.aisnet.org/acis2011/35

This material is brought to you by the Australasian (ACIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in ACIS 2011 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
A System to Support Accurate Transcription of Information Systems Lectures for Disabled Students

Miltiades Papadopoulos, Elaine Pearson
Accessibility Research Centre
Teesside University
Middlesbrough, United Kingdom
Email: M.Papadopoulos@tees.ac.uk, E.Pearson@tees.ac.uk

Abstract

Despite the substantial progress that has been made in the area of Automatic Speech Recognition, the performance of current systems is still below the level required for accurate transcription of lectures. This paper explores a different approach focusing on automation of the editing process of lecture transcripts produced by ASR software. The resultant Semantic and Syntactic Transcription Analysing Tool, based on natural language processing and human interface design techniques, is a step forward in the production of meaningful post-lecture materials, with minimal investment in time and effort by academic staff and responds to the challenge of meeting the needs of students with disabilities. This paper reports on the results of a study to assess the potential of SSTAT to make the transcription process of Information Systems lectures more efficient and to determine the level of correction required to render the transcripts usable by students with a range of disabilities.

Keywords

Accessibility, Automatic Speech Recognition (ASR), Natural Language Processing (NLP), Human-Computer Interaction (HCI), Information Systems Teaching.

INTRODUCTION

The traditional lecture remains the most dominant method of teaching, despite growing criticism of its efficiency, flexibility and accessibility (Smith et al. 2006). The lecture environment isolates students with hearing disabilities, who find it hard to follow speech and therefore are dependent on intermediaries. In addition, 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. There is a growing awareness of the need to improve the accessibility of the traditional lecture to fulfil the access requirements of students with disabilities.

Research has shown that Automatic Speech Recognition (ASR) technology can be employed to make lectures more flexible through transcription (Wald and Bain 2008). ASR can provide transcribed lecture notes as an alternative to traditional note taking, benefiting those learners, whose needs and preferences are not otherwise met. 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. Challenging environments, such as the lecture theatre, affect the efficiency of current systems and decrease the quality of the resultant transcripts (Papadopoulos and Pearson 2009). Moreover, Information Systems teaching could pose additional challenges to accurate lecture transcription as it is a technical subject matter, where students and lecturers are more focused on technology rather than activities requiring oral expression and it is normal to experience larger classes than other disciplines (Sixsmith et al. 2006).

The fundamental aim of this work is to respond to the challenges for inclusive learning, by suggesting an approach, which allows for the production of usable post-lecture material for Information Systems lectures in a timely and uncomplicated fashion. The research questions this work addresses are:

- Can an automated method be employed to assist academics in identifying inaccuracies in transcripts produced by ASR?
- Can a sufficient level of accuracy be achieved by such a system to create usable transcripts for disabled students?

The work presented in this paper takes a different approach to other major research projects in this area. Combining Human-Computer Interaction based solutions and Natural Language Processing research, it supports lecturers in the editing process and significantly reduces the time and effort required to produce accurate lecture transcripts. The innovative feature of this study is the novel approach we have taken to tackle the issue of providing time-poor academics with a system that minimises the time and effort required to produce usable transcriptions. The resultant mechanism is a step forward in producing meaningful support materials and
addressing the needs of students with disabilities and those studying in a foreign language. The paper reports on the results of a two-fold study designed to assess the potential of the proposed mechanism to make the transcription process of Information Systems lectures more efficient and to determine the level of accuracy required for the transcripts to be usable by students. Evaluations were conducted with 14 undergraduate and postgraduate Information Systems students. The experiment also involved 12 English Literature students, to test possible variations in students’ perceived usability of transcripts across different academic disciplines.

BACKGROUND RESEARCH

There have been a number of initiatives designed to supplement lectures through the use of real time captioning and asynchronous transcription. An early study (Leitch and MacMillan 2003), involving eight institutions and seventeen lecturers, revealed that the mean accuracy rate of the transcripts produced in real lecture situations was approximately 77%. Research into the readability and usability of speech recognition transcription has determined that an accuracy of at least 90% is required, a rate that a significant majority of students finds acceptable (Stuckless 1999; Hede 2002). Wald (2005) suggests that the poor results reported from those experiments are due to the fact that current speech recognition systems are based on models created from written documentation rather than spontaneous speech and that reasonable accuracy rates can only be achieved by committed lecturers after extensive training.

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 transcription. The system was evaluated for recognition accuracy and perceived accessibility and was tested in a controlled environment with pre-prepared lecture materials. The overall transcription accuracy of the trained system in the classroom setting was 85% (Kheir and Way 2007), however the researchers concluded that in order for the system’s accuracy rates to be satisfactory, extensive training by the academics delivering the lecture is necessary.

A sophisticated multimedia annotation system called Synote (Wald 2010) provides lecture recordings synchronised with transcripts, slide images and bookmarks to support learners that have difficulties taking notes. The system’s transcription accuracy rates vary between 70 – 85% for native English speakers. Manual editing is therefore required for the production of meaningful transcripts.

Previous work by the authors (Papadopoulos and Pearson 2009) measured the performance of trained and untrained ASR software in real lecture situations and evaluated the quality of the resultant transcripts. The study concluded that while the overall performance increases with extensive training, the quality of the transcription files does not improve significantly. Despite promising results, these studies demonstrate that current systems are not yet suitable for large-scale deployment in the university classroom. Extensive human post-editing would still be required for the production of truly usable lecture transcripts.

SSTAT – A PRAGMATIC APPROACH

Despite their efforts, major organisations have not managed to develop a break-through solution for accurate lecture transcription without extensive input by the lecturer. There still exists a significant gap between the desirable and actual transcription accuracy level. 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 NLP and HCI domains to achieve the goal of providing usable lecture transcripts to students with a range of disabilities. An approach that adopts computer-aided speech transcription (Luz et al. 2008), as opposed to fully automated transcription, together with effective user interfaces to support it, is proposed as an alternative. A tool that would simplify and improve the efficiency of the editing process, minimise the re-training process and reduce the time required to produce meaningful transcripts could be the step forward. Such a system should support lecturers by detecting incorrect sentences and reporting on the nature of the error in a user-friendly interface. It should support students by producing usable lecture transcripts in a timely fashion. Taking this as a starting point, the Semantic and Syntactic Transcription Analysis Tool (SSTAT) provides three basic levels of functionality; analyse text and identify erroneous syntactic and semantic transcription, classify errors so they can be easily identified and interpreted by academics and remove lexical inconsistencies, such as false starts, hesitations and repeated words.

SSTAT utilises Nuance NaturallySpeaking for the speech recognition component, one of the general-purpose speech-to-text applications that dominate the field of machine recognition. NaturallySpeaking can achieve impressive accuracy rates by trained speakers in controlled environments (Bennett et al. 2003). Users are required to carry out a brief initial training procedure in order to allow the software to get used to their voice, speech pace and accent. The training process involves dictation of social and subject-specific pre-prepared scripts. The process works as follows (Figure 1):

a) The lectures are recorded and the audio files are processed by the ASR software
b) The transcripts are passed to SSTAT for analysis

c) Lexical errors are removed to make the error recognition more readable (optional)

d) The syntactic and semantic mistranscriptions are presented to the academics in a user-friendly format

e) In addition, SSTAT produces a document, called ‘Retargeting Text’, which identifies all the semantic errors in the original transcript, and the number of their occurrences. This document is used as the basis for targeting the retraining process of the speech recognition software

f) The academic records the most frequent mistranscriptions and trains the software by dictating that list, utilising the ‘Add a single word’ feature in Nuance NaturallySpeaking

The process means that obvious lexical errors can be removed automatically leaving only those that need examination and judgement by the academics. Syntactic and semantic errors are highlighted in the text and can be easily pinpointed. The ‘Retargeting Text’ feature aims to improve the efficiency and speed of the re-training process.

**Error Categorisation**

To provide an effective evaluation of the transcripts SSTAT needs to handle both syntactic and semantic knowledge. Prolog was chosen as the programming language, since it is closely tied to the search for computational formalisms for expressing syntactic and semantic analyses for natural language sentences. Standard parsers are designed for grammar checking and cannot necessarily deal with the mixture of social and subject specific language and the spontaneous and creative speech of the lecturer. The tool needs to be able to identify errors that degrade the overall meaning of the sentences. We have, therefore, taken a pragmatic approach to achieve a usable method that can be utilised in the lecture setting.

The primary consideration was to categorise errors, according to their type: Semantic errors – mistranscriptions that corrupt or alter the intended meaning, syntactical errors – ungrammatical constructs that do not affect the meaning of the sentences, lexical inconsistencies – hesitations, false starts and repeated words.

**Vocabulary & Transcription Analysis**

University lectures contain a large proportion of technical and specialist terms. Therefore, ASR systems in the academic setting need to be able to recognise discipline-specific terminology successfully. Speech recognition engines use linguistic resources to recognise words; a pronunciation vocabulary that assigns words to typical pronunciations and language models, which try to predict the frequency of a word and word combination in a body of text.

A recent experiment (Marquard 2011) examined whether ASR vocabularies are able to recognise specialist terms in a particular domain. The reference that was used was a transcript of a lecture delivered to undergraduate Health Sciences students, which contained 7,810 words, using a vocabulary of 1,407 unique words. The vocabulary that was evaluated, Wikipedia 500K, includes a text corpus of the approximately 500,000 most frequently used terms from Wikipedia. The results suggested that the vocabulary was missing numerous domain-specific words, which are infrequent in general English, although they appeared in the lecture transcript.

Thus, SSTAT needs to be able to identify discipline-specific language. The tool utilises several vocabulary banks, which include subject-specific terminology of different disciplines. Words take the form of Prolog facts,
which include their assigned lexical category and form. Utilising Prolog grammar notation, a simple vocabulary bank could look like:

- noun(singular) --> [paradigm].
- noun(singular) --> [interpretivism].
- verb(singular) --> [evaluates].
- verb(plural) --> [evaluate].
- determiner --> [the].
- determiner --> [a].
- adjective --> [philosophical].
- adverb --> [rapidly].
- fullstop --> [.].

Vocabulary banks are text-based files; therefore they can be easily enriched with additional subject-specific and general English terms. Phrase structure and semantic interpretation rules are expressed in Prolog grammar notation as follows:

```prolog
/* Basic Sentence Structure*/
sentence --> np(X), vp(X), fullstop.

/* NP and VP rules */
np(X) --> determiner, noun(X).
np(x) --> noun(x).
np(X) --> determiner, adjective, noun(X).
vp --> verb(x).
vp --> verb(x), adverb.
```

'np' is a noun phrase and 'vp' is a verb phrase. By giving term arguments the plurality of the verbs is tied to that of the nouns. Following the same pattern a more elaborate grammar as well as semantic rules can be created.

Prolog provides the analysis of the lecture transcriptions and identifies syntactic and semantic mistranscriptions, while the reporting function is presented as a graphical interface developed in Java. In addition, Java is utilised for the identification and elimination of lexical inconsistencies in the transcripts.

**User Interface**

Errors need to be demonstrated as a graphical interface, in such a way that they can be easily identified and interpreted. Graphical user interfaces rely heavily on visual codes, as they can be particularly effective in supporting rapid access to information (Woods 1995). Research shows that highlighting has a significant impact on table searching and in addition, colour combinations for textual display can improve reading performance considerably (Wu and Yuan 2003). Inaccuracies in the analysed text are displayed in a colour-coded manner. This means that a particular colour is assigned to each error category. Mistranscribed words in the text are highlighted according to their error type, so that they can be easily interpreted.

<table>
<thead>
<tr>
<th>Type of Inaccuracy</th>
<th>Colour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic errors</td>
<td>Red</td>
</tr>
<tr>
<td>Syntactic errors</td>
<td>Green</td>
</tr>
<tr>
<td>False starts and hesitations</td>
<td>Yellow</td>
</tr>
<tr>
<td>Repeated words</td>
<td>Blue</td>
</tr>
</tbody>
</table>

Taking the simple sentence ‘The studies evaluates’ as an example, it can be easily identified that the sentence is syntactically incorrect. If this particular sentence is entered in SSTAT, Prolog analyses it and produces a text file, containing the error that has been identified and the incorrect words.

**Syntactic Error. Studies evaluates**

The Java programme reads the text, processes it and colour codes the words according to their error type (semantic, syntactic and lexical inconsistencies). In addition, the Java programme is responsible for identifying and removing repeated words, false starts and hesitations in the produced transcripts. Common disfluencies in the transcript, including filler words such as ‘um’ and ‘erm’, as well as discourse markers such as ‘you know’, ‘I mean’, and ‘sort of’ are matched against an extensive database and highlighted according to their error category.

The Analysis Panel (Figure 2) displays colour-coded information about the nature, as well as the number of inaccuracies in the text. Visual support provides memory links so that information is better retained. It will, therefore, enable users to visualise the error category that each colour is assigned to. Users can choose to view the exact number of errors for any category individually and opt to remove lexical inconsistencies. Finally, there
are options for them to print or save the analysed transcript as an HTML file (Figure 2). The design of SSTAT was kept minimal and straightforward to support a design goal of ease of use for academics.

![Image: Analysed text and analysis panel in SSTAT](image.png)

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 (2010). However, we acknowledge that there may be accessibility problems associated with colour-coding techniques and consequently with the SSTAT reporting mechanism. A software program that requires users to distinguish between identical shapes or different colours, could pose problems to people with vision impairments. A usability evaluation of the prototype needs to be conducted to examine the appropriateness of the interface and identify additional usability problems and design defects.

### SSTAT IN PRACTICE

In order to evaluate the potential of SSTAT in the production of truly meaningful lecture transcripts in a timely manner, a large-scale study was devised. The process combines aspects of experimentation and a multiple case study approach. The study is divided in two phases; the first phase aims to assess the effectiveness of SSTAT at reducing the editing time needed by lecturers to produce meaningful materials and improving the accuracy rates of transcripts, while the second phase was designed to determine the level of accuracy required for transcripts to be usable by students with a range of disabilities.

#### Phase One: Editing Efficiency of SSTAT

The first phase was designed to evaluate the success of SSTAT in the Information Systems discipline, as an error detection and editing tool for academics. The main objectives were to determine the reduction in the editing time of the produced transcripts, the level of improvement in their accuracy and finally the level of improvement in the efficiency and speed of the retraining process utilising the ‘Retargeting Text’ feature.

**Methodology**

Five lectures, three postgraduate level lectures in Information Systems, one undergraduate lecture in Information Systems and one undergraduate lecture in Computer Science, were recorded and transcribed. The lectures were presented by four lecturers; two native English speakers and two non-native English speakers. The duration of the lectures varied between 25 to 50 minutes. This set of experiments was divided into three tasks and followed an iterative process, to examine whether the accuracy of the transcripts improves over time through the use of the SSTAT. The procedure was repeated for each lecture and the results of the experiments were averaged.

**Task One** – Lecture recordings were submitted to the trained ASR system and the accuracy of the transcripts was calculated without any further processing.

**Task Two** – The transcript outputs from the ASR software were submitted to SSTAT for processing. The lexical inconsistencies were identified and automatically removed and the syntactic and semantic errors identified were colour-coded for easy identification and output to the ‘Retargeting Text’ file, to be used as a basis for the retraining process of the ASR software.

**Task Three** – The recordings were then re-transcribed by the speech recognition software. The resulting transcription files were analysed again by SSTAT and edited based on the errors identified. The accuracy of the final transcripts was compared with the previous iterations.
The retraining process and post-editing were conducted by three members of the research team. Two of them were responsible for two transcripts each, while the third one was responsible for the final transcript. The participants were not involved in any other stage of the experiment.

Results

The results demonstrated that the use of SSTAT increased the overall accuracy of the transcripts considerably. The mean accuracy of the original transcripts (before processing) was 72.8%. Automatic removal of lexical inconsistencies raised it to 77.9%. Despite the improvement in the accuracy of the files, they would still require further editing to be acceptable as post lecture materials, mainly due to numerous topic-specific mistranscriptions that could affect or degrade the intended meaning of the text.

The retraining process of the ASR software was based on the ‘Retargeting Text’ file produced during the second task of the experiment and the most common inaccuracies were added to the vocabulary of the ASR system. The recordings were then resubmitted to the ASR system. The retraining process resulted in a 4.5% increase in the mean accuracy of the transcripts, which was calculated to 77.3%. Several initially incorrectly transcribed subject-specific words were now transcribed correctly. The transcripts were submitted again to SSTAT with the editing process based on the semantic and syntactic errors identified by the tool. The mean accuracy of the final transcripts in the third task was 87.7%. This is within two per cent of the accuracy rate claimed by Hede (2002) as being required to achieve meaningful transcripts.

Table 2. Accuracy rates for all tasks

<table>
<thead>
<tr>
<th>Lecture</th>
<th>Task One Original Transcript</th>
<th>Task Two Automatic Removal</th>
<th>Task Three Post-editing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research Methods 1</td>
<td>77.47%</td>
<td>81.55%</td>
<td>90.34%</td>
</tr>
<tr>
<td>Research Methods 2</td>
<td>76.33%</td>
<td>80.19%</td>
<td>87.41%</td>
</tr>
<tr>
<td>Web Authoring</td>
<td>70.92%</td>
<td>75.69%</td>
<td>87.32%</td>
</tr>
<tr>
<td>Mobile Technologies</td>
<td>69.06%</td>
<td>74.83%</td>
<td>85.38%</td>
</tr>
<tr>
<td>Multimedia Development</td>
<td>70.17%</td>
<td>77.33%</td>
<td>88.19%</td>
</tr>
</tbody>
</table>

In order to determine whether the editing workload was reduced through the use of tool, the lecturers who gave the presentations were asked to edit the original transcripts, as well as the transcripts produced by SSTAT and calculate the time that was required to complete each task. Participants were asked to allow at least 7 days between the two conditions to avoid confounding due to order effects. Analysis of the results confirmed that the editing time for the final transcripts was significantly lower than that of the originals. Editing for each inaccuracy was mostly a straightforward process, however in cases where the meaning of a sentence was altered, editing was more challenging since users had to either guess the correct word or listen to the lecture recording.

Table 3. Editing time for transcripts before and after analysis

<table>
<thead>
<tr>
<th>Transcript</th>
<th>Editing Time (pre-SSTAT)</th>
<th>Editing Time (after SSTAT)</th>
<th>Decrease (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research Methods 1</td>
<td>90 minutes</td>
<td>52 minutes</td>
<td>42.2%</td>
</tr>
<tr>
<td>Research Methods 2</td>
<td>86 minutes</td>
<td>51 minutes</td>
<td>40.7%</td>
</tr>
<tr>
<td>Web Authoring</td>
<td>110 minutes</td>
<td>61 minutes</td>
<td>44.5%</td>
</tr>
<tr>
<td>Mobile Technologies</td>
<td>70 minutes</td>
<td>36 minutes</td>
<td>48.6%</td>
</tr>
<tr>
<td>Multimedia Development</td>
<td>125 minutes</td>
<td>69 minutes</td>
<td>44.8%</td>
</tr>
</tbody>
</table>

Summary

This experiment confirms that the use of SSTAT increases the accuracy of the resultant transcripts from an average of 72.8% before processing to an average of 87.7%. In addition, the editing time is reduced by an average of 44.2%.

Phase Two: Transcript Usability

The principle aim of this phase was to measure the level of accuracy required for lecture transcripts to be usable by students and to determine whether the quality of the transcripts produced by SSTAT matches students’ perceived level of usability. Despite the fact that the experiment was mainly focused on Information Systems
students, evaluations also included English Literature students, to test possible variations in students’ perceived usability of transcripts across different academic disciplines. The experiment addresses two research questions:

- How did participants value the improvements in each transcript?
- Are there any improvements in the perceived usability between the three transcripts?

In particular, Phase 2 will determine the level of correction required to render the transcripts usable by students and highlight the types of improvements that increase their usability, in order to demonstrate the potential of SSTAT in the production of meaningful post-lecture materials.

Methodology

Twenty-six participants were involved in the experiment, all of them university students studying at undergraduate and postgraduate level. The participants were divided into two separate groups according to their specific subject area. Fourteen of them were Information Systems (IS) students, while twelve were English Literature (EL) students. The experiments took place in the context of a real classroom situation and consisted of a 20-minute lecture presentation on research methods for the IS students, which for the purposes of the experiments contained technical language. The English literature students attended a 20-minute lecture on the history of English drama. Regarding the Information Systems group, eight participants were native English speakers, while six were non-native English speakers. The sample included four dyslexic students, one profoundly deaf learner, one student with hearing impairment, one student with mobility impairment, which affected their note-taking skills, and seven non-disabled students; six overseas students and one native English speaker. The English literature group consisted of twelve native English speakers and two non-native English speakers. The group included three dyslexic students and one student with minor hearing impairments, which affected their note-taking skills.

Each lecture was recorded and transcribed using the ASR system without the use of SSTAT. The transcription output (Transcript A) was saved to a text file and was distributed to the participants. Participants did not take manual notes during the presentation and were instructed to review the resulting transcript and complete a structured questionnaire, in order to assess their understanding of the lecture and elicit their views on the transcript’s quality, usability and readability.

The original transcript was then processed by SSTAT and the identified lexical inconsistencies were removed. In addition, the text was divided into paragraphs 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. The transcript was then edited manually, based on the semantic and syntactic errors reported by SSTAT and emailed again 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.

Students were asked to answer questions regarding the usability and comprehensibility of the transcripts. The aim of the experiments was to determine students’ attitudes and experiences utilising the transcripts, rather than identify the causes for their answers. Therefore, questionnaires followed a typical five-level Likert scale format, using continuous variables ranging from “Strongly Disagree” (1) to “Strongly Agree” (5).

Results

The questionnaires for all parts of the experiment were answered by all participants in both groups (N=26). All students who participated reported experiencing a number of problems during lectures, however only a small percentage (8%) uses technology to overcome the accessibility problems. Most students (77%) tend to lose focus and miss a lot of information due to problems with note taking, or fail to include key points in their notes (69%). Moreover, students report difficulties referring back to their notes (73%), as in many cases poor handwriting and incomplete sentences make them unusable.

A repeated measures ANOVA was conducted to measure the variance in perceived usability across the three different accuracy levels of the transcripts produced by SSTAT. The dependent variable, usability level, was measured by the participants’ responses in the questionnaires. It should be noted that the Likert-scale categories were assumed to constitute interval-level measurement. The results for the IS group demonstrated that SSTAT increased the overall usability of the transcripts; $F_{2,24} = 55.56$, $p < 0.001$. In order to check which of the three transcripts had the greatest impact on students’ perceived usability, two additional paired sample t-tests were conducted to make comparisons between the three conditions. The first t-test revealed a statistically significant increase in usability levels between transcript A ($M = 1.79$, $SD = 0.7$) and transcript B ($M = 2.36$, $SD = 0.93$); $t_{13} = -3.3$, $p = 0.006$. An even greater increase was revealed during the second t-test, which compared transcripts B
and C (M = 3.8, SD = 1.1); \( t_{13} = -7.1, p < 0.001 \). The results suggest that elimination of lexical inconsistencies and improvements in formatting do improve transcripts’ overall quality, while semantic and syntactic corrections have an even more significant impact on their usability (Figure 3).

Analysis of the results for the EL group showed that the overall usability of the transcripts was increased with the use of SSTAT; \( F_{2, 22} = 47.03, p < 0.001 \). However, the first t-test did not indicate a statistically significant increase in the scores between transcripts A (M = 1.33, SD = 0.49) and B (M = 1.67, SD = 0.65); \( t_{11} = -2.345, p = 0.039 \). This suggests that elimination of lexical inconsistencies and improvements in formatting did not affect students’ perceived usability of the transcripts greatly. This might be due to the fact that the original transcription accuracy of the History of English Drama lecture was higher (80.1%) than that of the Research Methods presentation (76.3%). On the other hand, the second t-test revealed a statistically significant increase in usability levels between transcripts B and C (M = 3.33, SD = 0.78); \( t_{11} = -6.504, p < 0.001 \). Figure 3 demonstrates that the mean average of perceived usability for the third transcript is significantly higher than that of transcripts B and C for both groups.

Usability entails a number of qualitative characteristics such as how well transcripts capture the meaning of what has been spoken and their usefulness to students, who rely on them to assist their note taking. To assess the usability 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 participants to avoid repetition and order effects. It was expected that semantic and syntactic corrections would need to be performed, in order for students to be able to answer the questions correctly, especially the topic-specific terminology ones. 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.

The answers of the participants were reviewed and ranked and a Friedman’s test was conducted to test students’ perceived understanding across the three conditions. There was a significant difference in correct responses for the questions that needed critical thinking between each of the three transcripts for both groups; \( \chi^2(2) = 13.00, p = 0.002 \). More specifically for the Information Systems group, 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 (\( Z = -1.414, p = 0.16 \), however an overall increase in the understanding of the topic was observed between transcripts B and C (\( Z = -2.449, p = 0.014 \)). Following an identical process for the topic-specific terminology questions, the results demonstrated a significant difference in the perceived understanding of the lecture depending on the transcript that was used; \( \chi^2(2) = 30.471, p < 0.005 \). A post-hoc analysis revealed that despite the fact that there were no significant differences between transcripts A and B (\( Z = -1.414, p = 0.157 \)), there was a statistically significant increase in correct responses, between transcripts B and C (\( Z = -3.873, p < 0.005 \)). Figure 4 illustrates the results.

For the English literature group, there was a significant difference in correct answers for the critical thinking questions; \( \chi^2(2) = 14.889, p = 0.001 \). A post-hoc analysis showed that similarly to the IS group results, there were no significant differences between transcripts A and B (\( Z = -1.414, p = 0.157 \)), however there was a significant increase between transcripts B and C (\( Z = -2.646, p = 0.008 \)). For the context specific terminology questions, an increase in perceived understanding was also recorded; \( \chi^2(2) = 21.385, p < 0.005 \). Comparing the transcripts (Figure 4), no statistical difference between transcripts A and B was observed (-1.732, \( p = 0.83 \)), while there was an increase between transcripts B and C (-3.162, \( p = 0.002 \)).
An interesting observation is that 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 (Figure 4). This is due to the fact that terminology for the two subjects is different. IS 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 almost the same for the two tasks for both groups.

Summary

Participants’ responses confirmed that only Transcript C produced an accuracy level sufficient to be usable. Considering the fact that the mean accuracy rate of the third transcript for both groups was 87.5%, it may be safe to assume that an accuracy level of 87.5% or greater can be considered sufficient for the production of usable post-lecture materials. It should be noted that most students (69%) agreed that both transcripts A and B are not suitable for use as post lecture materials, mainly because of mistranscribed topic-specific terminology that degrades the meaning of the sentences. Nonetheless, 54% felt that improvements in formatting and elimination of lexical inconsistencies in the second transcript increased its overall readability. The majority of students (73%) regarded both transcripts too inaccurate to be useful for revision purposes. One of the participants stated:

“Misspellings of subject words on key topics (research pyridine v research paradigm) would affect understanding if I read this text at a later time”.

Accurate transcription of subject specific words is essential for acceptable lecture transcripts. Transcripts at this accuracy levels would only be usable combined with additional manual notes:

“Combined with some note-taking this one could be enough”.

On the other hand, 73% of the participants felt that the semantic and syntactic corrections in Transcript C increased its usability considerably and 77% viewed this level of accuracy as sufficient for usable transcripts:

“This is pretty good. I’d definitely use it for revision purposes”.

In addition, one of the deaf students stated that transcripts at this level of accuracy “…could possibly be an alternative to note takers”.

CONCLUSION & FUTURE WORK

Improving the usability of automatically produced lecture transcripts is a task that can be achieved by combining Human-Computer Interaction based solutions and Natural Language Processing research. One of the main challenges to producing meaningful post-lecture materials is the poor performance of current speech recognition systems in the lecture environment. The solution proposed facilitates the editing process of ASR-generated transcripts and, according to the evaluation results, can lead to significantly improved transcripts, in terms of their accuracy and overall usability.

The results of this study suggest that SSTAT demonstrates a significant potential as a computer-aided speech transcription system for supporting the editing process of IS lecture transcripts. The evaluation results revealed that SSTAT produced a significant decrease of approximately 44% in the editing time required for the production
of transcription accuracy of almost 88% after only two passes through the system. The results of the second phase of the evaluation demonstrated that an accuracy level of at least 87.4% is sufficient for the production of usable post-lecture materials for IS students.

SSTAT could prove beneficial in diverse teaching scenarios. Additional experiments need to be conducted to assess the efficiency of the tool outside the traditional lecture environment and across different academic disciplines. A number of experiments have already been planned and are currently ongoing. In addition, a usability evaluation of the interface is underway with academics to identify possible usability issues and design defects.

REFERENCES


**COPYRIGHT**

[Papadopoulos & Pearson] © 2011. The authors assign to ACIS and educational and non-profit institutions a non-exclusive licence to use this document for personal use and in courses of instruction provided that the article is used in full and this copyright statement is reproduced. The authors also grant a non-exclusive licence to ACIS to publish this document in full in the Conference Papers and Proceedings. Those documents may be published on the World Wide Web, CD-ROM, in printed form, and on mirror sites on the World Wide Web. Any other usage is prohibited without the express permission of the authors.