An Ecologically Valid Evaluation of Speech Summarization in the University Lecture Domain

by

Anthony McCallum

A thesis submitted in conformity with the requirements
for the degree of Master of Science
Graduate Department of Computer Science
University of Toronto

Abstract

An Ecologically Valid Evaluation of Speech Summarization in the University Lecture Domain

Anthony McCallum

Master of Science

Graduate Department of Computer Science

University of Toronto

2012

The past decade has witnessed an explosion in the size and availability of online audio-visual repositories, such as entertainment, news, and lectures. Summarization systems have the potential to provide significant assistance with navigating such repositories. Unfortunately, automatically-generated summaries often fall short of delivering the information needed by users. This is due, in no small part, to the fact that the natural language heuristics used to generate summaries are often optimized with respect to currently-used evaluation metrics. Such measures simply score automatically-generated summaries against subjectively-classified gold standards without taking into account the usefulness of a summary in assisting a user achieve a certain goal or even overall summary coherence. Using current evaluation methods, which make use of subjective gold standards and evaluation that is not ecologically valid, even the most linguistically-complex summarization systems perform no better than basic heuristics, such as picking the longest sentences from a general-topic, spontaneous dialog, or the first few sentences from a news recording. With results like these, it is unclear that the speech summarization community can confidently move forward in developing more effective summarization systems.

To address this, we have conducted an ecologically valid evaluation of speech summarization in the university lecture domain. Results indicate that having access to the entire

lecture material is indistinguishable from either a primed or generic manually generated summary, indicating that summaries may be a suitable replacement for lectures in a time-constrained environment. However, automatic summaries generated using MMR result in significantly worse performance than manually generated summaries. Comparison to ROUGE, including using conventional context-free annotated gold standard summaries, failed to show any correlation, raising doubt on whether ROUGE can legitimately be used as a substitute for proper human evaluation.

Acknowledgements

I would like to thank everyone who either directly or indirectly contributed to the research described in this thesis. I would like to thank all of my colleagues in the Computational Linguistics group at the University of Toronto. You have made my time here enjoyable and academically stimulating. I would also like to thank my two collaborators at National Research Council Canada. Xiaodan Zhu has been consistently helpful with the more technical aspects of summarization and the various measures used for evaluation. Cosmin Munteanu has been a mentor to me, and I would like to thank him for sharing his wealth of HCI knowledge, and experience in conducting human subject studies and the resulting data analysis. Finally, I would like to thank my supervisor, Gerald Penn. His guidance has been vital to the success of this thesis. I am fortunate to have had the opportunity to learn from him and be a part of his research group.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

As one of the most natural means of communication between humans, it is not surprising that speech continues to be used effectively in a wide variety of applications. Widespread availability of recording equipment, affordable digital storage, and the pervasiveness of high bandwidth Internet connectivity has resulted in an ever-increasing amount of audio-visual material at our fingertips. Companies archive past conference calls and meetings, and universities often make recordings of lectures available online. An astounding 48 hours of video is uploaded to YouTube each minute [43]. Not surprisingly, the availability of methods to navigate speech has become essential. Given the quantity of speech data available, it is impossible and unnecessary to view all relevant speech in its entirety. Instead, demand exists for navigation techniques which can be used to help users effectively and efficiently extract the information that is relevant to achieving their desired goals.

The solution we choose to focus on is speech summarization [41]. Summarization maintains a representation of an entire spoken document, focusing on those utterances (sentence-like units) that are most important and therefore does not require the user to process everything that has been said. Current research focuses on extractive summarization where a selection of utterances is chosen from the original spoken document in order

to make up a summary. Extractive summaries can be generated using any raw audio source without the need for human transcription by relying automatic speech recognition (ASR) or simply using acoustic features. Extractive summaries can be displayed visually as text, or also audially. The latter may be most effective in situations with a relatively high ASR word error rate (WER), allowing users to listen to the original audio where reading an error-laden transcript may be unfavourable. WER is a method of measuring, at the word level, the difference between the ASR-generated text and what was actually said.

As potential for the use of speech summarization increases, this area has continued to see a strong level of research interest. Unfortunately, automatically-generated summaries often fall short of delivering the information needed by users [29]. This is due, in no small part, to the fact that the natural language heuristics used to generate summaries are often optimized with respect to currently-used evaluation measures. Evaluation is the means by which summary quality, and therefore the effectiveness of a given summarization algorithm, can be measured. If care is not taken in how evaluation is conducted, we may end up optimizing summarizers based on data that cannot accurately guide the development of algorithms to actually result in real improvement.

What is needed, is to place evaluation into an ecologically valid paradigm. Ecological validity is defined by Cohen [7] as "the ability of experiments to tell us how real people operate in the real world". An ecologically valid experiment immerses participants in an environment that either mimics reality, or somehow provokes behaviour that would have resulted from such an environment. If constructed properly, these experimental conditions will result in participants naturally acting how they would in the real world.

This thesis discusses a human evaluation in the university lecture domain. Summarization of undergraduate lecture material is conducted by participants placed under ecologically valid conditions. Similarly, evaluation is achieved through tasks analogous to those completed by students who would make use of summaries in the real world.

By doing this, we address the two fundamental flaws with current evaluation of speech summarization.

The first flaw with current methods of evaluation is the use of subjective gold standard summaries. Such metrics simply score automatically-generated summaries against subjectively-classified gold standards without taking into account the usefulness of a summary in assisting a user achieve a certain goal or even overall summary coherence. By putting participants under ecologically valid conditions, we can allow for the creation of summaries in a way that would happen naturally for people exposed to material in the lecture domain.

The second flaw is with the way in which summaries are evaluated with respect to these gold standards. Methods used often compare a system-generated summary to one or many gold standard summaries by matching sequences of words or phrases in various ways. In order to validate whether or not such a comparison actually results in an accurate measure of summary quality, an ecologically valid evaluation must be pursued.

Using current evaluation methods, which make use of subjective gold standards and evaluation that is not ecologically valid, even the most linguistically-complex summarization systems perform no better than basic heuristics, such as picking the longest sentences from a general-topic, spontaneous dialog, or the first few sentences from a news recording [35]. In addition, Penn and Zhu [35] show that current evaluation cannot distinguish state-of-the-art full-featured summarizers from MMR. With results like these, it is unclear that the speech summarization community can confidently move forward in developing more effective summarization systems. To address this we have conducted an ecologically valid evaluation where participants create and make use of summaries under conditions that can be expected in real-world tasks in the university lecture domain. It is only through studies like this, that the actual performance of summaries can be accurately measured. By legitimately casting doubt on conventional methods of evaluation, we are making a step in the right direction towards replacing subjective gold

standards and ecologically invalid measures with something that can accurately guide the development of future systems.

1.1 Contributions

This thesis makes the following research contributions to the field. We provide a brief review of speech summarization and the current evaluation techniques used. Most importantly, we design and conduct an ecologically valid experiment able to evaluate spontaneous speech. By doing so, we describe a method that can overcome the use of subjective gold standards as well as evaluation measures that are in no way representative of tasks that users may want to use summaries for. Finally, our results are used to examine the correlation between ecologically valid results and more conventional evaluation, and determine to what extent we can rely on these methods.

1.2 Outline

In this thesis, we begin by examining the topic of speech summarization. We then look at summarization techniques, as well as domains summarization can be applied to. We also review the current evaluation techniques used for speech summarization. Next, we lay out our experimental protocol, explaining the structure of our experiment and the hypotheses. Finally, results are presented along with analysis and discussion. We conclude with a brief look at future work, including a follow-up study.

Chapter 2

Background and Related Work

2.1 Speech Summarization

In the past, various methods have been proposed to navigate spoken documents, including time compression and playing two separate portions of audio simultaneously, one in each ear [5, 37]. Although some success has been reported in speeding up audio by up to two times, Tucker and Whittaker [41] have shown that subjective Likert [17] evaluation indicates that users found this to be too fast. We choose to focus instead on speech summarization. Summarization maintains the ability to review a representation of an entire spoken document but focuses on those utterances (sentence-like units) that are most important and therefore does not require the user to process everything that has been said.

2.1.1 Extractive versus abstractive summarization

Current research focuses primarily on extractive summarization where a selection of utterances is chosen from the original spoken document in order to make up a summary. This is in contrast to abstractive summarization where new utterances are generated. Extractive speech summarization is currently more feasible and as such makes up the

majority of current speech summarization research. In fact, it has be shown that extractive speech summarization is quite effective for tasks requiring the use of spontaneous speech archives, and qualitative evaluation indicates extractive summaries are intuitive and efficient [29].

2.1.2 Query-based summarization

Most summarization is generic in the sense that summaries are meant to extract the most important or salient content from the original document. However, work has also been done on query-based summarization, where the content of a summary is determined by queries or the general needs of the users of the summaries. Hsueh and Moore [15] conducted a user study showing that compressing summaries of meeting recordings in order to only include information relative to the user's task resulted in improved performance in a decision debriefing task. Zhu et al. [48] used hypertext that is hyperlinked to a news report as a query into the linked document. The assumption here is that the user wants to find content relevant to the hypertext description. Portions of our study use primed summaries, where human summarizers take into account a set of questions that the summary should help answer. The result are examples of query-based summaries.

2.1.3 Summary presentation

Extractive summaries can be displayed visually as text, or also audially. The latter may be most effective in situations with relatively high ASR word error rates, allowing users to listen to the original audio where reading an error laden transcript would be unfavourable. Extractive summaries can use the original audio, which is arguably more natural than the text-to-speech audio that an abstractive method would likely have to rely on. This is a case where extractive summarization may be preferable to abstractive summarization.

2.1.4 Extractive speech summarization approaches

Extractive summarization involves extracting portions of the original content for inclusion in a summary. Although there has been some work done on using words as the extraction unit, most research involves extracting larger units, called utterances. Utterances are sentence-like units that can be determined by pause-length, sentence breaks, or dialogue acts in multi-speaker documents. If not otherwise specified, all techniques covered in this section pertain to utterance level extractive summarization. Maskey et al. [25] show that intonational phrases make the best extraction unit, outperforming sentences and two pause-based segmentations. The authors note that, in the corpus used, intonational phrases tended to be shorter than sentences, but held more semantic meaning.

We now discuss the various features that can be used to determine which utterances in the original spoken document should be included in an extractive summary. We split our discussion into two parts based on the type of feature; transcript features and acoustic features. It is important to note that multiple features can be combined into a summarization system from one or many of these categories.

Transcript features

Extractive speech summarization can be thought of as an extension of text summarization. Under this paradigm, the speech is first transcribed using ASR, and then summarization is performed on the resulting transcript. Although ASR is still far from perfect, transcript summarization is still effective with transcripts containing errors. In addition, it is expected that ASR technology will continue to improve. After being generated using ASR, the transcript is then processed appropriately, using some or all of the techniques mentioned below. Zechner and Waibel [44, 45] present a typical summarization framework (Figure 2.1). In each stage, various features of the transcript are taken advantage of to either modify the transcript or select appropriate utterances.

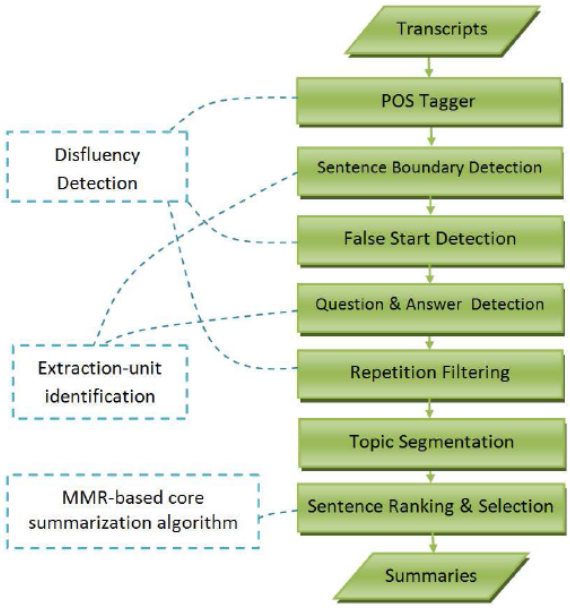

Figure 2.1: Transcript-based summarization

Disfluency processing Many advanced systems perform disfluency processing steps. Disfluencies, such as false starts, hesitations, repetitions, filled pauses, and interruptions, are common in spontaneous speech, representing 15-25% of the total words spoken in a document [44]. This is a problem not present in text summarization, or even summarization of non-spontaneous speech, such as broadcast news. Disfluency detection is an important component of Zechner's [44] summarization system for multi-party conversations in unrestricted domains. Repetitions of word sequences one to four words long are identified by a repetition detection algorithm. Repetitions of longer than four words are rare in natural speech, and as such, are not dealt with. False starts are incomplete clauses. These can occur for a variety of reasons, including in the event that the current speaker is interrupted by someone else. A C4.5 decision tree trained on SWITCHBOARD Penn Treebank was used to recognize false starts. A part-of-speech tagger is trained and used to identify filled pauses, which are broken down into either non-lexicalized filled pauses (e.g. *uh*, *um*) or lexicalized filled pauses (e.g. *like*, *you know*). Other advanced algorithms also use similar methods for disfluency removal as Zechner [35]. For less

spontaneous domains, such as broadcast news, disfluency processing is not necessary.

Extraction unit detection Summarization systems that define utterances based on pause length can easily determine how to detect utterance boundaries. Zechner [44] uses sentence boundaries as well as question-answer pairs to determine extraction units. A decision tree is used as a classifier and trained using part-of-speech tags and trigger-word status as features. The four words before and after a boundary are used. Similarly, a decision tree is trained in conjunction with a speech-act-tagging method to detect question-answer pairs. The purpose behind this is to ensure that questions and their corresponding answers are kept in a single unit, to increase coherence. This is applicable to domains containing multi-speaker conversation.

Core summarization The core summarization module of many summarization systems uses the popular maximal marginal relevance (MMR) algorithm [35, 44]. MMR was first proposed by Carbonell and Goldstein [3] as a means to incorporate the notion of "novelty" into relevance-centric approaches in information retrieval. The marginal relevance of an utterance is a linear combination of how relevant an utterance is to the document as well as how novel it is, that is, how dissimilar it is from previous utterances already in the summary. The second term is referred to as redundancy, as an utterance that is similar to one already in the summary is, to some extent, redundant.

MMR is a vector-space model where relevance and redundancy are often both computed using cosine similarity [30]. For a document D, the relevance of an utterance, U_i, is $Sim(U_i, D)$, where Sim is the cosine similarity or some other similarity metric. D is the average document vector. Similarly, if $Summ$ is the average of the vectors of the utterances already added to the summary, then the redundancy term can be written as $Sim(U_i, Summ)$. Note that Sim for the relevance and redundancy terms can potentially be different metrics. Given a weight, λ, the MMR score of a given utterance, U_i, is:

$$Score_{MMR}(i) = \lambda(Sim(U_i, D)) - (1 - \lambda)(Sim(U_i, Summ)) \tag{2.1}$$

Penn and Zhu [35] point out that while MMR has recently been set aside in pursuit of more advanced features, most of these new techniques perform no better than simply measuring the length of utterances. Furthermore, Murray et al. [30] show that MMR outperforms feature-based approaches in the meeting domain. In this work, the best performing approach, latent semantic analysis (LSA), performs no better than MMR, but is not as suitable for query-based and multiple-document summarization, which the authors believe is a disadvantage. MMR can easily be adapted to query-based summarization by replacing the general relevance term with a metric that measures how relevant or related an utterance is to a given query. To do this we can replace $Sim(U_i, D)$ with $Sim(U_i, Q)$.

In the lecture domain, Penn and Zhu [35] show that, using current evaluation metrics, MMR does not perform significantly worse than summarizers making use of every conceivable feature available. An ecologically valid evaluation, such as the one described in this thesis, is necessary to confirm whether this result is due to the actual merit of MMR or, instead, to flawed evaluation measures. To set the stage for accomplishing this goal, we choose MMR as the automatic summarization algorithm in our evaluation. As mentioned above, MMR is suitable for query-based and multiple-document summarization, making it an ideal approach to build upon. Furthermore, it is unsupervised and does not require labeled training data.

Gurevych and Strube [11] instead use a shallow knowledge-based approach to select in-summary utterances. WordNet was used as a knowledge source to conceptually represent the appropriate word sense of the nouns of both the individual utterances and the entire document. The semantic distance between each utterance and the document is defined as the average pairwise distance between all concepts contained in each. The summary is then iteratively created by taking the most similar utterance that has not yet been added to the summary. This approach depends on word sense disambiguation which was manually labeled. However, Zhu and Penn [49] show that simply choosing the dominant sense of each noun underperforms TF*IDF-based MMR. This shallow knowledge-based

approach was originally applied to single document summarization, which conceivably has less redundancy, perhaps explaining why a redundancy component was not thought to be needed by Gurevych and Strube.

Acoustic features

The second approach involves making use of acoustic features from the audio of the original document. This can be especially useful in situations where ASR word error rates are high. Although we discuss transcript features and acoustic features separately, they can be combined in various ways. For example, transcript-based summarization can be augmented by taking into account prosodic features of utterances.

Transcribing speech using ASR requires human annotated data for training. To perform well in new environments or for adaptation to a specific speaker, this is an expense, and may not always be feasible. In addition, ASR error is increased by the presence of out-of-vocabulary words. Munteanu et al. [27] show that the WER threshold for useful ASR transcripts is 25%. While this is a relatively high value, there are situations where we don't need transcripts, even if they can be made available, such as when the resulting summary is expected to be listened to instead of read.

Prosody-based features Prosody is the aspect of speech that represents how the given speech is presented, irrelevant of the content. Examples of prosodic features include pitch, speaking rate, loudness, intonation, stress, and rhythm.

Chen and Withgott [4] use a syllable-level hidden Markov model to identify emphasized speech. The purpose of the model is to determine excerpts of at least 15 seconds in duration for use in a summary of spoken conversation. The hidden Markov model framework was trained using data where emphasis was ranked on a three-point scale, or labeled as non-emphasized, by human annotators. Chen and Withgott [4] originally extracted emphasized words, but discovered that these words did not correspond to content

words, resulting in poor summaries. However, the use of 15 second phrases containing large amounts of emphatic speech showed improvement. The results of their model show no significant difference from a gold standard, suggesting that this method works well.

Maskey and Hirschberg [24] use a sentence-level hidden Markov model to extract sentences based on prosodic features for inclusion in a summary, in the broadcast news domain. The features used are speaking rate; F0 minimum, maximum, and mean; F0 range and slope; minimum, maximum, and mean energy; energy slope; and sentence duration. This model was shown to perform better than a random baseline. However, there is no comparison to a positional baseline, which has been known to perform very well for broadcast news. In addition, the authors include sentence duration as a prosodic feature. Sentence duration conceivably correlates with sentence length. As sentence length is known to be effective in broadcast news, it may be difficult to determine the isolated effectiveness of the more conventional prosodic features [6]. Maskey and Hirschberg [23] also compare these prosodic features with lexical, structural, and discourse features for use in summarization, concluding that the best summaries are achieved by using a combination of feature types.

Acoustic-pattern features In contrast to prosody-based models, acoustic-pattern-based models focus on the actual lexical content of the spoken document. Current summarization techniques, such as MMR, do not actually require understanding of the speech, but simply involve comparing identical words. The idea behind acoustic-pattern-based models is to be able to find identical words by finding similar portions of the speech signal itself, instead of first transcribing it to text.

Zhu et al. [51] propose a method that uses a modified version of Park and Glass's [34] segmental dynamic time warping (SDTW) algorithm. This algorithm itself is a segmental variant of dynamic time warping. SDTW is an unsupervised method that finds sequences of speech data containing similar patterns which correspond to lexical units such as words.

Zhu et al. use this modified version of SDTW to find acoustic patterns that occur across utterances. These patterns are expected to correspond to words contained in more than one utterance. Using these patterns, it is possible to determine the relatedness of any pair of utterances. This is captured in a relatedness matrix. Finally, the authors use this relatedness matrix to choose utterances which maximize relevance and minimize redundancy; essentially performing MMR. Empirical results show that performance of this model is as good as using ASR with a WER of 33% to 37%. The authors note that this is actually quite promising, especially for situations where ASR is not effective, including domains containing a high amount of out-of-vocabulary words. Zhu et al.'s [51] is an example of a model that uses both prosodic-based and acoustic-pattern features.

2.2 Speech summarization domains

Speech summarization research has typically been conducted on four categories of domains; broadcast news, meeting recordings, telephone conversations, and presentations. While each presents a somewhat open domain, some are more suited to future practical application than others.

2.2.1 Broadcast news

Most speech summarization research has been conducted on broadcast news. This is a highly structured domain, where speech is read off of well-formed transcripts, resembling written text more than conversational speech. Furthermore, in broadcast news, nearly perfect studio acoustic conditions and professionally trained readers results in low ASR WER making it an easy domain to summarize. However, a positional baseline performs very well in summarizing broadcast news [6]. That is, those utterances occuring near the beginning of a broadcast are often more suitable for inclusion in a summary than those occuring later, meaning that simply taking the first N utterances provides a very

challenging baseline, questioning the value of summarizing this domain. In addition, the widespread availability of written sources on the same topics means that there is not a strong use case for speech summarization over simply summarizing the equivalent textual articles on which the news broadcasts were based. It is, however, interesting to note that unlike written news, the first several utterances serve to capture listeners' attention [6]. Although one of the major reasons for working with broadcast news is its low ASR WER, there has been work that aims to eliminate the need for ASR completely [24, 51].

Most real world applications for speech summarization will conceivably involve unplanned, spontaneous speech. Zhu, et al. [50] point out several important characteristics of spontaneous speech for speech summarizers, including that it is often not linguistically well-formed, and contains disfluencies and false starts. Additionally, spontaneous speech is more vulnerable to ASR error, resulting in higher WER. Although broadcast news may contain spontaneous speech in the form of interviews or informal reports, spoken news reports are for the most part very far from spontaneous. This raises the question of whether research done in this domain is relevant to other, arguably more applicable domains.

2.2.2 Meeting recordings

Meetings are held by all types of organizations and are often the medium used to make important decisions. Whether held physically, or via the Internet or phone, meetings can be recorded and summarized using speech summarization techniques. Meeting summaries are useful as both a memory aid for the meeting attendees themselves, and as an information source for individuals who were not present at the original meeting. They can also be used to bring participants who are joining a meeting after it has already commenced, up to speed. This domain is more applicable than broadcast news, but at the same time contains additional challenges including lower quality recording equipment and spontaneous conversation, both resulting in a lower ASR WER. In addition, multiple

speakers add another layer of complexity to the problem.

A prominent example of work in this domain is a study by Murray et al. [28] who conducted a decision audit task where participants were asked to find information by navigating various meeting summaries. The study shows that summarization is a useful tool even in domains with higher ASR error, such as meetings. In addition, qualitative results suggest that extractive summarization is an intuitive method for navigation in this domain.

2.2.3 Telephone conversations

Telephone conversations can easily be recorded from call centers. Perhaps the future will bring the ability to record personal calls in order to retain a searchable and navigable record of past conversations. Alternatively, such as in the case of the widely-used SWITCHBOARD corpus, participants can be paid to generate artificial conversations. As telephone conversations are an example of spontaneous conversation, this domain has been useful for developing methods to deal with disfluencies, a phenomenon prevalent in spontaneous speech. For example, Zechner [44] has worked on methods to detect and remove three types of disfluencies; filled pauses, false starts, and repetitions.

2.2.4 Presentations

The presentation domain is unique in the sense that it is structured, but still consists of spontaneous speech. Unlike broadcast news, presenters are not necessarily professionally voice-trained and do not read from a script. Although recording conditions are often better than in meetings, they are not at the same level as broadcast news. Furthermore, although many presentations may be based off of written sources, such as textbooks or reports, the lecturers often deviate from these sources, meaning that there is interesting content to summarize that could not have been extracted simply from the non-speech sources themselves, unlike in the case of broadcast news. Furthermore, in this domain

the positional baseline performs very poorly. Presentations often include props, such as slides. He et al. [12] have experimented with using slide transitions as a method of summarization for corporate technical presentations. Other sources of presentation data include university lectures and lecture talks from broadcast television [16].

2.3 Lecture webcasting

Availability of high bandwidth Internet connectivity has resulted in an increasing amount of audio-visual content available online. Some of this content takes the form of lectures. All sorts of lecture material is available on sites like YouTube, which itself sees more than 48 hours of video uploaded each minute [43]. More structured and focused talks are available on sites such as TED [39]. Further down this path, it is increasingly commonplace for university lectures and academic presentations to be broadcasted online in the form of webcasts [1]. Various webcasting systems have been developed for research purposes to tackle some of the issues involved with lecture webcasting, including systems developed by Microsoft Research [47], MIT [10], University of Toronto [1], and University of California [38]. Lecture webcast interfaces provide basic audio/video navigation as well as optional features such as a table of contents or access to slides. University lecture webcasts have been shown to positively affect students. Brotherton and Abowd [2] show that access to lectures online encourages review activities. Furthermore, webcasted lectures did not result in lower class attendance.

As pointed out by Munteanu [26], current webcasting interfaces are lacking in several respects. For example, transcripts of the spoken content are rarely provided. This is particularly an issue for lectures, as actual live lectures are significantly more interactive when compared to the traditional equivalents of other webcasted content, such as television shows. User studies conducted by Dufour et al. [8] show that text, whether transcripts or an abstract, is a useful tool for performing information tasks using web-

casts. Munteanu et al. [27] have conducted user studies to determine to what extent ASR transcripts can be used, showing that transcripts with a WER of up to 25% were still useful. Munteanu [26] also notes that complimenting webcasted lectures with transcripts allows for more traditional text-based tools and techniques to be used, including searching, indexing, and summarization.

2.4 Evaluation of speech summarization

Evaluation is perhaps the most important active research area in speech summarization. In order to determine whether a given method of summarization is effective or to compare two different methods, we must have some means of assessing the quality of the produced summaries. The best way to evaluate how well a summarization algorithm performs is to look at the summaries produced by the algorithm. Evaluation can be broken down into two categories; intrinsic and extrinsic. Intrinsic evaluation looks at the contents of the summary itself whereas extrinsic evaluation uses a criterion external to the summary to evaluate it; often by embedding the summary in a task. In addition, subjective surveys are often used to gauge non-objective aspects of summaries, such as coherence or fluency in intrinsic evaluation, or how intuitive or enjoyable using a particular type of summary was in an extrinsic evaluation. A common method for gathering this data is using questionnaires consisting of Likert scale [17] questions, where participants are asked to rate their level of agreement with various statements.

2.4.1 Intrinsic evaluation

Intrinsic evaluation measures the quality of a given summary by examining the contents of the summary itself. Using various measures, it is possible to assign the summary a score representing how "good" the summary is. While intrinsic evaluation is often more time- and cost-effective than extrinsic evaluation, there is a certain level of arbitrariness

about what the actual score means, and whether or not a higher scoring summary will actually be more useful in a real world application.

Evaluating summaries intrinsically still requires human effort. If we evaluate each summary independently, this effort is linear in the number of summaries. However, most recent approaches decouple human effort by having human annotators first create their own gold standard summaries, or alternatively rank each utterance in the original document. Following this, all newly generated automatic summaries can simply be compared to these gold standards automatically using predefined metrics. Most methods of intrinsic evaluation are borrowed from text summarization, although there have been attempts to take an ASR-like approach by measuring summary accuracy at the word level. We discuss some of the most common approaches below.

F-Measure

F-measure is an evaluation metric that balances the effects of *precision* and *recall*. Note that in this case, we are evaluating a summary created by the system, that is, a set of utterances selected by the system. The human summary, or the set of utterances selected by the human, is the gold standard.

$$precision = \frac{|\text{utterances selected by both system and human}|}{|\text{utterances selected by system}|} \tag{2.2}$$

$$recall = \frac{|\text{utterances selected by both system and human}|}{|\text{utterances selected by human}|} \tag{2.3}$$

F-measure is defined as below. Precision and recall are most often treated equally. When this is the case, the parameter β is set to 1.

$$\text{F-measure} = \frac{(\beta + 1) * precision * recall}{\beta * precision + recall} \tag{2.4}$$

Nenkova [32] describes three major issues with using a precision/recall evaluation metric such as F-measure.

Human variation: Different human annotators tend to choose different summaries as being appropriate to include in a summary. This means that depending on which annotator a given automatic summary is being compared to, a vastly different score may be assigned. This is reason to consider placing more weight on recall rather than precision.

Granularity: The heart of this issue is that utterances vary in length. Since any precision/recall metric gives an equal weighting to each in-summary utterance, we can have two summaries that are assigned an equal score, but where the utterances in one are significantly longer than the other and therefore contain more information.

Semantic equivalence: Multiple utterances may convey the same meaning. However, if a summary contains an utterance that is different, but contains equivalent semantic content to that in the human annotated summary, it will be penalized instead of rewarded.

ROC curves

Receiver operating characteristic (ROC) curves have also been commonly used for intrinsic evaluation of speech summaries [6]. ROC curves are especially useful for comparing the effectiveness of various features, allowing several summary types to be plotted as separate curves on the same graph. An ROC curve is plotted in two dimensions with the x-axis representing the false-positive rate, or $1 - specificity$, and the y-axis representing true-positive rate, or *sensitivity*. ROC curves are essentially a variation of precision and recall and as such suffer from many of the same issues identified for the F-measure metric [32].

$$1 - specificity = \frac{|\text{utterances selected by system but not human}|}{|\text{utterances not selected by human}|} \tag{2.5}$$

$$sensitivity = \frac{|\text{utterances selected by both system and human}|}{|\text{utterances selected by human}|} \tag{2.6}$$

Relative utility

Relative utility was proposed in order to address the issues of human variation and semantic equivalence common to precision/recall metrics [36]. In this technique, multiple human annotators are asked to score all utterances on a uniform utility scale. The higher the ranking, or utility, the more likely the utterance should be included in a summary. An automatically generated summary is assigned a utility calculated based on the scores assigned to the summary's utterances by the human annotators. The overall score of the summary is the summary's utility score relative to the maximum achievable utility.

ROUGE

ROUGE is a recall-oriented evaluation metric, first described by Lin and Hovy [19]. It has become a standard in text summarization and has also been applied to speech summarization. ROUGE measures the number of overlapping units between an automatic summary and a set of reference, or gold standard, summaries. Being recall-oriented, ROUGE addresses the human variation criticism of precision/recall metrics. In addition, ROUGE makes use of multiple reference summaries. The other criticisms are also partially solved by the fact that ROUGE is able to look at units smaller than utterances. Semantically similar utterances will likely share certain words or phrases, and will thus receive credit from ROUGE. By not looking at utterances as a whole, utterance length is also less important. When the units being compared are n-grams, sequences of n words, Lin [18] defines the ROUGE-N score of a candidate summary as below, where $gram_n$ is an n-gram of length n, and $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in the automatic summary and the set of reference summaries.

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \tag{2.7}$$

In addition to ROUGE-N which computes overlapping n-grams, ROUGE can be calculated using several other variations [18].

ROUGE-L computes the longest common subsequence between two summaries.

ROUGE-W is a weighted version of ROUGE-L which gives higher weights to consecutive common subsequences.

ROUGE-S computes skipped n-grams.

ROUGE-Sx is a variation of ROUGE-S where each skip-bigram is no more than x words apart.

ROUGE-SU is a version of ROUGE-S where unigram matching is also considered.

ROUGE has gained considerable credibility in the text summarization community owing to the fact that ROUGE scores tend to correlate well with human evaluation [18]. Zhu and Penn [49] also show that ROUGE scores agree with those computed using Relative Utility and F-measure. However, Murray et al. [31] demonstrate that ROUGE may not be as suitable for speech summarization, where human evaluation disagreed with ROUGE scores on whether MMR-generated summaries were better than those produced using an LSA approach. Liu and Liu [20, 21] also show that in the multiparty meeting domain, ROUGE scores do not correlate well with human evaluation. Human evaluation consisted of ranking statements about the summaries on a five-point scale. However, they noted that removing disfluencies improved correlation. Although the method of human evaluation in these studies may be questionable, this data nevertheless further questions the validity of using ROUGE, particularly in speech summarization.

ASR-inspired methods

Some authors [14] have relied on variations of Word Error Rate where summaries are compared word by word, taking into account the number of insertions, substitutions and deletions. These methods essentially calculate how many word-level errors an automatically generated summary contains when compared to a gold standard summary. These approaches are largely based on commonly used measures in ASR. Zechner and Waibel [46] compute a Summary Accuracy score where for each speaker turn, human annotators

scored phrase-level segments. For each word, a relevance score was computed based on the average of the phrase-level scores. From there, the Summary Accuracy is calculated as the sum of the word level scores relative to the maximum achievable relevance score of the summary. Sharing some similarities to Zechner's method, there has been an attempt by Galley [9] to use a variation of the Pyramid method [33] that has gained popularity in text summarization. Instead of n-grams which used in ROUGE, the pyramid method looks for various clauses and phrases identified by annotators, known as summary content units (SCUs), repeated in different model summaries [33]. Galley [9] performs Pyramid using words and their location in a document instead of SCUs, resulting in a measure similar to Summary Accuracy.

Zhu and Penn [49] conducted an experiment comparing various intrinsic evaluation techniques. Their results show significant differences in summary rankings evaluated using ASR-inspired methods and more traditional methods borrowed from text summarization such as ROUGE and F-measure. In light of these results, they advise caution in relying on ASR-inspired evaluation metrics.

2.4.2 Extrinsic evaluation

Extrinsic summarization does not evaluate a summary based on its content, but rather, determines how useful a summary is in the context of an external criterion. If the summary being evaluated is embedded in a task, its value is measured by how well the task is performed, or alternatively, the time taken to complete the task. Based on Cohen's [7] definition that ecological validity is "the ability of experiments to tell us how real people operate in the real world", extrinsic evaluation is capable of being ecologically valid if an appropriate task or measure is employed, representing how the summary might be used in the real world. The same cannot be said for intrinsic evaluation. Zhu and Penn's [49] comparison of intrinsic evaluation metrics confirms that using different metrics results in significant differences in rankings of summaries. There is no way of confirming which of

the intrinsic measures produces the "correct" scores without performing a task-embedded evaluation. However, extrinsic evaluation normally requires a task to be repeated for each type of summarization system being evaluated. In addition, even for a single system, a large number of human subjects are needed to be able to say something conclusively. This expense is the major reason why intrinsic evaluation is still the dominant method used. In addition, extrinsic evaluation is sensitive to experimental setup, meaning that if the task is not well designed, then the results may not be applicable to other tasks or domains. We now discuss some of the more prominent extrinsic evaluations that have been performed on speech summarization and navigation.

Evaluation of audio-video presentation summaries

He et al. [12] performed an extrinsic evaluation in the presentation domain. Presentations and corresponding powerpoint slides were used to generate automatic video summaries which 24 participants used in a fact-finding task. Each participant completed a quiz before viewing the summaries in order to gauge background knowledge. Each participant was then permitted to view a summary on a fully functional interface for as long as he or she desired. Automatic summaries were generated using a simple slide-transition approach, a pitch-based method, and a method combining the two techniques as well as incorporating data about which portions were accessed by previous users. Following this, access to the summary was removed and the participant filled out a final quiz which was used to evaluate the usefulness of the summary. Participants indicated that they preferred the author-generated manual summaries to any of the automatic ones, which showed no significant difference from each other. Quiz scores also reflected this. It is also worth noting that all final quiz scores outperformed the initial background quiz, indicating that all summaries were useful.

A follow-up study was also performed in the same domain which compared the amount of information users acquired from various sources, including slides, audio-video, and

transcripts [13]. Results show that although users prefer audio-video, author-generated transcripts containing highlights are just as effective. The value of slides varies depending on the type of content the slides consist of. Slides containing more detailed content are more effective than those containing only main points. In this study, slides are treated as a type of summary, which they essentially are. These results put pressure on the value of summarization under the condition where good slides are already available, such as is often the case in the university lecture domain.

Evaluation of speech compression

Extrinsic evaluation has also been used to measure speech compression techniques. Tucker and Whittaker [41, 42] exposed participants to speech compressed using various techniques. Following this, the participants were asked to rate utterances from the original spoken content based on importance. Alternatively, participants were also asked to use a browser to listen to extracts before rating utterances. The participants' ratings were compared to gold standard ratings generated by annotators without time constraints. The various compression techniques as well as extracts were evaluated based on how well the corresponding participant-generated ratings matched those of the gold standard. The results suggest that extracts outperformed other compression techniques, showing support for extractive summarization.

Decision audit task

Murray et al. [28] have used the meeting domain to perform a study involving 50 particiapnts who were asked to perform a decision audit task. Such a task required a general understanding of the spoken documents and required the gathering of information that could not easily have been found by simply performing a search on a transcript. The task involved finding decisions made during a set of meetings as well as alternative proposals discussed and the reasons supporting the final decision as well as the alternative

proposals.

Each participant was given a fixed amount of time to use an interface to create a summary-report, which addressed the question of what the final decision was, what alternatives were proposed, and supporting reasons for all options. This task was conducted using five conditions, with ten participants completing the task for each condition. Each condition determined what additional summary material the subject had available to him or her while creating his or her summary-report. Conditions included combinations of human-generated and automatic summaries, both extractive and abstractive. Subjective evaluation of the participants' report was conduction by a panel of judges, who examined the created summary-reports. This subjective evaluation also included a single objective question which required judges to count the number of key points included from a gold standard list, which itself was created subjectively by another set of judges. Evaluation also consisted of questionnaires as well as analysis of the browsing habits of participants. Although this is a task-based evaluation, the result of the task was a summary-report, which was mostly subjectively evaluated. Results indicate that automatic summaries are useful, and extractive summaries performed well. Not surprisingly, the best performance was achieved by manual abstractive summaries.

'Time travel' in meetings

Tucker et al. [40] developed a system that provides the gist of a meeting up to a certain point in time, allowing participants who join late to be quickly brought up to speed. The gist was established using a variation of a simple TF*IDF algorithm allowing the system to operate in real time. Using 41 participants the authors conducted an evaluation where late-comers to a meeting were either given the gist of the missed content or nothing at all. Factual questions on the content of the meeting were asked and used to evaluate the effectiveness of the provided gist. The results indicate that the system not only improves knowledge of the missed content, but also improves participants' knowledge of content

subsequent to the gist; that is, the portion of the meeting where they were present. The authors suggest that by providing a gist of the entire meeting, users can have the flexibility of skipping portions of the meeting with the option of effectively reviewing parts of the meeting that they were or were not present for. In this study a gist is simply a very brief summary.

Chapter 3

Experimental Design

3.1 Motivation

Current speech summarization research makes use of intrinsic evaluation metrics such as F-measure, Relative Utility and ROUGE, scoring summaries against subjectively-classified gold standard summaries obtained using human annotators. To compute these scores, annotators are often asked to rank utterances on a numerical scale, and in doing so commit to relative salience judgements with no attention to goal orientation and no requirement to synthesize the meanings of larger units of structure into a coherent message. The utterances in automatically generated summaries are then evaluated by either summing the annotator-assigned scores or by counting the number of annotator-selected utterances that appear in the automatic summary. Given this subjectivity, current intrinsic evaluation metrics are unable to properly judge which summaries are useful for real world applications. For example, using these intrinsic measures has failed to show that summaries created by algorithms based on complex linguistic and acoustic features are better than baseline summaries created by simply choosing the first utterances occurring in a spoken document or the longest utterances [35, 50]. Using this type of evaluation, MMR is also shown to perform just as well as these full-featured summariz-

ers [35]. What is needed is an ecologically valid evaluation metric that determines how valuable a summary is when embedded in a task, rather than how closely a summary matches the subjective utterance level scores assigned by annotators.

3.1.1 Spontaneous speech

Speech summarization has the most potential for domains consisting of spontaneous speech (e.g. lectures, meeting recordings). Spontaneous speech is often not linguistically well-formed, and contains disfluencies, such as false starts, filled pauses, and repetitions. Additionally, spontaneous speech is more vulnerable to ASR errors, resulting in higher WER. Unfortunately, these domains are not easy to evaluate compared to highly structured domains such as broadcast news. Furthermore, in broadcast news, nearly perfect studio acoustic conditions and professionally trained readers results in low ASR WER making it an easy domain to summarize. The result is that most research has been conducted in this domain. However, a positional baseline performs very well in summarizing broadcast news [6], meaning that simply taking the first N utterances provides a very challenging baseline, questioning the value of summarizing this domain. In addition, the widespread availability of written sources on the same topics means that there is not a strong use case for speech summarization over simply summarizing the equivalent textual articles on which the news broadcasts were based.

University lectures present a much more relevant domain, with less than ideal acoustic conditions and a structured presentation in which deviation from written sources (e.g., textbooks) is commonplace, and yet a positional baseline performs very poorly. The lecture domain also lends itself well to a task-based evaluation metric; namely university level quizzes or exams. This constitutes a real-world problem in a domain that is also representative of other spontaneous speech domains that can benefit from speech summarization.

3.1.2 Task-orientated evaluation

As pointed out by Penn and Zhu [35], current speech summarizers have been optimized to perform an utterance selection task that may not necessarily reflect how a summarizer is able to capture the goal orientation or purpose of the speech data. In our study, we follow the prevailing trend in HCI towards extrinsic summary evaluation, where the value of a summary is determined by how well the summary can be used to perform a specific task rather than comparing the content of a summary to an artificially created gold standard.

The university lecture domain is an example of a domain where summaries are an especially suitable tool for navigation. Simply performing a search will not result in the type of understanding required of students in their lectures. Lectures have topics, and there is a clear communicative goal. By using actual university lectures as well as university students representative of the users who would make use of a speech summarization system in this domain, any results obtained are ecologically valid.

3.2 Overview

Our study is a within-subject experiment where participants are provided with first year sociology university lectures on a lecture browser system installed on a desktop computer. For each lecture, the browser makes accessible the audio, manual transcripts, and an optional summary. Evaluation of a summary is based on how well the user of the summary is able to complete a quiz based on the content of the original material.

3.3 Research questions and hypotheses

The purpose of this research is to determine how various types of summaries perform when measured using an ecologically valid evaluation metric in a relevant, spontaneous speech domain. We entertained the following hypotheses:

3.3.1 No summary hypothesis

Using no summary will perform at least as well as using a summary under ecologically valid settings. Under this hypothesis, the *no summary* condition, where participants have access to the entire lecture content will perform at least as well as the conditions where participants have access only to a summary.

3.3.2 Primed summary hypothesis

Primed summaries do not lead to better performance on lecture quizzes. It is expected that summaries created manually by participants who have access to all of the questions in the evaluation quiz may result in better performance than those where participants had no access to the questions at the time of summarization. In the event of such an outcome, this hypothesis can be rejected.

3.3.3 Automatic summary hypothesis

Automatic summaries generated using MMR will not perform significantly better than manual summaries. It is not certain whether summaries automatically generated using MMR will perform worse, better, or the same as those manually created.

3.4 University lecture domain

The university lecture domain is better structured than meeting proceedings and has a far better ASR WER because the lecturer can be wired with a head-worn microphone and adapted to his or her own acoustic model, but it also contains speech that is more spontaneous than an anchored news broadcast and has higher error rates. The lecture domain also lends itself well to a task based evaluation metric, namely university level exams or quizzes. Furthermore, university lectures are a domain that can benefit from speech summarization. Increasingly, universities make lecture videos available online for

review purposes as well as for individuals who were not present at the lecture altogether. The target of these webcasts are not limited to university students, and there are methods for distributing lecture material to anyone who is interested in learning the content.

Sociology 101 was selected as the course from which to select lectures from. The course has no pre-requisites enabling us to gather participants from a variety of backgrounds. As a bonus, participants often commented that the lecture content was interesting; something that kept participants alert and engaged. Lectures from the course were recorded live, and four excerpts (referred to simply as lectures from here on in) of approximately 40 minutes each were selected for use in the study, each portion covering a different topic. In addition, one short segment from a fifth topic was selected for use in familiarizing participants with the lecture browser interface used in the study. The topics of the four lectures were chosen such that the content of each lecture was not dependent on the content from any other lecture, allowing for proper counterbalancing with respect to the order each participant used the set of lectures.

All lectures were transcribed manually. The resulting manual transcripts were used by the summarization algorithm as well as by the participants for creating manual extractive summaries. Transcripts were broken down into utterances, defined by 200 millisecond pauses in spoken content. The lectures consisted of an average of 983 utterances each.

3.4.1 Slides

As with many university courses, the Sociology 101 lecturer made use of slides throughout the lecture. These slides were captured directly from source and synchronized to the audio track. The slides were made available to participants in the lecture browser interface. There was an average of 22 slides per lecture, consisting mainly of text outlining key points.

In their user study, He et al. [13] show that users' preference for summaries decreases when slides containing detailed information are provided. Slides are a natural part of

most lectures and other presentations, and often serve as a type of summary themselves. Having good slides provides an additional challenge for extractive summaries. To make the experiment realistic, we included access to slides in all conditions. For a summary to perform well, it will have had to provide additional value over these slides.

3.5 Human subjects

Participants were recruited from the University of Toronto, St. George campus using posters advertising the study. These posters were displayed on bulletin boards in university buildings containing a high volume of undergraduate student traffic. An effort was made to distribute these posters widely in order to get a diverse set of participants that would be representative of typical undergraduates.

The study was open only to undergraduate students in 2nd, 3rd, and 4th year. The aim was to attract participants who were familiar with the format of university lectures and exams, but not participants who had progressed past the undergraduate level and would thus have a possible unfair advantage intellectually. Participants were required to have never taken a sociology, anthropology, social psychology, or university-level civics course. This was an effort to reduce the amount of previous knowledge of the lecture content. A total of 48 participants took part in the experiment; 25 female and 23 male. An additional 4 participants were involved in a modified experiment, completing a single condition for all four lectures. This data was used to normalize lectures for differences in difficulty and content. Each participant was compensated $70 for the entire experiment which consisted of a maximum six hour time commitment spread over three sessions.

3.6 Summaries

Our strategy for extractive speech summarization can be viewed as an utterance selection task. The original audio is broken down into a set of utterances which are identified by

200 millisecond pauses. The result are utterances that roughly correspond to natural sentences or phrases, although Liu and Xie [22] note a difference in summarization performance when using utterances defined in this manner as compared to sentences defined by human annotators. From this point forward, the task of summarization consists of choosing a subset of these utterances for inclusion in a summary. In line with current literature, we chose to set the target length of each summary at 20% of the words in the entire lecture. Although lectures were conducted by the same lecturer, using the same speaking style and rate, certain summarization methods may show a preference for longer or shorter utterances meaning that this 20% word length may not necessarily correspond to a summary containing 20% of utterances. Automatic summaries corresponding to approximately 20% of words can be created easily, by selecting the highest rated utterances until the summary length has been reached. However, a pilot study demonstrated that this is not as easily accomplished by human summarizers. Trying to hit the 20% mark with accuracy could be excessively time consuming as well as influence the choice of utterances. To avoid this, we allowed human summaries to range from 17% to 23% of words. There was an average of 983 utterances, or 6757 words per lecture.

3.7 Lecture browser interface

A lecture browser interface (Figure 3.1) was developed for use in this experiment. The interface allowed users to play university level sociology lectures. There is a pause feature as well as the ability to skip forward or backward through the lecture using an audio timeline, located at the center of the screen. The timeline also has markings to indicate slide transitions. Slides are visible at the top center of the interface. The manual transcript is displayed at the bottom center and is split and numbered by utterance. The utterances are highlighted according to the current utterance being spoken by the lecturer. Double clicking another utterance in the transcript causes the audio to jump either forward or

backward to that utterance. There is an option to toggle an autoscroll function, which causes the transcript to scroll automatically so that the currently playing utterance is always visible. Disabling this feature allows a participant to read portions of the lecture not directly adjacent to the currently playing audio. There is a table of contents on the left side of the interface. This displays the titles of the slides contained in the lecture. Clicking on an item in the table of contents causes the slide, audio, and transcript to jump to the corresponding location.

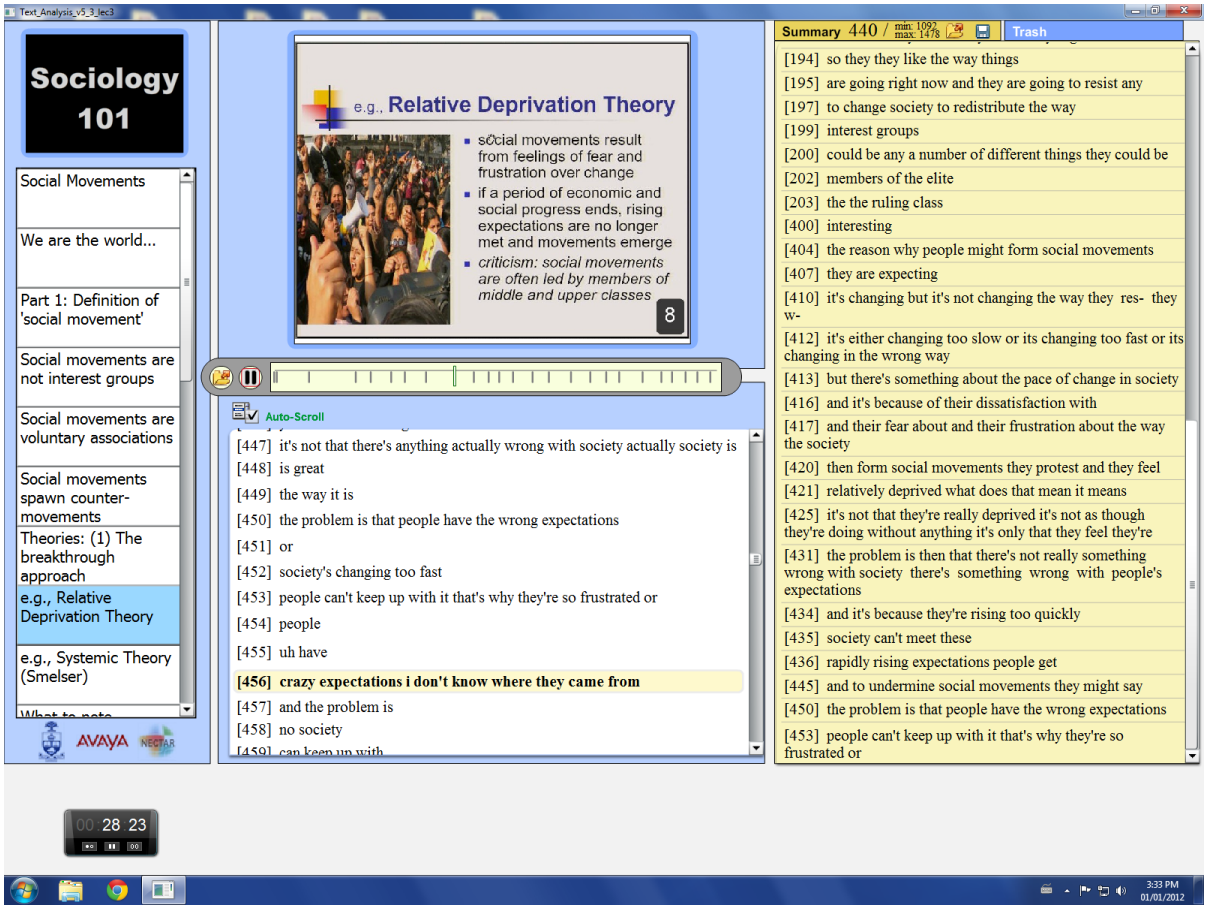

Figure 3.1: Lecture browser interface

The right portion of the interface is used for summary creation. Utterances can be dragged from the manual transcript into the summary pane. Dragging an utterance out of the summary pane results in it being removed from the summary and placed in the trash, which can be accessed by clicking a tab. Above the summary pane is a counter showing

the current number of words in the summary as well as the maximum and minimum number of utterances. Participants were allowed to overshoot the maximum summary length, provided they remove enough utterances to return to the required length range within the allotted time. A separate timer application was positioned at the bottom left of the display and used to time tasks.

3.8 Evaluation task

A teaching assistant for the sociology class from which the lectures were drawn was employed to generate quizzes for use in evaluation (Appendix A.4, A.7, A.10, A.13). These quizzes were designed to incorporate factual questions only, avoiding those questions that would require comprehension of the topic. This ensures that variation in participant intelligence has a minimal impact on our results. In addition, we ensured that questions involved information that was distributed equally throughout the lecture, but at the same time not linearly in the transcript or audio slider, which would have allowed participants to predict where the next answer might be located. Additionally, the order of questions was randomized. All questions were non-trivial to minimize the chance of the participant having previous knowledge of the answer. After answering the quiz, participants were asked to identify any questions for which they felt they already knew the answer to before the start of the experiment.

All questions were short answer or fill-in-the-blank. Each quiz contained four distinct types of questions. Question types do not appear in any particular order on the quiz and were not grouped together.

Type 1 These questions can be answered simply by looking at the slides. As such, these questions could be answered correctly with or without a summary. Each quiz contains three of these questions.

Type 2 Slides provide an indication of where the content required to answer the ques-

tions are located. Access to the corresponding utterances is still required to find the answer to the questions. Each quiz contains three of these questions.

Type 3 Answers to the questions can only be found in the transcript and audio. The slides provide no hint as to where the relevant content is located. Each quiz contains three of these questions.

Type 4 These questions are more complicated and require a certain level of topic comprehension. These questions often require connecting concepts from various portions of the lecture. These questions are more difficult and were included to minimize the chance that participants would already know the answer to questions without watching the lecture.

In this experiment, we simulated a scenario where someone has heard a lecture in the past, and may or may not remember the entire content while completing an open book exam or quiz. Quizzes are a relevant evaluation metric for this domain, as from a student's perspective, often the entire purpose of learning lecture content is to be able to perform well on evaluation tasks, like quizzes and exams. The higher level goal of learning the content is also ensured by a well designed quiz. That is, even in the real world academic context, instructors must design quizzes in such a way that performing well on a quiz correlates to a strong grasp of the lecture content. By using an actual teaching assistant for the class, we have achieved this. As such, a well designed quiz is an ecologically valid evaluation metric for summarization in this domain.

3.9 Procedure

Each participant was asked to attend three sessions. Each session took place in a private room, with an experimenter conducting the study. Participants were seated in front of a desktop computer with the lecture browser interface loaded. Participants had access to a mouse in order to use the interface as well as paper and pens. A keyboard was not

required, as there was no keyword search functionality in the interface. The experimenter was seated next to the participant and took qualitative notes on how the interface was used. These notes included the general strategy used by each participant in both phases, as well as the approximate time taken for various tasks. Every effort was made to ensure the participants were comfortable and at ease throughout the study. The study was broken into two main phases; the summarizing phase and the evaluation phase.

3.9.1 Summarizing phase

The summarizing phase of the experiment consisted of two sessions spaced one to seven days apart. During each session, the participant was asked to listen to two lectures and either create a summary using the interface or take notes using pen and paper, depending on the condition. The participant was permitted to take a break between the two lectures. For each 40-minute lecture, the participant was given a maximum of 60 minutes to listen to the lecture and either create a summary or take notes. The participant was free to interact with the interface in any way, including pausing or skipping. However, there was a strict requirement that each lecture must be listened to in its entirety during the time permitted. In the event that 60 minutes had elapsed and either the lecture had not been completely played, or the created summary was not within the required length limits, an additional 9 minutes, or 15% of time, was provided. Participants were not informed of this optional addition ahead of time, and only 7 participants required this extension.

At the beginning of the first session, participants were given an overview of the study and asked to sign a consent form (Appendix B.9 - B.10). Prior to listening to the first lecture, participants were instructed on the use of the interface and were asked to listen to a 6-minute warm-up lecture and create a primed summary (Appendix A.1). The purpose of this was to get the participants familiar with using the interface. The order of the remaining four lecture-condition pairs were counterbalanced using a Latin square.

3.9.2 Evaluation phase

The evaluation phase consisted of a third session scheduled a minimum of one week following the second session. This gave time for memory effects to stabilize. This final session began with a brief reminder of the functionality of the interface. The participant was then given a short 4-question quiz to complete in 4 minutes (Appendix A.2). A summary was provided based on the short warm-up lecture given at the beginning of the first session of the study. All slides were provided in the interface. The purpose of this quiz was to re-familiarize the participant with the interface as well as allow the participant to get used to using a summary. During this warm-up quiz and the remainder of the session, the participant had access to a modified version of the lecture browser interface (Figure 3.2) where the summarization pane was disabled. In addition, the manual transcript was replaced by a summary and the provided audio consisted only of the in-summary utterances concatenated together in sequence, according to the summary transcript.

Following the warm-up, the participant was asked to complete a quiz for one lecture for each of four conditions; a total of four quizzes. Twelve minutes were given for each quiz. Lectures and conditions were mapped and ordered using the same counterbalancing scheme used for that particular participant in the summarizing phase. Quizzes were designed as described above. During this time, the participant was able to browse the summary audio, all the slides, and manual transcript of the summary. During the *no summary* condition, participants were provided with the entire audio and transcript instead. As each original lecture was around forty minutes in length, there was a strong established time constraint that could make using a well-prepared summary beneficial to completing the task at hand. In fact, He et al. [13] show that providing users with a summary improved performance of a quiz-solving task compared to simply providing transcripts. An opinion survey also shows that there is a qualitative benefit to summarization in this context [13].

After each quiz, the participant was given a form (Appendix A.5, A.8, A.11, A.14)

and asked to check off any questions for which it was felt that he or she already knew the answer to before the start of the experiment. This was used to gauge an estimate of previous knowledge. Following this, the participant was asked to fill out a short questionnaire (Appendix B.1) which included questions on his or her experience and perceived performance on the quiz. At the very end of the study, the participant was asked to fill out a longer questionnaire (Appendix B.2 - B.8) eliciting basic demographic information, as well as qualitative questions about his or her experience during the study and suggestions for future applications.

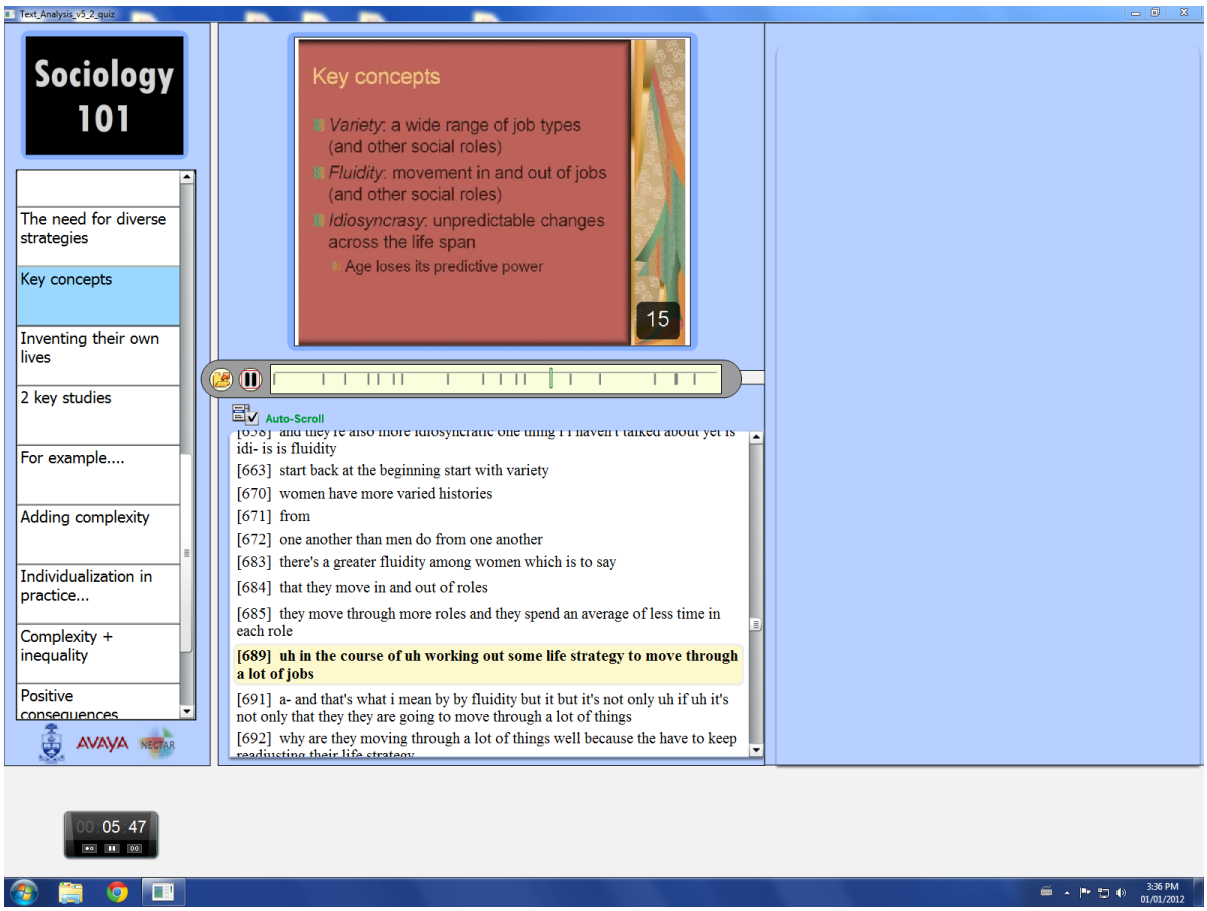

Figure 3.2: Modified lecture browser interface

3.9.3 Pilot study

A pilot study was conducted using five compensated participants in order to confirm appropriate times for each portion of the experiment as well as appropriate lengths for quizzes. Prior to this, a pre-pilot study was completed using five graduate student volunteers. The purpose of this was to perform trials on more basic aspects of the experiment, including features of the interface itself and the setup of the room and equipment. Piloting and pre-piloting was also useful for determining appropriate lengths for summaries. In the end, we decided on summaries consisting of a selection of utterances resulting in 20% of the words in the original document, a number standard in literature.

3.10 Conditions

Participants were asked to perform both the summarizing phase and the evaluation phase for all four lectures, one for each of four conditions. Each participant was subjected to all of the conditions. The lecture-condition combinations varied by participant, and the order of the four lectures were also rotated appropriately. A Latin square was used to create a counterbalancing scheme that resulted in 16 lecture-condition-order combinations. Each of these combinations was repeated three times, requiring a total of 48 participants. The four conditions were as follows:

A1: No summary

A2: Generic manual summary (17-23% of words)

A3: Primed manual summary (17-23% of words)

A4: Automatic MMR summary (20% of words)

3.10.1 A1: No summary

In this condition, no summary is used. During the summarizing phase of the experiment, participants were instructed to take notes using pen and paper. They were asked to take

notes in the same manner as they would in a real lecture where the course instructor had told them that these notes could be brought into an open book final exam. This is meant to resemble as closely as possible the existing real world scenario that we are applying summarization to. During the evaluation phase, participants had access to the entire audio, transcript, and slides. In addition, the previously created notes were available for use. While all lecture material was provided, the twelve-minute time constraint made it impossible to listen to the lecture in its entirety. If any of the other conditions are able to achieve or exceed the results of this condition, then we can confirm that summaries are a suitable replacement for lectures in a time-constrained situation.

3.10.2 A2: Generic manual summary

This condition made use of manual summaries consisting of 17-23% of the words in the original lecture. During the summarizing phase, participants were asked to listen to the lecture and use the lecture browser interface to create a summary of the required length. Summaries were created manually by selecting a set of utterances from the lecture transcript. Participants were told they would listen to an actual lecture and that a summary could be taken with them into their final exam as an aid. Under this assumption, the participants were advised to use their best judgement on which utterances were appropriate to include in the summary. In the evaluation phase, the same summary was provided for use in answering the evaluation quiz. Only those utterances in the summary were provided in the transcript and audio. However, all slides were available.

3.10.3 A3: Primed manual summary

This condition is similar to the *generic manual summary* condition with the following difference. During the summarizing phase, participants were given a set of priming questions (Appendix A.3, A.6, A.9, A.12). Participants were told they were listening to an actual lecture where they were able to create a summary that could be taken

into their final exam as an aid. They were told that the instructor had provided a set of questions where the final exam would consist of a subset of these questions. The participants were not told how many of these question would appear on the evaluation quiz. Twenty-one priming questions were provided per quiz, while each evaluation quiz consisted of twelve of these questions. Priming questions were not ordered according to lecture content, meaning that using these questions effectively required effort on the part of the participant.

3.10.4 A4: Automatic summary

The procedure for this condition is identical to the *generic manual summary* condition from the point of view of the participant. However, during the evaluation phase, an automatically generated summary was provided instead of the summary which the participant created him or herself in the summarizing phase. The algorithm used to generate this summary was an implementation of MMR. There was only one MMR summary for each lecture, meaning that multiple participants made use of identical summaries. The automatic summary was created by adding the highest scoring utterances one at a time until the sum of the length of all of the selected utterances reached 20% of the number of words in the original lecture. MMR was chosen as it is a successful and widely known algorithm and, as such, allows for comparison to literature. As explained in the previous chapter, using current evaluation methods, MMR has been shown to be a very competitive baseline, even among state-of-art summarization algorithms, which tend to correlate well with it [35]. In the last chapter of this thesis, we discuss the use of a more advanced algorithm for a subsequent study.

3.11 Lecture variation

Even with counterbalancing, there is still legitimate concern that variation in lecture content may have an effect on performance. A given lecture can naturally only be used in one condition for each participant. Although we are counterbalancing among 48 participants, a number that allows lecture variation to be minimized, we have decided to perform an extra step to further mitigate the confounding effect of lecture variation. The method used was to conduct the experiment on an additional four participants, where for each lecture, only the *no summary* condition was used. The purpose of the resulting scores was to gain an idea of how the difficulty of each lecture-quiz pair varied while keeping all other variables fixed. The counterbalancing scheme used in the original experiment consisted of a total of four lecture orderings. In line with this, each of these orderings was used for one participant in this normalizing experiment. We averaged the four scores for each lecture and used these to adjust the raw quiz scores to see if lecture variation had any effect.

Chapter 4

Results and Data Analysis

4.1 Overview

The study was carried out over a period of six months. In addition to pre-pilot and pilot participants, 48 participants completed the main experiment and 4 participants completed a mini experiment used to compensate for lecture variation. The result was a total of 208 completed quizzes, divided equally among the four lectures. After the study was completed, these quizzes were marked by a Sociology 101 teaching assistant and the data was recorded and analyzed.

4.2 Task completion times

Each participant was asked to watch four lectures while completing the summarization tasks. Sixty minutes were given per task, however, if the time elapsed and the task was not yet completed (i.e. the summary length was too long or too short), an extra 9 minutes, or 15% of the time, was provided. A total of 7 participants required this extra time. As the lectures were 40 minutes in length and each participant was instructed to watch each lecture in its entirety, the remaining participants took between 40 and 60 minutes to complete each lecture. The time taken is summarized in Figure 4.1. For

condition A1, time in excess of the length of the lecture was generally due to pausing the lecture in order to take notes. In the remaining conditions, aside from pausing the lecture, participants who overshot the summary length often used time after the lecture had completed playing in order to remove excess utterances. Some participants also went back through the lecture content to add additional utterances. This was particularly relevant for condition A3, where some participants went back to search for answers to priming questions that they had not previously found.

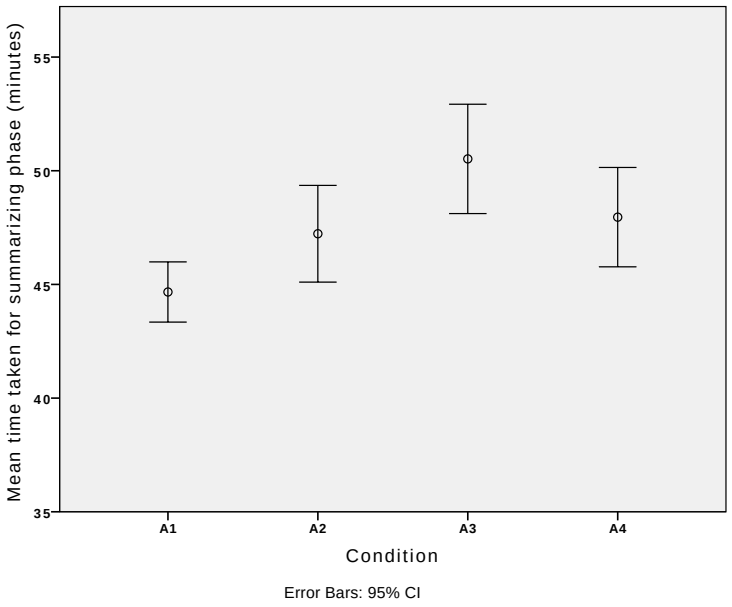

Figure 4.1: Time taken to complete summarizing phase

4.3 Quiz scores

Quizzes were scored by a teaching assistant for the sociology course from which the lectures were taken. Quizzes were marked as they would be in the actual course and each question was graded with equal weight, each out of two marks. The scores were then converted to a percentage. The resulting scores are $49.3 \pm 17.3\%$ for condition A1 (no summary), $48.0 \pm 16.2\%$ for condition A2 (generic manual summary), $49.1 \pm 15.2\%$ for

condition A3 (primed manual summary), and $41.0 \pm 16.9\%$ for condition A4 (automatic MMR summary). These scores are lower than averages expected in a typical university course. This can be partially attributed to the presence of a time constraint.

Execution of the Shapiro-Wilk Test confirmed the scores are normally distributed and Mauchly's Test of Sphericity indicates that the sphericity assumption holds. Skewness and Kurtosis tests were also employed to confirm normality. A repeated measures ANOVA determined that scores varied significantly between conditions ($F(3, 141) = 5.947$, $P = 0.001$). Post-hoc tests using the Bonferroni correction indicate that conditions A1, A2, and A3 resulted in higher scores than condition A4. The difference is significant at $P = 0.007$, $P = 0.014$ and $P = 0.012$ respectively. Although normality was assured, the Friedman Test further confirms a significant difference between conditions $\chi^2(3) = 11.684$, $P = 0.009$.

4.3.1 The impact of slides

In order to partially isolate the effect of the summary content versus material found in the slides, we removed those questions that could be answered using only slides (Type 1 questions). The logic behind this was that such questions could be answered equally well regardless of the content of the summary, or if a summary was provided at all. There are three of these questions in each quiz. As such, we now analyze the quiz scores using only the remaining nine questions.

Again, with these scores, the sphericity assumption holds and the data is confirmed to be normally distributed. The resulting mean scores are $43.8 \pm 21.0\%$ for condition A1, $42.5 \pm 18.1\%$ for condition A2, $46.3 \pm 15.7\%$ for condition A3, and $35.2 \pm 17.7\%$ for condition A4. Similar to the raw scores, a repeated measures ANOVA shows significant difference between conditions ($F(3, 141) = 6.637$, $P < 0.0005$). Post-hoc tests using the Bonferroni correction show that conditions A1, A2, and A3 resulted in significantly higher scores than A4, at $P = .021$, $P = 0.034$, and $P = 0.001$ respectively.

The removed questions should theoretically be able to be answered equally well in all conditions, since each condition provides the participant with full access to the slides. A more significant ANOVA P value suggests that for those questions actually requiring use of the non-slide material, the type of summary provided is actually more significant than the raw scores show.

Slides are also a type of summary in themselves. He et al. [13] show that slides containing a larger amount of information are preferable to those only showing the main points. In our experiment, slides are essentially competing with our explicit summaries. However, this is valid, as in actual university lectures slides are provided and a summary must provide additional value to these slides to be worth using at all.

4.3.2 Adjustment for lecture variation

As noted in the previous chapter, a mini experiment was conducted where four separate participants completed condition A1 for each of the four lectures. The average scores for each lecture (Figure 4.2) are 56.3% for Lecture 1, 53.1% for Lecture 2, 60.4% for Lecture 3, and 65.6% for Lecture 4.

It is evident that even when using the same participants, the same conditions and proper counterbalancing, different lectures result in slightly different scores. This could be due to the inherent level of difficulty of the lecture content, or bias caused by the questions chosen in the quiz. Regardless, counterbalancing on the 48 participants used to generate scores above means that the effect of lecture variation should result in minimal bias of the results. However, we use this mini experiment to further investigate any possible effect.

To accomplish this, for each quiz completed during the full study, we calculated the difference between the quiz score and the average score for that lecture as determined by the mini experiment. That is, we subtracted the mini experiment average from the the original score. Averaging these results by condition, we get $-9.55 \pm 17.7\%$ for condition

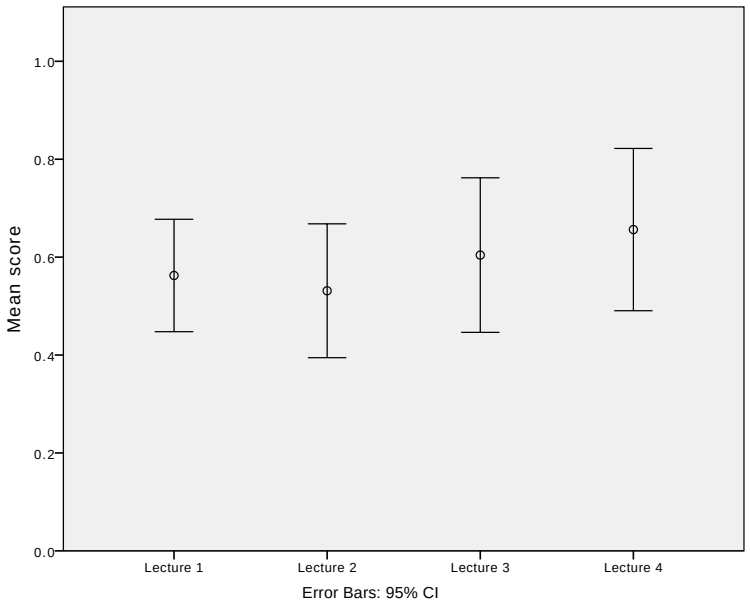

Figure 4.2: Quiz scores for mini experiment by lecture

A1, $-10.8 \pm 16.1\%$ for condition A2, $-9.72 \pm 15.7\%$ for condition A3, and $-17.9 \pm 17.2\%$ for condition A4.

Sphericity and normality were confirmed to hold. A repeated measures ANOVA indicates that these values are significantly different between conditions ($F(3, 141) = 5.417$, $P = .001$). Post-hoc tests using the Bonferroni correction show that condition A4 is significantly lower than A1 ($P = 0.011$), A2 ($P = 0.022$), and A3 ($P = 0.014$). Condition A1 is used for all lectures in the mini experiment. In condition A1, the entire transcript, notes, and slides are available. Additionally, the raw scores of the study show that condition A1 results in the highest average score. As such, we can use the mini experiment average scores as a sort of top line. The fact that all of the differences are negative reflects this idea; meaning that we are measuring how much worse than the top line each condition is. The fact that A4 performs the lowest simply means that using the summary provided in condition A4 results in the greatest negative deviation from the top line as established by the four mini experiment subjects.

These results generate the same conclusions as those calculated using the raw scores.

This further confirms that the results obtained can be attributed to the summary conditions and not to lecture variation.

4.4 Hypotheses and discussion

Based on the results above, we now address the hypotheses outlined in the previous chapter.

4.4.1 No summary hypothesis

Using no summary will perform at least as well as using a summary under ecologically valid settings.

This hypothesis can only be rejected for condition A4, that is, the automatic MMR summary. For all other conditions, using a summary provides no additional benefit compared to having access to the entire transcript, audio, and hand-written notes. More importantly, using a manual summary does not result in worse performance. This means that a well-designed summary is capable of being as effective as having access to the entire lecture content as well as notes. This suggests that extractive summaries have a lot of potential, as they are essentially as good as the original content, even after removing those questions that slides have the highest chance of helping with. As such, we conclude that summaries are potentially a suitable replacement for lectures in a time-constrained situation.

4.4.2 Primed summary hypothesis

Primed summaries do not lead to better performance on lecture quizzes.

This hypothesis cannot be rejected, as our results have been unable to demonstrate that providing priming questions improves summary performance. This may be surprising at first, and cannot be used to refute the effectiveness of priming in all contexts.

However, the fact that there is no significant difference in performance between primed and generic manual summaries is promising, as it suggests that not having access to evaluation quizzes does not hinder summary performance. If humans can create effective summaries unaided, when placed under ecologically valid experimental conditions, then automatic summarization systems may someday also have this potential.

4.4.3 Automatic summary hypothesis

Automatic summaries generated using MMR will not perform significantly better than manual summaries.

Our results have demonstrated that MMR results in summaries that perform significantly worse than manual summaries, or no summary. As such, this hypothesis can be rejected. Penn and Zhu [35] show that, when evaluated using conventional evaluation measures which are not ecologically valid, the performance of MMR is indistinguishable from state-of-the-art summarizers that make use of every conceivable feature available. Showing that, under an ecologically valid evaluation paradigm, MMR fares worse than manual and no summary conditions puts even more doubt on conventional evaluation measures that make use of subjective gold standard summaries. Future ecologically valid evaluation of the same full-featured summarizer used by Penn and Zhu [35] may further confirm this by also resulting in poor scores for MMR.

4.5 ROUGE

ROUGE [18] is often used to evaluate summarization. Although Lin [18] claimed to have demonstrated that ROUGE correlates well with human summaries, both Murray et al. [31], and Liu and Liu [21] have cast doubt upon this. It is important to acknowledge, however, that ROUGE is actually a family of measures, distinguished not only by the manner in which overlap is measured (1-grams, longest common subsequences, etc.), but

by the provenience of the summaries that are provided to it as references. If these are not ecologically valid, there is no sense in holding ROUGE accountable for an erratic result.

To examine how ROUGE fairs under ecologically valid conditions, we calculated ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-SU4 on our data using the standard options outlined in previous DUC evaluations. ROUGE scores were calculated for each of the generic manual summary, primed manual summary, and automatic summary conditions. Each summary in a given condition was evaluated once against the generic manual summaries and once using the primed manual summaries. Similar to Liu and Liu [21], ROUGE evaluation was conducted using leave-one-out on the model summary type and averaging the results.

In addition to calculating ROUGE on the summaries from our ecologically valid evaluation, we also followed more conventional ROUGE evaluation and employed three annotators to create gold standard summaries using binary selection. Annotators were not primed in any sense, did not listen to the lecture audio, and had no sense of the higher level purpose of their annotations. We refer to the resulting summaries as context-free, as they were not created under ecologically valid conditions. Using these context-free summaries, the original generic manual, primed manual, and automatic summaries were evaluated using ROUGE. The result of these evaluations are presented in Table 4.1 (Average ROUGE scores for individual lectures can be found in Appendix C.1 - C.2).

Looking at the ROUGE scores, we can see that when evaluated by each type of model summary, MMR performs worse than either generic or primed manual summaries. This is consistent with our quiz results, and perhaps shows that ROUGE may be able to distinguish human summaries from MMR. Looking at the generic-generic, primed-primed, and context-free-context-free scores, we can get a sense of how much agreement there was between summaries. It is not surprising that context-free annotator summaries showed the least agreement, as these summaries were generated with no higher purpose in mind. This suggests that using annotators to generate gold standards in such a manner

Table 4.1: Average ROUGE scores

Peer Type	Model Type	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-SU4
generic	generic	0.75461	0.48439	0.75151	0.51547
primed	generic	0.74408	0.46390	0.74097	0.49806
mmr	generic	0.71659	0.40176	0.71226	0.44838
generic	primed	0.74457	0.46432	0.74091	0.49844
primed	primed	0.74693	0.46977	0.74344	0.50254
mmr	primed	0.70773	0.38874	0.70298	0.43802
generic	context-free	0.72735	0.46421	0.72432	0.49573
primed	context-free	0.71793	0.44325	0.71472	0.47805
mmr	context-free	0.69233	0.37600	0.68813	0.42413
context-free	context-free	0.70707	0.44897	0.70365	0.48019

is not ideal. In addition, real world applications for summarization would conceivably rarely consist of a situation where a summary was created for no apparent reason. More interesting is the observation that, when measured by ROUGE, primed summaries have less in common with each other than generic summaries do. The difference, however, is less pronounced when measured by ROUGE than by F-measure. This is likely due to the fact that ROUGE can account of semantically similar utterances.

4.5.1 Correlation between ROUGE and quiz scores

In order to assess the ability of ROUGE to predict quiz scores, we measured the correlation between ROUGE scores and quiz scores on a per participant basis. Similar to Murray et al. [31], and Liu and Liu [21], we used Spearmans rank coefficient (rho) to measure the correlation between ROUGE and our human evaluation. Correlation was measured both by calculating Spearman's rho on all data points ("all" in Table 4.2) and by performing the calculation separately for each lecture and averaging the results ("avg"

in Table 4.2). Significant rho values (p-value less than 0.05) are shown in bold.

Note that there are not many bolded values, indicating that there are few (anti-) correlations between quiz scores and ROUGE. The rho values reported by Liu and Liu [21] correspond to the all row of our generic-context-free scores (Liu and Liu [21] did not report ROUGE-L), and we obtained roughly the same scores as they did. In contrast to this, our "all" generic-generic correlations are very low. It is possible that the lectures condition the parameters of the correlation to such an extent that fitting all of the quiz-ROUGE pairs to the same correlation across lectures is unreasonable. It may therefore be more useful to look at rho values computed by lecture. For these values, our R-SU4 scores are not as high relative to R-1 and R-2 as those reported by Liu and Liu [21]. It is also worth noting that the use of context-free binary selections as a reference results in increased correlation for generic summaries, but substantially decreases correlation for primed summaries.

With the exception that generic references prefer generic summaries and primed references prefer primed summaries, all other values indicate that both generic and primed summaries are better than MMR. However, instead of ranking summary types, what is important here is the ecologically valid quiz scores. Our data provides no evidence that ROUGE scores accurately predict quiz scores.

Table 4.2: Correlation (Spearman's rho) between quiz scores and ROUGE

Peer Type	Model Type		ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-SU4
generic	generic	all	0.017	0.066	0.005	0.058
		lec1	0.236	0.208	0.229	0.208
		lec2	0.276	0.280	0.251	0.092
		lec3	0.307	**0.636**	0.269	0.428
		lec4	0.193	-0.011	0.175	0.018
		avg	0.253	0.278	0.231	0.187
primed	generic	all	-0.097	-0.209	-0.090	-0.192
		lec1	-0.239	-0.458	-0.194	-0.458
		lec2	-0.306	-0.281	-0.306	-0.316
		lec3	0.191	0.142	0.116	0.255
		lec4	**-0.734**	**-0.780**	**-0.769**	**-0.780**
		avg	-0.272	-0.344	-0.288	-0.325
generic	primed	all	0.009	0.158	-0.004	0.133
		lec1	0.367	0.247	0.367	0.162
		lec2	**0.648**	0.425	**0.634**	0.304
		lec3	0.078	0.417	0.028	0.382
		lec4	0.129	0.079	0.115	0.025
		avg	0.306	0.292	0.286	0.218
primed	primed	all	0.161	0.042	0.161	0.045
		lec1	0.042	-0.081	0.042	-0.194
		lec2	0.238	0.284	0.259	0.284
		lec3	0.205	0.120	0.205	0.120
		lec4	0.226	0.423	0.314	0.423
		avg	0.178	0.187	0.205	0.158
generic	context-free	all	0.282	**0.306**	0.265	**0.347**
		lec1	-0.067	0.296	-0.004	0.325
		lec2	0.414	0.414	0.438	0.319
		lec3	0.410	0.555	0.410	0.555
		lec4	0.136	0.007	0.136	0.054
		avg	0.223	0.318	0.245	0.313
primed	context-free	all	-0.146	-0.282	-0.151	**-0.305**
		lec1	0.151	-0.275	0.151	-0.299
		lec2	-0.366	**-0.611**	-0.366	**-0.636**
		lec3	0.273	0.212	0.273	0.202
		lec4	**-0.815**	-0.677	**-0.825**	**-0.755**
		avg	-0.189	-0.338	-0.192	-0.372

Chapter 5

Conclusions and Future Work

5.1 Overview

We have conducted an ecologically valid evaluation of speech summarization in the university lecture domain. Results indicate that having no summary, a manual generic summary, or a primed generic summary do not affect scores on an evaluation quiz. However, automatic summaries generated using MMR result in significantly worse performance. As manual summaries were indistinguishable from having full access to the lecture material, we conclude that summaries are potentially a suitable replacement for lectures in a time-constrained environment. Furthermore, the failure of primed summaries to outperform generic summaries demonstrates that humans can create effective summaries unaided in the lecture domain, when placed under ecologically valid experimental conditions. This is similar to how high calibre students essentially have to predict what the exam questions will be based on lecture content. Whether this finding can be generalized to other domains is unclear. The fact that MMR has been shown to be less effective than manual summaries provides further evidence that Penn and Zhu [35] have raised a legitimate charge against ecologically invalid evaluation making use of subjective gold standards. A future evaluation making use of their advanced summarization algorithm

may further confirm this.

Finally, our results have been compared to ROUGE, including using conventional context-free annotated gold standard summaries. Results failed to show a correlation between ROUGE and human evaluation under ecologically valid conditions. This raises considerable doubt on whether ROUGE can legitimately be used as a substitute for proper human evaluation.

5.2 Future data analysis

Qualitative questionnaires were presented to participants after completing each quiz as well as at the end of the entire evaluation. Future work includes analyzing the collected data, which consists mostly of Likert scale questions [17] as well as some short answer questions. This may determine how well participants' subjective beliefs about quiz difficulty and perceived performance on the quizzes are reflected in the actual quiz scores. Related to this, participants were asked to check off questions that they believed they already knew the answer to. Using the results of this previous knowledge assessment, we can determine whether participants were accurately able to assess their own level of understanding. In addition, questionnaire responses can be used to inform the design of future studies. For example, participants were asked how similar the evaluation quizzes were to actual university level quizzes and exams. Finally, short answer questions may provide information regarding how intuitive the interface was and whether or not participants felt that the interface was hindering their use of summaries.

He et. al [12] report that the order in which summaries are presented to participants has an effect on the perceived quality of the summary. Although, like our study, counterbalancing resulted in each summary being placed equally often in each sequence position, participants rated the last summary shown to them as being significantly clearer, less choppy and of higher quality than the rest [12]. Our quiz scores can also be broken

down based on the order in which each participant completed each quiz. Analysis can be carried out to determine if order has any effect on not just the qualitative perception of the summary, but the resulting quiz scores as well.

5.3 The next experiment

In conjunction with the experiment conducted, we have designed a second experiment that will act as an extension to the work described in this thesis. This experiment follows a very similar protocol, aiming to test similar hypotheses under sligtly different conditions.

Our current experiment has simulated a scenario where someone has heard a lecture at least a week in the past and may or may not remember the content. In our next experiment, we aim to simulate a scenario in which someone wants to extract information from a lecture source that he or she has not previously heard. This is applicable for situations where a student has missed a class and may even be able to determine whether, under certain conditions, a summary is able to replace listening to a lecture entirely.

In this experiment, participants will be asked to make use of summaries to complete the same tasks; that is, quizzes covering the content of university level sociology lectures. However, this time participants will not be creating manual summaries, but instead using manual summaries created either by other participants or annotators. In fact, they will not be viewing any of the lectures prior to completing the task.

5.3.1 Advanced summarization algorithm

In this second experiment, in addition to MMR, we will use a state-of-the art automatic summarizer. The automatic summarizer we will use in this study is the one described by Penn and Zhu [35] (Figure 5.1). This summarizer makes use of a wide variety of speech features and is representative of the state-of-the-art. The summarizer takes either

ASR or manual transcripts as well as an audio file as input which it uses to process disfluenceis and extract various features important to identifying sentences. False starts and repetitions, which occur commonly in spontaneous speech, are detected and removed. A binary logistic regression classifier is used to train an utterance selection module that can make use of various lexical (MMR score, utterance length, etc.), structural (utterance position, etc.), and acoustic (pitch, energy, speaking rate, etc.) features, among others.

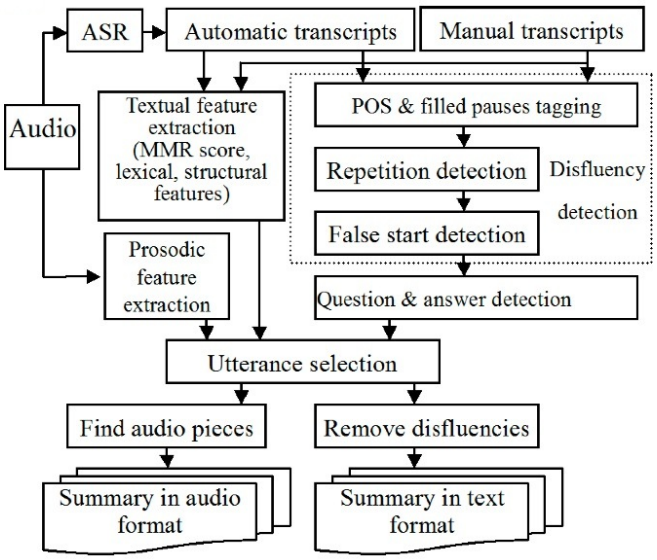

Figure 5.1: Automatic summarization for spontaneous conversation

By doing this, we can see where the performance of a full-featured summarizer actually fits with respect to MMR and manual summaries. Since we have already shown that primed summarization is no better than generic summarization under these conditions, we will exclude a primed manual summary condition.

5.4 Future work

The study presented in this thesis provides a first step in tackling the flaws present in current speech summarization evaluation. We have placed doubt on the practice of employing subjective gold standards and intrinsic evaluation metrics such as ROUGE.

However, we must recognize the importance of using gold standards as a method to quickly evaluate experimental summarization systems. Although a study like ours is able to accurately measure the value of a summary, it is time-consuming and expensive. Future work should focus on finding methods to turn results from large-scale studies, such as this one, into gold standards that can be repeatedly used for tuning and developing new systems. This type of work can involve developing novel methods for automatic evaluation that are grounded in the ecologically valid paradigm, but without the need to repeat the corresponding human subject studies.

Bibliography

[1] RM Baecker, G. Moore, and A. Zijdemans. Reinventing the lecture: Webcasting made interactive. In *Proc. HCI International 2003*, volume 1, pages 896–900, 2003.

[2] J.A. Brotherton and G.D. Abowd. Lessons learned from eclass: Assessing automated capture and access in the classroom. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 11(2):121–155, 2004.

[3] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336. ACM, 1998.

[4] F.R. Chen and M. Withgott. The use of emphasis to automatically summarize a spoken discourse. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 229–232. IEEE, 1992.

[5] E.C. Cherry and WK Taylor. Some further experiments upon the recognition of speech with one and with two ears. *Journal of the Acoustical Society of America*, 1954.

[6] H. Christensen, B.K. Kolluru, Y. Gotoh, and S. Renals. From text summarisation to style-specific summarisation for broadcast news. *Advances in Information Retrieval*, pages 223–237, 2004.

[7] P.R. Cohen. *Empirical methods for artificial intelligence*, volume 55. MIT press Cambridge, Massachusetts, 1995.

[8] C. Dufour, E.G. Toms, J. Lewis, and R. Baecker. User strategies for handling information tasks in webcasts. In *CHI'05 extended abstracts on Human factors in computing systems*, pages 1343–1346. ACM, 2005.

[9] M. Galley. A skip-chain conditional random field for ranking meeting utterances by importance. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 364–372. Association for Computational Linguistics, 2006.

[10] J. Glass, T.J. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay. Recent progress in the mit spoken lecture processing project. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[11] I. Gurevych and M. Strube. Semantic similarity applied to spoken dialogue summarization. In *Proceedings of the 20th international conference on Computational Linguistics*, page 764. Association for Computational Linguistics, 2004.

[12] L. He, E. Sanocki, A. Gupta, and J. Grudin. Auto-summarization of audio-video presentations. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, pages 489–498. ACM, 1999.

[13] L. He, E. Sanocki, A. Gupta, and J. Grudin. Comparing presentation summaries: slides vs. reading vs. listening. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 177–184. ACM, 2000.

[14] C. Hori, T. Hori, and S. Furui. Evaluation method for automatic speech summarization. In *Eighth European Conference on Speech Communication and Technology*, 2003.

[15] P.Y. Hsueh and J.D. Moore. Improving meeting summarization by focusing on user needs: a task-oriented evaluation. In *Proceedings of the 14th international conference on Intelligent user interfaces*, pages 17–26. ACM, 2009.

[16] A. Inoue, T. Mikami, and Y. Yamashita. Improvement of speech summarization using prosodic information. In *Proc. of Speech Prosody*, pages 599–602, 2004.

[17] R. Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[18] C.Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the workshop on text summarization branches out (WAS 2004)*, pages 25–26, 2004.

[19] C.Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 71–78. Association for Computational Linguistics, 2003.

[20] F. Liu and Y. Liu. Correlation between rouge and human evaluation of extractive meeting summaries. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 201–204. Association for Computational Linguistics, 2008.

[21] F. Liu and Y. Liu. Exploring correlation between rouge and human evaluation on meeting summaries. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(1):187–196, 2010.

[22] Y. Liu and S. Xie. Impact of automatic sentence segmentation on meeting summarization. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 5009–5012. IEEE, 2008.

[23] S. Maskey and J. Hirschberg. Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[24] S. Maskey and J. Hirschberg. Summarizing speech without text using hidden markov models. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 89–92. Association for Computational Linguistics, 2006.

[25] S.R. Maskey, A. Rosenberg, and J. Hirschberg. Intonational phrases for speech summarization. In *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[26] C. Munteanu. *Useful Transcriptions of Webcast Lectures*. PhD thesis, University of Toronto, 2009.

[27] C. Munteanu, R. Baecker, G. Penn, E. Toms, and D. James. The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 493–502. ACM, 2006.

[28] G. Murray, T. Kleinbauer, P. Poller, S. Renals, T. Becker, and J. Kilgour. Extrinsic summarization evaluation: A decision audit task. *ACM Transactions on SLP*, 6(2), 2009.

[29] G. Murray, T. Kleinbauer, P. Poller, S. Renals, J. Kilgour, and T. Becker. Extrinsic summarization evaluation: A decision audit task. *Machine Learning for Multimodal Interaction*, pages 349–361, 2008.

[30] G. Murray, S. Renals, and J. Carletta. Extractive summarization of meeting recordings. 2005.

[31] G. Murray, S. Renals, J. Carletta, and J. Moore. Evaluating automatic summaries of meeting recordings. In *Proc. of the ACL 2005 MTSE Workshop, Ann Arbor, MI, USA*, pages 33–40, 2005.

[32] A. Nenkova. Summarization evaluation for text and speech: issues and approaches. In *Ninth International Conference on Spoken Language Processing*, 2006.

[33] A. Nenkova and R. Passonneau. Evaluating content selection in summarization: The pyramid method. In *Proceedings of HLT-NAACL*, volume 2004, pages 145–152, 2004.

[34] A.S. Park and J.R. Glass. Unsupervised pattern discovery in speech. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(1):186–197, 2008.

[35] G. Penn and X. Zhu. A critical reassessment of evaluation baselines for speech summarization. *Proceedings of ACL-HLT. Columbus, OH*, 2008.

[36] D.R. Radev and D. Tam. Summarization evaluation using relative utility. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 508–511. ACM, 2003.

[37] A. Ranjan, R. Balakrishnan, and M. Chignell. Searching in audio: the utility of transcripts, dichotic presentation, and time-compression. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 721–730. ACM, 2006.

[38] L.A. Rowe, D. Harley, P. Pletcher, and S. Lawrence. Bibs: A lecture webcasting system. 2001.

[39] TED. http://www.ted.com/, 2012.

[40] S. Tucker, O. Bergman, A. Ramamoorthy, and S. Whittaker. Catchup: a useful application of time-travel in meetings. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 99–102. ACM, 2010.

[41] S. Tucker and S. Whittaker. Time is of the essence: an evaluation of temporal compression algorithms. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 329–338. ACM, 2006.

[42] S. Tucker and S. Whittaker. Temporal compression of speech: An evaluation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(4):790–796, 2008.

[43] YouTube. http://www.youtube.com/t/press_statistics/, 2012.

[44] K. Zechner. Automatic generation of concise summaries of spoken dialogues in unrestricted domains. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 199–207. ACM, 2001.

[45] K. Zechner and A. Waibel. Diasumm: Flexible summarization of spontaneous dialogues in unrestricted domains. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 968–974. Association for Computational Linguistics, 2000.

[46] K. Zechner and A. Waibel. Minimizing word error rate in textual summaries of spoken language. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 186–193. Morgan Kaufmann Publishers Inc., 2000.

[47] C. Zhang, Y. Rui, J. Crawford, and L. He. An automated end-to-end lecture capture and broadcasting system. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 4(1):6, 2008.

[48] X. Zhu, S. Kazemian, and G. Penn. Identifying salient utterances of online spoken documents using descriptive hypertext. In *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*, pages 173–176. IEEE, 2008.

[49] X. Zhu and G. Penn. Evaluation of sentence selection for speech summarization. In *Workshop of Crossing Barriers in Text Summarization, RANLP-2005, Bulgaria*. Citeseer, 2005.

[50] X. Zhu and G. Penn. Summarization of spontaneous conversations. In *Proc. of Interspeech*, pages 1531–1534. Citeseer, 2006.

[51] X. Zhu, G. Penn, and F. Rudzicz. Summarizing multiple spoken documents: finding evidence from untranscribed audio. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 549–557. Association for Computational Linguistics, 2009.

Appendices

Appendix A

Quizzes

Population Summary Guide **Participant ID: _____**

Please take into account the following questions when designing a summary.

What are people who study population called?
What is the rate of increase in the world population per year?
How do population theories compare to other types of sociological theories?
Population growth has been most drastic in the last how many years?

Figure A.1: Familiarization lecture priming questions

Population Quiz **Participant ID: _____**

Please answer the questions below.

How do population theories compare to other types of sociological theories?
Why is focusing on human populations more interesting than focusing on other types of populations, such as used car populations?
What are people who study population called?
American sociology, which has dominated the study of sociology, is most concerned with what factors?

Figure A.2: Familiarization lecture evaluation quiz

Gender Relations Summary Guide **Participant ID: _____**

Please take into account the following questions when designing a summary.

We see the most patriarchy in societies with high _____.
What main point do the changing trends for university enrollment and completion illustrate?
What are higher depression rates for women highly correlated to?
What positive consequence does the entrance of mothers in the Canadian workforce have on Canadian society?
What does it mean to say that there is greater fluidity in women's lives compared to men's?
What does the individualization of women's lives mean?
What does Dennis Hogan's finding about the changed ordering of education, work, and marriage suggest about how men are sequencing their lives?
Which two factors mediate the stress and health problems that women experience as a result of competing work/family demands?
Considering the lives of married women, provide an illustration of how patriarchy is in operation today?
What factors additionally constrain women's lives, but are relatively less impactful on men's lives?
What does idiosyncrasy refer to?
What was Buchman's main finding about the ordering of women's lives?
Why is the women's movement a paradox?
To what extent can we compare the lives of two women in Toronto, who are both currently 40 years old and hold a bachelor degree?
Why is there a low correspondence between women's age, education, experience and daily activity?
In order to predict where a particular man is at a given afternoon, what factors should we consider?
What implications does the declining birth rate have for gender relations?
If we were to compare the life histories of men in law school, what would we find?
At which life stage are women's lives most similar to men's?
According to the Globe and Mail, what proportion of students enrolling in university do not finish?
In hunter-gatherer societies, how were gender roles differentiated according to child-bearing functions?

Figure A.3: Lecture 1 priming questions

Gender Relations Quiz **Participant ID: _____**

Please answer the questions below.

In hunter-gatherer societies, how were gender roles differentiated according to child-bearing functions?
In order to predict where a particular man is at a given afternoon, what factors should we consider?
What implications does the declining birth rate have for gender relations?
Why is there a low correspondence between women's age, education, experience and daily activity?
Why is the women's movement a paradox?
To what extent can we compare the lives of two women in Toronto, who are both currently 40 years old and hold a bachelor degree?
What does idiosyncrasy refer to?
What factors additionally constrain women's lives, but are relatively less impactful on men's lives?
Which two factors mediate the stress and health problems that women experience as a result of competing work/family demands?
Considering the lives of married women, provide an illustration of how patriarchy is in operation today?
What does it mean to say that there is greater fluidity in women's lives compared to men's?
We see the most patriarchy in societies with high _____.

Figure A.4: Lecture 1 evaluation quiz

Previous Knowledge Assessment **Participant ID: _____**

Please check off the questions for which you already knew the answer to before listening to the lecture last week.

❑ In hunter-gatherer societies, how were gender roles differentiated according to child-bearing functions?
❑ In order to predict where a particular man is at a given afternoon, what factors should we consider?
❑ What implications does the declining birth rate have for gender relations?
❑ Why is there a low correspondence between women's age, education, experience and daily activity?
❑ Why is the women's movement a paradox?
❑ To what extent can we compare the lives of two women in Toronto, who are both currently 40 years old and hold a bachelor degree?
❑ What does idiosyncrasy refer to?
❑ What factors additionally constrain women's lives, but are relatively less impactful on men's lives?
❑ Which two factors mediate the stress and health problems that women experience as a result of competing work/family demands?
❑ Considering the lives of married women, provide an illustration of how patriarchy is in operation today?
❑ What does it mean to say that there is greater fluidity in women's lives compared to men's?
❑ We see the most patriarchy in societies with high _____.

Figure A.5: Lecture 1 previous knowledge assessment

Sexuality Summary Guide **Participant ID: _____**

Please take into account the following questions when designing a summary.

What is an important role that sexual communities (e.g. the gay community) play?
How does having children affect the sex life of married couples?
Whether or not women initiate sex on the first date might depend on:
How have people been dealing with the increased gender inequality that comes with having children?
What is heteronormativity?
What usually happens to women's satisfaction after babies are born?
Through their commitment to unpaid domestic work, women contribute to the _____ economy.
Why might inequality be a sexual turn-off?
What happens when someone's sexual tastes and desires do not follow the socially-accepted etiquette?
Women earn about 72% of what men earn, holding equal job type. What are some of the strategies that women can take in attempt to ensure equal pay for equal work?
Why do men resist changes to traditional gender relations more than women?
Why do women use tranquilizers at a higher rate than men?
Are men or women more likely to view sex as recreational?
What is intimacy?
In which two areas of sexuality are the social inequalities of race, gender, and class reflected?
That women who have sex with a lot of men are called "sluts" but men who have sex with a lot of women are called "studs" reflects which social process?
What does it mean when we say that we are ethnocentric about sex?
Why is it important to examine gender inequality when talking about sexuality?
List three types of costs are associated with having children.
What proportion of the population self-identify as homosexual?
Why do people who are socially unequal tend to avoid interaction?

Figure A.6: Lecture 2 priming questions

Sexuality Quiz **Participant ID: _____**

Please answer the questions below.

Why do people who are socially unequal tend to avoid interaction?
What proportion of the population self-identify as homosexual?
What does it mean when we say that we are ethnocentric about sex?
Why is it important to examine gender inequality when talking about sexuality?
What is intimacy?
Why do men resist changes to traditional gender relations more than women?
What happens when someone's sexual tastes and desires do not follow the socially-accepted etiquette?
Women earn about 72% of what men earn, holding equal job type. What are some of the strategies that women can take in attempt to ensure equal pay for equal work?
Through their commitment to unpaid domestic work, women contribute to the _____ economy.
Whether or not women initiate sex on the first date might depend on:
What is an important role that sexual communities (e.g. the gay community) play?
How does having children affect the sex life of married couples?

Figure A.7: Lecture 2 evaluation quiz

Previous Knowledge Assessment **Participant ID: _____**

Please check off the questions for which you already knew the answer to before listening to the lecture last week.

❑ Why do people who are socially unequal tend to avoid interaction?
❑ What proportion of the population self-identify as homosexual?
❑ What does it mean when we say that we are ethnocentric about sex?
❑ Why is it important to examine gender inequality when talking about sexuality?
❑ What is intimacy?
❑ Why do men resist changes to traditional gender relations more than women?
❑ What happens when someone's sexual tastes and desires do not follow the socially-accepted etiquette?
❑ Women earn about 72% of what men earn, holding equal job type. What are some of the strategies that women can take in attempt to ensure equal pay for equal work?
❑ Through their commitment to unpaid domestic work, women contribute to the _____ economy.
❑ Whether or not women initiate sex on the first date might depend on:
❑ What is an important role that sexual communities (e.g. the gay community) play?
❑ How does having children affect the sex life of married couples?

Figure A.8: Lecture 2 previous knowledge assessment

Social Movements Summary Guide **Participant ID: _____**

Please take into account the following questions when designing a summary.

Which major political party evolved from a religious social movement in Alberta?
What type of boundaries would we have to cut through in order to take a social movement global?
According to the conflict paradigm, whether or not a social movement will happen depends on what?
Which theory states that a precipitating factor is necessary for a social movement to form?
Give an example of the type of social goods that social movements are interested in redistributing.
What did first wave feminism focus on achieving?
What role does social class play in identity-based movements and New Social Movements?
What is the difference between social movements and interest groups?
What is a social movement?
Social movements are voluntary organizations, but why is a voluntary organization like a chess club not a social movement?
What is the major assumption of the conflict paradigm?
Why has third-wave feminism gone "underground"?
The gay and environment movement are examples of which type of social movement?
Which theory argues that people in the underclass are more likely to mobilize than the upper class because they have more to gain?
The Women's Movement is an example of which type of movement?
Why do most social movements fail?
What is the fundamental difference between the functionalist and the conflict views of why social movements occur?
What is the major critique of the relative deprivation theory?
According to functionalists, if society is working well, how often should social movements occur?
Relative Deprivation Theory argues that the problem rests not with society but with people's _____.
Why was the Relative Deprivation Theory so popular in the mid-20th Century?

Figure A.9: Lecture 3 priming questions

Social Movements Quiz **Participant ID: _____**

Please answer the questions below.

According to functionalists, if society is working well, how often should social movements occur?
Why do most social movements fail?
What is the fundamental difference between the functionalist and the conflict views of why social movements occur?
Which theory argues that people in the underclass are more likely to mobilize than the upper class because they have more to gain?
The gay and environment movement are examples of which type of social movement?
What is the major assumption of the conflict paradigm?
What is a social movement?
Social movements are voluntary organizations, but why is a voluntary organization like a chess club not a social movement?
What did first wave feminism focus on achieving?
Which theory states that a precipitating factor is necessary for a social movement to form?
Give an example of the type of social goods that social movements are interested in redistributing.
What type of boundaries would we have to cut through in order to take a social movement global?

Figure A.10: Lecture 3 evaluation quiz

Previous Knowledge Assessment **Participant ID: _____**

Please check off the questions for which you already knew the answer to before listening to the lecture last week.

❑ According to functionalists, if society is working well, how often should social movements occur?
❑ Why do most social movements fail?
❑ What is the fundamental difference between the functionalist and the conflict views of why social movements occur?
❑ Which theory argues that people in the underclass are more likely to mobilize than the upper class because they have more to gain?
❑ The gay and environment movement are examples of which type of social movement?
❑ What is the major assumption of the conflict paradigm?
❑ What is a social movement?
❑ Social movements are voluntary organizations, but why is a voluntary organization like a chess club not a social movement?
❑ What did first wave feminism focus on achieving?
❑ Which theory states that a precipitating factor is necessary for a social movement to form?
❑ Give an example of the type of social goods that social movements are interested in redistributing.
❑ What type of boundaries would we have to cut through in order to take a social movement global?

Figure A.11: Lecture 3 previous knowledge assessment

Immigration Summary Guide **Participant ID: _____**

Please take into account the following questions when designing a summary.

Aside from Canada, which four other countries also have high rates of immigration?
What is the American official policy on immigrants?
What was the assumption about people who did not come from hardy Northern countries?
Why was hardiness a factor that Canada initially looked for in its immigrants?
Today, the route for upward mobility is most open for ethnic minorities of which race?
What are the factors that predict the job status, job security and income level of immigrants?
Immigrants' experiences in Canada depend on racial origin, economic integration, cultural origin AND on _____.
What impact did immigration have on aboriginals?
For how many years after migration can immigrants expect to experience the discrepancy between their incomes and their education?
What does it mean when Porter says that Canada is the Great Railroad Station?
Compared to immigrants of the 1960s, second generation immigrants of today are much more upwardly mobile. What accounts for this difference?
Currently, ethnicity has less influence than _____ on one's position in the class structure.
The Polish Peasant demonstrates how immigrants in the new country used to rely on _____ to convince family friends in the old country to join them in migration.
In 1960, which groups were at the top of the "patchwork hierarchy"?
What accounts for the spike in immigration during the late 19th century/early 20th century?
Why is the Theory of Biculturalism problematic?
Originally, Canada most preferred immigrants from which parts of the world?
Fluctuations in immigration numbers reflect fluctuation in _____.
Immigrant status, along with _____, continues to predict SES.
How many different ethnic groups are in Canada?
Immigrants are the litmus test for the _____ of the host country.

Figure A.12: Lecture 4 priming questions

Immigration Quiz **Participant ID: _____**

Please answer the questions below.

Originally, Canada most preferred immigrants from which parts of the world?
In 1960, which groups were at the top of the "patchwork hierarchy"?
What accounts for the spike in immigration during the late 19th century/early 20th century?
The Polish Peasant demonstrates how immigrants in the new country used to rely on _____ to convince family friends in the old country to join them in migration.
What does it mean when Porter says that Canada is the Great Railroad Station?
Compared to immigrants of the 1960s, second generation immigrants of today are much more upwardly mobile. What accounts for this difference?
For how many years after migration can immigrants expect to experience the discrepancy between their incomes and their education?
What impact did immigration have on aboriginals?
Immigrants' experiences in Canada depend on racial origin, economic integration, cultural origin AND on _____.
Today, the route for upward mobility is most open for ethnic minorities of which race?
What are the factors that predict the job status, job security and income level of immigrants?
What is the American official policy on immigrants?

Figure A.13: Lecture 4 evaluation quiz

Previous Knowledge Assessment **Participant ID: _____**

Please check off the questions for which you already knew the answer to before listening to the lecture last week.

❑ Originally, Canada most preferred immigrants from which parts of the world?
❑ In 1960, which groups were at the top of the "patchwork hierarchy"?
❑ What accounts for the spike in immigration during the late 19th century/early 20th century?
❑ The Polish Peasant demonstrates how immigrants in the new country used to rely on _____ to convince family friends in the old country to join them in migration.
❑ What does it mean when Porter says that Canada is the Great Railroad Station?
❑ Compared to immigrants of the 1960s, second generation immigrants of today are much more upwardly mobile. What accounts for this difference?
❑ For how many years after migration can immigrants expect to experience the discrepancy between their incomes and their education?
❑ What impact did immigration have on aboriginals?
❑ Immigrants' experiences in Canada depend on racial origin, economic integration, cultural origin AND on _____.
❑ Today, the route for upward mobility is most open for ethnic minorities of which race?
❑ What are the factors that predict the job status, job security and income level of immigrants?
❑ What is the American official policy on immigrants?

Figure A.14: Lecture 4 previous knowledge assessment

Appendix B

Questionnaires and Forms

Post-quiz Questionnaire **Participant ID: _____**
Condition: _____

Please choose the appropriate answer for the following statements:

I think my answers on the quiz were:
[ ] All correct [ ] Mostly correct [ ] Some correct [ ] Mostly wrong [ ] All wrong

Compared to typical assignments in my classes, solving this quiz was:
[ ] Much easier [ ] Easier [ ] Same [ ] Harder [ ] Much harder

Compared to the introductory quiz from the beginning of this experiment, solving this quiz was:
[ ] Much easier [ ] Easier [ ] Same [ ] Harder [ ] Much harder

Having previously seen this lecture during a previous session was very useful in solving the quiz.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

For the questions that I completed, I am confident that my answers are correct.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I left some questions blank because of lack of time.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I left some questions blank because I could not find the correct answers.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I guessed the answers for questions where I could not find the correct answers.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

There was sufficient time to complete the quiz.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

Finding the answers using the summary / hand-written notes was easy.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

Finding the answers using the slides was easy.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I already knew the answer to _____ of the questions.
[ ] None [ ] Some [ ] About half [ ] Most [ ] All

I remembered the answer to _____ of the questions from listening to the lecture last week.
[ ] None [ ] Some [ ] About half [ ] Most [ ] All

Figure B.1: Post-quiz questionnaire

Post-experiment Questionnaire **Participant ID: _____**

What is your gender? _____

What is your field of study? _____

What year of study are you in?
[ ] 1st [ ] 2nd [ ] 3rd [ ] 4th [ ] 5th [ ] graduate

Are you currently enrolled in a sociology course?
[ ] Yes, university level [ ] Yes, college level [ ] No

Have you previously enrolled in a university or college level sociology course?
[ ] Yes, I enrolled and completed at least one course
[ ] Yes, I enrolled but did not complete the course
[ ] No

Have you previously enrolled in a university or college level anthropology, civics or social psychology course?
[ ] Yes, I enrolled and completed at least one course
[ ] Yes, I enrolled but did not complete the course
[ ] No

Have you ever attended a university or college level sociology lecture?
[ ] Yes, I have sat in on 1 or 2 lectures
[ ] Yes, I have sat in on 3 or more lectures
[ ] Yes, I have sat in on at least one semester worth of lectures
[ ] No

Have you ever attended a university or college level anthropology, civics or social psychology lecture?
[ ] Yes, I have sat in on 1 or 2 lectures
[ ] Yes, I have sat in on 3 or more lectures
[ ] Yes, I have sat in on at least one semester worth of lectures
[ ] No

Have you previously taken in a high-school level sociology course?
[ ] Yes, I enrolled and completed at least one course
[ ] Yes, I enrolled but did not complete the course
[ ] No

Have you previously taken in a high-school level anthropology, civics, or social psychology course?
[ ] Yes, I enrolled and completed at least one course
[ ] Yes, I enrolled but did not complete the course
[ ] No

Figure B.2: Post-experiment questionnaire (page 1 of 7)

Are you currently enrolled in an anthropology, civics or social psychology course?
[ ] Yes, university level [ ] Yes, college level [ ] No

Are you currently enrolled in a sociology course?
[ ] Yes, university level [ ] Yes, college level [ ] No

Please choose the appropriate answer for the following statements:

During the lecture where I took notes, I took more notes than I would have taken in a regular class:
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

During the lecture where I took notes, I took less notes than I would have taken in a regular class:
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

When creating summaries, I used the slides to determine which content to add to my summary:
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

When creating summaries, listening to the audio was very useful:
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

When creating summaries, reading the transcript was very useful:
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I had trouble keeping my summaries short enough to meet the length requirements:
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I had trouble keeping my summaries long enough to meet the length requirements:
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

There was sufficient time given to create good summaries:
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

The interface allowed me to easily create summaries:
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

When solving the quizzes, I found the answers easily by using the audio/video playback.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

When solving the quizzes, I found the answers easily by using the slides.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

When solving the quizzes, I found the answers easily by using the transcripts.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

When solving the quizzes, I found the answers easily by using the timeline.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I preferred using shortened lecture summaries to answer the quizzes compared to full-length lectures.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I preferred using shortened lecture summaries to answer the quizzes compared to my handwritten notes.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

Figure B.3: Post-experiment questionnaire (page 2 of 7)

Shortened lecture summaries were easy to listen to.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

Having previously watched the lectures was very helpful in completing the quizzes.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I was able to answer some questions from memory.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I was able to answer most questions from memory.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

Listening to the audio was very useful in solving the quizzes.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

Looking at the slides was very useful in solving the quizzes.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

Using the transcript was very useful in solving the quizzes.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

Using the audio timeline bar was very useful in solving the quizzes.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I found making the summaries myself useful in learning the lecture content.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I preferred using the summaries created by myself compared to the one created automatically.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I would consider using a lecture summarization system similar to the one provided if it was available.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I would consider using summarized lectures as a substitute for going to class.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I would consider using summarized lectures to aid in completing assignments.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I would consider using summarized lectures to prepare for an exam or quiz.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I would consider using summarized lectures to make up for a missed class.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I would prefer summarizing a lecture myself as opposed to having one automatically summarized.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

When solving the quizzes, I had trouble finding the answers using the audio/video playback.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

When solving the quizzes, I had trouble finding the answers using the slides.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

When solving the quizzes, I had trouble finding the answers using the transcripts.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

Figure B.4: Post-experiment questionnaire (page 3 of 7)

When solving the quizzes, I had trouble finding the answers using the timeline.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I would have preferred using full-length lectures to complete the quizzes instead of shortened summaries.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I preferred using my handwritten notes to complete the quizzes instead of shortened summaries.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

Shortened lecture summaries were difficult to listen to.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

Having previously watched a lecture was not helpful in completing the quizzes.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I was unable to answer most questions from memory.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I was unable to answer any questions from memory.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

Listening to the audio was not useful in solving the quizzes.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

Looking at the slides was not useful in solving the quizzes.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

Using the transcript was not useful in solving the quizzes.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

Using the audio timeline bar was not useful in solving the quizzes.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I did not find making the summaries myself useful in learning the lecture content.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I preferred using the automatically created summary compared to the ones created by myself.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I would not consider using a lecture summarization system similar to the one provided if it was available.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I would not consider using summarized lectures as a substitute for going to class.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I would not consider using summarized lectures to aid in completing assignments.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I would not consider using summarized lectures to prepare for an exam or quiz.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I would not consider using summarized lectures to make up for a missed class.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

I would prefer having an automatically generated summary as opposed to summarizing a lecture myself.
[ ] Strongly agree [ ] Agree [ ] Neutral [ ] Disagree [ ] Strongly disagree

Figure B.5: Post-experiment questionnaire (page 4 of 7)

Was the software interface provided effective at displaying summarized lectures?

What features of the system would you like to see improved?

Are there any features you feel are missing from the system that would help?

What did you like or dislike about the system?

What did you like or dislike about using summaries?

What other purposes would you use a summary for?

Figure B.6: Post-experiment questionnaire (page 5 of 7)

Compare the summaries you created with the one generated automatically.

What could be done to improve the effectiveness of automatically generated summaries?

Was the software system adequate for taking advantage of the full value of summaries?

What strategy did you use for selecting which utterances to place in a summary?

Do you think the summaries created and used in this study were of an appropriate length? Explain.

Please add any other comments about your experience in this study and with summaries?

Figure B.7: Post-experiment questionnaire (page 6 of 7)

In general how do you approach learning – how do you learn best?

I need to write it down:
[ ] Always [ ] Sometimes [ ] Never [ ] Don't know

I need to read it over:
[ ] Always [ ] Sometimes [ ] Never [ ] Don't know

When someone explains it to me:
[ ] Always [ ] Sometimes [ ] Never [ ] Don't know

Please choose the appropriate answer for the following statements:

I learn best if I can watch a video related to the material.
[ ] Always [ ] Sometimes [ ] Never [ ] Don't know

I learn best while working with others.
[ ] Always [ ] Sometimes [ ] Never [ ] Don't know

I learn best on my own without distractions.
[ ] Always [ ] Sometimes [ ] Never [ ] Don't know

I learn best with pictures, graphics, and charts.
[ ] Always [ ] Sometimes [ ] Never [ ] Don't know

I learn best by taking my own notes.
[ ] Always [ ] Sometimes [ ] Never [ ] Don't know

I learn best by summarizing the material.
[ ] Always [ ] Sometimes [ ] Never [ ] Don't know

I learn best by using summaries of the material prepared by the instructor.
[ ] Always [ ] Sometimes [ ] Never [ ] Don't know

I learn best by using lecture slides.
[ ] Always [ ] Sometimes [ ] Never [ ] Don't know

I learn best by reviewing the material at home after the lecture.
[ ] Always [ ] Sometimes [ ] Never [ ] Don't know

I learn best by taking notes during the lecture.
[ ] Always [ ] Sometimes [ ] Never [ ] Don't know

I learn best by completing assignments.
[ ] Always [ ] Sometimes [ ] Never [ ] Don't know

I learn best by completing quizzes or exams.
[ ] Always [ ] Sometimes [ ] Never [ ] Don't know

I learn best while working with others.
[ ] Always [ ] Sometimes [ ] Never [ ] Don't know

Figure B.8: Post-experiment questionnaire (page 7 of 7)

CONSENT FORM

I agree to participate in a study ("Ecologically valid evaluation of automatic lecture summarization") that determines if and how people benefit from using computer-generated text transcripts and summaries of audio recordings from lectures to complement the traditional playback of the audio/video lecture recording. I understand that my participation is entirely voluntary. The following points have been explained to me:

1. The purpose of this research is to compare the use of various summaries of recorded lectures accompanied by playback of the recordings. I understand that I will participate in both of the following tasks. The first task will take place over two separate visits, and the second task will take place during a third visit.
    a. Summarization
        i. I will be provided with a computer-based interface that allows playback of recorded lectures and display of slides and/or textual transcripts. I will be instructed on the use of this interface.
        ii. I will have to listen to a portion of a university lecture.
        iii. I will have to locate a set of segments from a textual transcript of lecture in order to create a summary.
    b. Evaluation
        i. I will be provided with a computer-based interface that allows playback of recorded lectures and display of slides and/or textual transcripts. I will be instructed on the use of this interface.
        ii. I will make use of a summary in order to complete a quiz on the subject of the lecture.

2. Upon completion of both tasks, I will be offered $70 as compensation.

3. The procedure will be as follows: I will use the computer provided by the investigator, and perform the tasks as indicated. For task (a) I will have to read the transcription provided on the computer, along with playing the indicated audio/video file. For task (b) I will complete a quiz related to the information presented in the lecture, with the aid of a provided summary.

4. The investigator does not foresee any risk greater than that of using a computer for everyday academic use.

5. I understand that I may withdraw from the study at any time.

6. I understand that I will receive a copy of this consent form.

7. All of the data collected will remain strictly confidential. Only people associated with the study will see my responses. My responses will not be associated with

Figure B.9: Consent form (page 1 of 2)

my name; instead, my name will be converted to a code number when the researchers store the data.

8. The experimenter will answer any other questions about the research either now or during the course of the experiment. If I have any other questions or concerns, I can address them to the faculty supervisor**,** Prof. Gerald Penn of the Department of Computer Science. He can be contacted by phone: 416-978-7390, email: gpenn@cs.utoronto.ca, or in person at his office: Room 396B, Pratt Building, 97 St. George Street, Toronto, Ontario.

9. Upon completion of my participation, I will receive an explanation about the rationale and predictions underlying this experiment.

_____
Participant's Printed Name

_____
Participant's Signature

_____
Date

_____
Experimenter Name

_____
Participant's Numerical Code

Figure B.10: Consent form (page 2 of 2)

Appendix C

Additional Data

Table C.1: Average ROUGE scores

Peer Type	Model Type		ROUGE-1		ROUGE-2		ROUGE-L		ROUGE-SU4	
			avg	stdev	avg	stdev	avg	stdev	avg	stdev
generic	generic	lec1	0.77673	0.015	0.52456	0.027	0.77425	0.015	0.54620	0.023
		lec2	0.75261	0.024	0.48352	0.037	0.74936	0.023	0.51409	0.032
		lec3	0.74575	0.018	0.45840	0.032	0.74281	0.018	0.49477	0.028
		lec4	0.74334	0.013	0.47106	0.025	0.73962	0.013	0.50680	0.020
		avg	0.75461	0.022	0.48439	0.039	0.75151	0.022	0.51547	0.032
primed	generic	lec1	0.75779	0.039	0.48912	0.057	0.75513	0.039	0.51669	0.048
		lec2	0.73846	0.019	0.45927	0.030	0.73487	0.019	0.49359	0.024
		lec3	0.74523	0.024	0.45895	0.030	0.74272	0.024	0.49603	0.026
		lec4	0.73483	0.020	0.44825	0.038	0.73115	0.020	0.48593	0.030
		avg	0.74408	0.027	0.46390	0.042	0.74097	0.028	0.49806	0.035
mmr	generic	lec1	0.73718	-	0.43328	-	0.73267	-	0.46848	-
		lec2	0.71161	-	0.40983	-	0.70766	-	0.45397	-
		lec3	0.70570	-	0.39795	-	0.70231	-	0.44249	-
		lec4	0.71188	-	0.36598	-	0.70641	-	0.42857	-
		avg	0.71659	0.014	0.40176	0.028	0.71226	0.014	0.44838	0.017

Table C.2: Average ROUGE scores (continued)

Peer Type	Model Type		ROUGE-1		ROUGE-2		ROUGE-L		ROUGE-SU4	
			avg	stdev	avg	stdev	avg	stdev	avg	stdev
generic	primed	lec1	0.75944	0.011	0.49149	0.020	0.75610	0.011	0.51872	0.017
		lec2	0.73859	0.015	0.45880	0.032	0.73486	0.015	0.49324	0.026
		lec3	0.74565	0.018	0.45894	0.035	0.74245	0.018	0.49607	0.029
		lec4	0.73460	0.021	0.44803	0.038	0.73022	0.021	0.48572	0.031
		avg	0.74457	0.019	0.46432	0.035	0.74091	0.019	0.49844	0.028
primed	primed	lec1	0.74964	0.028	0.46994	0.036	0.74616	0.028	0.50002	0.031
		lec2	0.74556	0.011	0.47930	0.023	0.74201	0.012	0.51031	0.020
		lec3	0.74833	0.021	0.46544	0.032	0.74546	0.021	0.50131	0.027
		lec4	0.74419	0.011	0.46439	0.019	0.74013	0.011	0.49849	0.018
		avg	0.74693	0.019	0.46977	0.028	0.74344	0.019	0.50254	0.024
mmr	primed	lec1	0.72874	-	0.41605	-	0.72387	-	0.45593	-
		lec2	0.69491	-	0.37508	-	0.69031	-	0.42589	-
		lec3	0.70032	-	0.39849	-	0.69719	-	0.44303	-
		lec4	0.70696	-	0.36533	-	0.70054	-	0.42722	-
		avg	0.70773	0.015	0.38874	0.023	0.70298	0.015	0.43802	0.014
generic	context-free	lec1	0.70598	0.012	0.47215	0.027	0.70312	0.013	0.49161	0.024
		lec2	0.72641	0.017	0.46163	0.041	0.72320	0.017	0.49254	0.033
		lec3	0.74041	0.017	0.46317	0.044	0.73816	0.017	0.49893	0.036
		lec4	0.73661	0.012	0.45990	0.026	0.73279	0.012	0.49985	0.021
		avg	0.72735	0.019	0.46421	0.035	0.72432	0.020	0.49573	0.028
primed	context-free	lec1	0.69261	0.017	0.43523	0.046	0.68976	0.017	0.46234	0.034
		lec2	0.71659	0.014	0.44321	0.027	0.71273	0.015	0.47726	0.022
		lec3	0.73911	0.025	0.45589	0.046	0.73699	0.025	0.49478	0.040
		lec4	0.72342	0.022	0.43867	0.041	0.71939	0.023	0.47781	0.033
		avg	0.71793	0.026	0.44325	0.040	0.71472	0.026	0.47805	0.034
mmr	context-free	lec1	0.68269	-	0.38211	-	0.67779	-	0.41776	-
		lec2	0.68528	-	0.37159	-	0.68055	-	0.41905	-
		lec3	0.70100	-	0.39443	-	0.69718	-	0.43813	-
		lec4	0.70034	-	0.35585	-	0.69698	-	0.42159	-
		avg	0.69233	0.010	0.37600	0.016	0.68813	0.010	0.42413	0.009
context-free	context-free	lec1	0.66387	0.012	0.41048	0.015	0.66104	0.009	0.43451	0.009
		lec2	0.69815	0.005	0.44315	0.006	0.69400	0.006	0.47419	0.005
		lec3	0.72541	0.019	0.46184	0.036	0.72227	0.021	0.49785	0.030
		lec4	0.74085	0.008	0.48041	0.008	0.73731	0.008	0.51422	0.008
		avg	0.70707	0.034	0.44897	0.030	0.70365	0.034	0.48019	0.035