SpEx: A Tool for Visualising and Navigating Speech Audio

by

Fahmi Abdulhamid

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Master of Engineering
in Software Engineering.

Victoria University of Wellington
2013

Abstract

Audio is a ubiquitous form of information that is usually treated as a single, unbreakable, piece of content. Thus, audio interfaces remain simple, usually consisting of play, pause, forward, and rewind controls. Spoken audio can contain useful information across multiple topics and finding the information desired is usually time consuming. Most audio players simply do not reveal the content of the audio. By using the speech transcript and acoustic qualities of the audio, I have developed a tool, SpEx, which enabled search and navigation within spoken audio. SpEx displayed audio as discrete segments and revealed the topic content of each segment using mature Information Visualisation techniques. Audio segments were produced based on the acoustic and sentence properties of speech to identify topically and aurally distinct regions. A user study found that SpEx allowed users to find information in spoken audio quickly and reliably. By making spoken audio more accessible, people can gain access to a wider range of information.

Acknowledgments

I would like to thank my family for their amazing ongoing support while I completed my thesis. Mavis, Hani, and Asim, your support is truly appreciated. I also thank Stuart Marshall, my supervisor, for his continual input and support of my work. My work would not have been as thorough as I present in this thesis, nor have it taken the direction which it has.

I also thank the members of the Human Computer Interaction group for their feedback and valuable advice on my work: Daniel Cope for his input and proof-reading of my thesis; Craig Anslow for his ongoing support and valuable advice; Roman Klapaukh and Kah Chan for their design suggestions, and Bridget Johnson for showing me her work on multi-touch music composition. Also, to my friends for their moral support, in no particular order: Arindam Bhakta, Timothy Jones, Julian Mackay, Daniel Atkins, Diane Strode, Michael Waterman, Chris Green, Jonathan Hart, and to everyone else, too numerous to list.

I owe much gratitude to David Brebner, of Fingertapps, for his support and contribution of ideas to my project. The original idea for my work stemmed from him. I also appreciate all those who took part in my user study for their time and helpful feedback. Finally, I would like to thank Victoria University of Wellington and the Victoria Masters by Thesis Scholarship for providing the facilities and financial aid to complete my work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Audio is a ubiquitous form of digital information. The ease with which audio can be recorded and played back has lead to the popularity of audio to store recordings of speeches, meetings, and personal notes to name a few. The modern "Audio Document" containing speech is easily recorded with today's mobile devices and can be quickly shared with a community over the Internet, making the Audio Document ideal for sharing knowledge. Audio provides extra depth to information. Audio stores not just *what* was said, but also *how* it was said and even background events.

Despite the ease of recording and sharing audio, software interfaces to play audio have not advanced far beyond the traditional tape recorder. In fact the tape recorder metaphor still exists today on our digital devices: audio controls still consist of play, pause, forward, and rewind. Finding that specific piece of information that you had forgotten or even understanding which parts are worthwhile listening to are still not easy tasks. Lectures make good examples of recordings that do not always have to be listened to sequentially. Students may wish to listen to only a part of a lecture, to revise an unclear section of their notes for example. Most interfaces simply do not provide users with structural information about the content recorded.

Advanced interfaces which do provide structural information are sim-

ply not able to display enough information for effective audio retrieval. Interfaces which navigate by structure do not enable fine-grain access to the underlying content. Interfaces which navigate by searching do not display enough information about the overall content of an audio recording. Interfaces which navigate by structure and by search require external information. Recording external information such as presentation slides and who is speaking increases the work involved to record the audio in the first place. Moreover, visual elements used are often sub-optimal. Elements are minimal and do not stray far from standard user interface components.

In this thesis, I am proposing a novel interface design for spoken content retrieval in the university domain. One which makes use of mature Information Visualisation techniques to exploit our powerful perceptual abilities. Users will be able to identify patterns and trends to find information in audio recordings. I call my system *SpEx*.

SpEx makes uses of the Treemap Information Visualisation technique to display audio in segments. Audio segments are built from characteristics of prosody and acoustic conditions. Prosody such as which words were stressed and the speaker's emotion coupled with background acoustic events are properties not found in text. They can be used to isolate topically and aurally distinct regions of audio to mark regions for navigation. Word Clouds are applied to the Treemap segments to display the topical flow of audio recordings. Meanwhile, a search facility and fine-grain interaction techniques promote precise access to audio at the sentence level. Treemaps have not before been used for audio navigation, nor have the interaction techniques I have extended the Treemap with. SpEx allows effective search and navigation within audio documents without the need for recording external information. Hence, SpEx is simple to use and to deploy. A user study demonstrates the high performance of SpEx for information retrieval in audio.

1.1 Goals and Objectives

The breadth of tasks audio has been applied to, even when we consider audio containing only speech, is vast. I constructed the goals and objectives to target the more narrow scope of common audio navigation tasks found in educational settings. Lectures provide a good source of spoken content and I believe that systems which can assist in the learning process are useful to help students to learn and assist lecturers to become more productive. The narrowing of the scope also ensures that some significant contributions can be made within the time allocated for a Masters thesis while still contributing knowledge to the general domain of spoken content navigation.

I identified three major goals: a feature extraction system to provide the visualisation with computed information from the spoken word and acoustic events found in audio; the construction of a visualisation to support a similar task taxonomy for audio navigation as described by Whittaker et. al. [93] and modified by Dufour et. al. [29]; and a user study to evaluate the developed visualisation. The detailed goals and objectives are listed below:

1. Design and build a system to extract textual features from transcripts and acoustic features from audio recordings to be used for Information Visualisation purposes.

2. Design and implement an interactive Information Visualisation prototype capable of displaying textual and acoustic features about audio recordings. The visualisation should support the following audio navigation tasks:

   T1 **Section Selection**: Users should identify relevant segments in audio recordings.

   T2 **Fact Finding:** Users should locate the location of specific items of information in an audio recording.

T3 **Summarisation**: Users should comprehend the content of an audio recording.

3. Evaluate the Information Visualisation prototype in a user study to understand the following:

   - Does the visualisation effectively support the audio navigation tasks described above?
   - What strategies lend themselves to the effective use of the visualisation?

1.2 Thesis Structure

This thesis will discuss the reasoning, design, and analysis of SpEx, and its audio processing companion application called TAFE. I start by describing my research methodology and form my requirements in Chapter 3. I then use my requirements to guide the visual design and interaction design of SpEx in Chapter 4. The audio feature extraction component is delegated to TAFE. TAFE is described and evaluated in Chapter 5. I design a user study to measure the performance of SpEx against my requirements in Chapter 6. The results of my user study are described and analysed in Chapter 7 before I conclude my thesis in Chapter 8.

Appendices are found at the end of the thesis. Appendix A contains my application for human ethics approval for my user study and consent forms given to participants. Appendix B contains the question sheet given to participants for my user study and answers. Appendix C contains anonymised raw results from my user study.

Chapter 2

Background and Related Work

In this chapter, I begin by describing what is Information Visualisation and provide a discussion on background and related work. In particular about audio visualisation, text visualisation, and speech transcripts.

2.1 Information Visualisation

Information Visualisation is a technique for displaying abstract data visually to support decision making [19]. We humans have a powerful perceptual system which is able to identify objects, spot patterns, and categorise the world around us into what we think is important and what is not. Unfortunately, the data that we are given are often presented in either textual or tabular form, a form where patterns and trends are difficult to assimilate. It is up to us to make decisions by mentally processing this information which becomes more difficult as the amount of data computers record and store increases. Much like how we may draw a brainstorm to organise our thoughts, Information Visualisation is a visual aid to organise our information.

Information Visualisation exploits the powerful visual display and processing capability of computers to display large collections of information as visual diagrams. The visual diagrams are designed such that our per-

ceptual system can find patterns, trends, and anomalies in data which may otherwise be too difficult to find. For example, a calender can display which days are most busy and a mind-map can display complex relationships between ideas. By revealing the characteristics of data, Information Visualisation serves to aid decision making by reducing the time it takes to extract meaningful information from data. Besides the effective use of design, Information Visualisation makes use of data mining techniques to process large amounts of data to reveal high-level characteristics for visualisation and also Human Computer Interaction (HCI) to allow users to easily consume, filter, and manipulate information more easily [20]. By manipulating data, users are able to find sought after information more efficiently and more effectively.

The application of data mining, visual design, and appropriate interaction technique has seen success [80], but only after careful consideration of design. Buxton states that products should be designed and tested "For the Wild" rather than in a laboratory because how a product is used by real users can differ from its intended purpose [18]. Consequently, those creating Information Visualisations must balance design and analytics in relation to their target audience. The correct choice of visual mapping and metaphor must be chosen to effectively present information. Poor design choice or an unattractive design may hinder the effectiveness of a visualisation [67]. One must also focus on how data is processed before it is presented. Visualising raw data may not have the same impact as inferring and visualising high-level patterns or trends found in data [52]. Displaying data using inappropriate design decisions can obscure the patterns and trends necessary for users to correctly explore and reason about data.

The young field of Multimedia Analytics advocates the use of Multimedia Information Retrieval (MIR) across multiple media sources to be used with Visual Analytics (pre-processing data for visualisation) to support decision making [21]. Indeed, Information Visualisation techniques

have been applied to video, audio, and supporting material to help users navigate multimedia content with notable success. Sections 2.2 and 2.3, below, give an overview of some of the efforts made to navigate audio and text material.

2.2 Spoken Audio Navigation

In the following sections I discuss prior work on audio navigation, both in the literature and in commercial industry. I identify three types of audio navigation interfaces: Navigation by Structure (Section 2.2.1) where the structure of the recorded conversation or speech is used for navigation; Navigation by Spoken Words (Section 2.2.2) where what has been spoken is used for navigation; and Navigation by Structure and Spoken Words (Section 2.2.3) where the two methods are used together. I follow with a short survey of text visualisation methods which are relevant to audio navigation (Section 2.3).

2.2.1 Navigation by Structure

Systems which navigate audio by the structure of speech typically identify and distinguish discrete segments of audio which correspond to topics or coherent units of speech. Systems which navigate by structure may display these identified segments to allow users to build an understanding of the holistic structure and important parts of audio. By comprehending the structure of audio, users may identify which portions of the audio are important and how different portions may be related to assist in navigation. Furthermore, discrete segments allow users to listen to portions of audio non-sequentially and out of order to distil only the parts which are perceived as important.

Speech skimming is one example of navigating audio by structure, though without segments. Speech skimming systems such as SpeechSkim-

mer [5] and the elastic audio slider [47] utilise the capabilities of the human auditory system to comprehend content and structure quickly by playing audio faster than normal speed. However, we can reliably interpret less information as the speed of audio increases which places an upper bound on the efficiency to which audio structure can be processed by speech skimming. SpeechSkimmer increases the upper bound of audio play speed by compression methods such as removing repeated tones (such as 'o' in 'book'), shortening pauses between words, removing background noise, and increasing play speed while maintaining speaker pitch. In contrast, the elastic audio slider performs no audio manipulation, but rather gradually increases then decreases the speed audio is played for comprehension.

Speech skimming can be useful if users are familiar with the audio they are listening to and the audio duration is short. By quickly skimming over audio, users can identify if an audio recording is relevant and, if so, extract important information. If users are not familiar with the audio or the audio duration is long, speech skimming can be difficult. Listening to a lengthy recording can still be a lengthy process, particularly if the structure is not known beforehand to skip known irrelevant segments. In which case, speech skimming is best suited for tasks to quickly comprehend the main points of audio recordings, i.e., summarising the audio content.

In contrast, visualising identified segments in audio can support navigation in longer recordings, even if users are not completely familiar with the audio. One such technique is to visualise speaker turn-taking behaviour. Systems such as The Listener [41] and RadioActive [96] segment audio by change of speaker. The Listener and RadioActive make the assumption that each speaker turn represents a coherent unit of speech which can be listened to in isolation from other segments and can be used to indicate boundaries between important facts and events. A similar technique used by MegaSound [41] segments audio between pauses because pauses may likewise separate sentences and coherent unit of speech. The

Listener, RadioActive, and MegaSound allow users to see the segments which make up an audio recording and listen to each segment.

Listening to each segment can increase navigation efficiency. Users can quickly make a relevance decision on a segment before moving onto the next. Therefore, important information can be found by skipping unimportant segments. Visualising turn-taking behaviour may also be useful if users are interested in particular speakers. For instance, by choosing to ignore the irrelevant speakers. But visualising turn-taking behaviour is not suitable for locating when particular topics occur (Section Selection) or locating specific pieces of information when searching for facts (Fact Finding).

We must accept that systems which automatically create segments are only useful if the segments produced match with users' search queries. For instance, identifying who is angry in an audio recording is not suited to a system which segments speech on speaker turn-taking behaviour or pauses. In contrast, systems which choose to visualise acoustic features leave the segmentation of audio to the powerful perceptual system in humans. We can identify patterns, trends, and anomalies in visual graphics efficiently. One such system for audio structure navigation produced colour-maps of audio recordings [64]. The colour-mapping technique has been used for non-speech audio prior [76] and can be used for spoken audio using acoustic features characteristic of speech. The colour-map technique splits an audio recording into thousands of frames of millisecond duration. Each frame is coloured based on its acoustic properties and all frames are put together to produce a pattern to describe the changing acoustic conditions of the audio such as speaker changes and intermissions.

I do not know of a formal study which had analysed the effectiveness of colour-maps for spoken audio navigation. I believe colour-maps are useful to aid users who are unfamiliar with an audio recording to assimilate its structure. But even for those who are familiar with an audio record-

ing, it would be difficult to identify when specific information is spoken for Fact Finding tasks. Colour simply does not map to what has been said. Furthermore, I am sceptical about the effectiveness of a colour-map for identifying patterns visually. Segments may not always be easy to distinguish because the human perceptual system cannot easily distinguish small changes in colour, let alone distinguish changes in the individual red, green, and blue components of colour [82] which may correspond to different acoustic features.

Systems which use the structure of spoken audio for navigation can improve retrieval performance, but later recall of audio structure deteriorates with such systems [94]. Users must re-familiarise themselves with the audio structure before locating the information they desire which increases the cognitive overhead of use. As a consequence, systems which support audio information retrieval without requiring users to understand the structure of audio recordings have been developed and will be discussed in Section 2.2.2, below.

2.2.2 Navigation by Spoken Words

The tools mentioned in Section 2.2.1 support audio navigation by visualising the structure, but not the content, of audio recordings. While visualising structure can be useful for certain search tasks (such as skipping intermissions in a radio show), it can also be a deterrent if a search task was related to what had been said. Tools which navigate by spoken words typically mimic text-based search tools. Spoken word based tools use the text transcript derived from the audio for navigation. The transcript contains the spoken words and the time the words occurred. The transcript is used for search and navigation typically without the need to comprehend the structure of an audio recording.

Systems for online audio retrieval such as SpeechBot [84], Speechfind [54], and MAVIS [66] have taken to displaying the automatically gener-

ated speech transcripts of audio much like traditional text-based website search engines. SpeechBot, Speechfind, and MAVIS allow users to find information in audio by entering search queries and displaying candidate results as text summaries. Terms in text summaries which appear in the search query highlighted in bold. Highlighting search terms in the sentences they occur is an effective way of finding information regardless of whether or not users understand the structure of an audio recording. But consequently they are not useful for those who want to identify the main topics or produce a summary of an entire audio recording. Important topics and segments are not displayed and if users don't know what an audio recording contains, search queries are not effective (one must ask "What do you search for?"). Linking summaries to the structure of audio recordings remains an open challenge [57, Ch. 6].

The automatic summaries do help with building a rough understanding of an audio recording, but a user study has found that summaries are not sufficiently descriptive unless they are manually generated [53]. MAVIS does additionally include manual summaries alongside automatic summaries at the expense of human time and cost, which may not always be feasible due to time and cost limitations.

Displaying a time-line of candidate audio recordings with markers to indicate where search terms are found in the audio is another method to aid audio navigation. Gradients [12], BBSearch [78], GAudi [2], and VoiceBase [87] are three systems which utilise a time-line to position search terms. Systems which annotate time-lines are useful for locating where specific information are likely to occur in audio recordings because the position of search terms highlight relevant portions of the audio, but the relevance of audio portions are not always clearly displayed. Time-line markers may point to words used out of context (such as "Internet" and "address" instead of "residential" and "address") and don't necessarily make it clear if all search terms are located nearby without careful inspection by the user. Audio navigation systems aim to reduce navigation effort, hence

systems should minimise work performed by users. VoiceBase and GAudi resolve term ambiguity by displaying the sentence each marker occurs in as context. Whereas Gradients highlights co-occurring terms by highlighting strongly related portions of audio stronger than others.

A similar technique utilises search term Chains [74]. Chains are sequences of nearby search term matches. Not only do Chains visually group co-occurring terms on a time-line, but also overlap to resolve term ambiguity. Grouping co-occurring terms assists in finding significant portions of audio that match a search query. Multiple nearby term matches are mentally treated as a single entity rather than having users mentally group nearby markers as described by the law of proximity in Gestalt physiology [65]. Notwithstanding, time-line based audio navigation with word markers is still not suitable for Section Selection or Summarisation unless users are already familiar with the audio. Without a known search query, no information can be presented to users. I cannot expect users to understand the structure of audio recordings by experimentally entering a search query and storing the result into short-term memory before entering another query. The process would be time consuming and prone to error.

While such spoken word based tools are useful for performing precise search queries, the lack of any structural display can be a hindrance. For instance, such tools are not useful for browsing or skimming audio because almost no structural information is present. In which case, assimilating the content of a recording is difficult. Typically, users who know what they want will search while users who only have a general idea of what they want will browse in exploration [10]. As discussed in Section 2.2.3, tools which utilise both the structural information as well as the spoken word are apt for tasks which involve both precise search queries and less precise browsing based tasks.

2.2.3 Navigation by Structure and Spoken Words

Tools which navigate by structure and spoken words both distinguish discrete segments of audio recordings and provide facilities to search the text transcript. While displaying audio structure assists in Section Selection and Summarisation tasks by providing important visual cues for browsing and scanning, querying tools support Fact Finding by providing precise access to specific information. An ideal audio navigation tool supports Section Selection, Fact Finding, and Summarisation in equal measure.

A popular method in the literature for visualising both structure and spoken words emphasises the display of automatically generated transcripts formatted as printable documents for search and navigation. The first such tool is SCAN [93] which introduced WYSIAWYH, "what you see is (almost) what you hear". WYSIAWYH prescribes the use of a visual analogue to display audio as formatted text documents. SCAN was improved to become SCANMail [92] and influenced the development of JotMail [90]. The speech-as-data methodology and interface [83] similarly attempted to focus on text, but instead made use of annotations, pause detection, and key-word spotting to support low-power devices. Low power devices, such as mobile phones, were (at the time) not capable of performing speech recognition due to limited performance, power, and time users were expected to wait for processing to complete.

Designed properly, such tools are apt for Fact Finding and Summarisation. The ability to search a transcript and to allow users to read an entire transcript, particularly around locations where search terms occur, fosters quick access to specific audio locations for Fact Finding and provides the ability to support strategic fixation [91] (spotting important words) to scan text efficiently for Summarisation. On the other hand, it is time consuming to read a long piece of text to understand what the major topics are and such tools rely on search queries for navigation. In which case, the inability to format a transcript with structural elements such as headings due to the free-form nature of speech hinders the effectiveness of the method to

support Section Selection.

An alternative method for navigating audio by both structure and spoken words, one which more effectively supports Section selection, makes use of external cues to provide structural overviews for audio recordings. External cues are typically domain-dependent. For instance, presentation slides can be used for navigating lecture recordings [45, 51], life events such as email and calender entries can be used for navigating recorded conversations in iRemember [85], and speaking and writing turn-taking behaviour can be used for navigating meeting recordings in HANMER [59]. External cues become indices to support explorative navigation in audio and serve to provide context to search queries by matching search terms to nearby cues. By displaying context, users can ignore irrelevant content more efficiently than manually listening to the audio for each candidate audio location. Furthermore, users are not expected to read through potentially long passages of text transcript which may be a deterrent.

External cues are particularly apt for determining the relevance of different portions of the audio. HANMER presents context by highlighting nearby and related segments, but HANMER does not make clear to users how segments are related. Navigation efficiency may be reduced because more audio must be listened to. In contrast Hürst's lecture browser [45] darkens lecture slide indicators to display slide relevance and leaves the task of determining related slides to users, while Lecture Video Browser [51] displays slide titles as an index to the audio. iRemember uses the calender metaphor to display when conversations occurred and which conversations are related to which life events. Although the use of external cues can be effective, the process of recording the presence of external factors may not always be easy. Recording slide transitions, logging all personal information, or having each participant in a meeting record what he or she says are not tasks which can be done without notable effort or investment in recording infrastructure. Furthermore, techniques which

make use of external cues cannot be applied to pre-existing recordings.

Using a multimodal approach with video in addition to audio can assist with audio navigation without recording external events. For instance, a video retrieval system was developed to retrieve videos by searching for spoken content and visual cues [35]. The video retrieval interface displays a bar-graph to indicate the relative correspondence of search queries to videos. The emphasis was on text because video cues were not always good discriminators of information. Similarly, a presentation viewer utilised significant changes in presentation video streams to segment audio for navigation [37]. The presentation viewer accompanies each segment with a collection of keywords which appear in the respective segment. The Ferret browser performs advanced video analysis to synchronise physical events with the spoken words [89] and allows users to customise the interface with information he or she deems relevant.

Although video is used for navigation, particularly to display structure, the audio transcript is consistently emphasised and used for search and navigation tasks. Furthermore, although video Summarisation systems do exist, such the video tapestry [9], the display of video does not seem to help summarise the audio content of the media. I believe that video is useful for navigation, but the bulk of data for domains such as lecture recordings are found in the audio stream. Interfaces which do support video for lecture recording domains must focus navigation on audio content.

It is clear from the above examples of work in the literature that existing tools do not yet support the Section Selection, Fact Finding, and Summarisation tasks in equal and sufficient measure without the use of external cues which require additional effort to record. I propose a new system and user interface which supports Section Selection, Fact Finding, and Summarisation with only the audio stream. The system will be simple to deploy because it only requires recorded audio and the interface will be simple to consume due to the utilisation of mature Information Visuali-

sation techniques. Furthermore, I am to support Section Selection not by changes of topic or acoustic qualities alone, but by changes of both topic and acoustic condition together. A multimodal approach to Section Selection may identify both topically and aurally distinct sections for effective audio retrieval.

In the following section (Section 2.3), I provide an overview of the literature on text visualisation. Text visualisation is a field close to audio navigation because text-based methods can be applied to the extracted audio transcripts.

2.3 Text Visualisation

Text visualisation is the visual display of text documents for tasks such as producing summaries, highlighting main topics, identifying significant themes, displaying search results, and comparing documents for efficient assimilation by users. Text documents can range in size from small, such as abstracts of articles, to very large, such as entire books or collections of news articles spanning multiple years. Large amounts of text, particularly unfamiliar text, can be difficult to comprehend quickly due to the amount of word content and complex relationships between passages. Text documents may be structured with headings (such as news articles) or may be unstructured (such as email). In particular, text transcripts extracted from audio (be it via speech recognition or manual dictation) can be regarded as a form of unstructured text. Unstructured because there are not guaranteed to be any clear headings or topic boundaries in free-form (spontaneous) speech. Therefore text visualisation, although not always targeted to spontaneous speech, offers techniques which may be relevant to the visualisation of speech audio. I believe two categories of text visualisation are particularly relevant to audio navigation: static and temporal text visualisation.

Static text visualisation methods focus on visualising common words

and their relationships to build an overall model of a text document, typically by examining word frequencies. Examples include DocuBurst [22], TextArc [72], and Word Clouds such as Wordle [86]. Visualisations which visualise word frequency make the assumption that words which appear most frequently in a document (excepting stop-words such as "a" and "the") represent the most important information and, in-turn, produce good overviews of documents. Words which occur more frequently are displayed in larger text than words which occur less frequently. Adjusting word size has been found to improve recall [77]. As far as document contents are concerned, the usefulness of displaying words in isolation is limited. Isolated words can appear ambiguous and can be interpreted incorrectly without an understanding of the context they are used in.

While Word Clouds such as Wordle typically do not mitigate contextual issues, DocuBurst and TextArc do. DocuBurst visualises word hierarchy with the hyponymy ("is-a") relationship to disambiguate word meaning whereas TextArc displays words near the passages of text in which they appear. A related static visualisation called The Word Tree [88] displays words and phrases in context by surrounding selected words and phrases with neighbouring sentence fragments. A tree structure is used to represent all possible sentence fragments as branches which allows frequency to be displayed as the number of branches rather than the size of the text. Unfortunately, the Word Tree does not highlight important words and phrases. Users must perform search queries to explore the text document which makes the Word Tree difficult for those unfamiliar with the text. Nevertheless, static text visualisations can still be misleading. For instance when word size is augmented, longer words can appear larger than smaller but more frequent words and topic trend is not visualised which can impact a document's relevance against users' goals.

In contrast, temporal text visualisations display patterns and trends over time in either a single text document or across multiple text documents. TileBars [40] visualise term frequency, term position, and term

overlap for search queries in individual text documents. Documents are segmented and segments are displayed as cells in a table. Segments are highlighted proportional to how well they match the user's search query. Word Clouds can also be used for displaying discrete changes over time [26] by displaying multiple Word Clouds where the positions of words stay relative to each other between Word Clouds. However effective TileBars and Word Clouds can be for displaying document trends, discrete changes require more work to assimilate by users than changes that are displayed as continuous (despite the underlying changes being discrete).

ThemeRiver [38] and TextFlow [25] use the river metaphor to display trends in a collection of documents as a smooth progression rather than discrete changes. The rivers are made of multiple layers which widen where topic activity grows and narrow where topic activity shrinks. Although visually pleasing, analysing topics which are not popular is difficult because the respective layers can appear too thin to be legible. Topic interaction can also be important for text analysis. Unlike ThemeRiver, TextFlow visualises topic splitting and merging behaviour to provide extra depth to the information provided which can help users to analyse a wide range of topic behaviour.

Hence, just as text visualisation methods are used to support understanding text, text visualisation may also be applicable to support understanding audio. The following section will briefly discuss what information text transcripts of speech audio contains and how they can be procured.

2.4 Source of Speech Transcripts

An important concept for speech audio retrieval is the origin and role of the speech transcript. The quality of a transcript can affect the performance of any speech retrieval system. A transcript is a text document which contains the words spoken. For speech audio retrieval the transcript

additionally contains when the words were spoken so that the transcript text can be time-aligned with the audio during playback. Speech retrieval systems commonly use transcripts where each sentence is marked with a start and end time. However, other applications may use more course (paragraph level) or more fine (word or syllable level) granularity where needed.

There are three common methods for producing speech transcripts: manual transcription, Automatic Speech Recognition (ASR), and ASR with manual correction. Manual transcription involves the effort of a real person to listen to the audio and copy the words spoken into text. ASR uses a computer system to automatically produce a text transcript from audio. ASR is trained on sample audio recordings and uses machine learning algorithms and probability techniques to create a best guess of what was said, typically with errors. ASR transcripts can be manually corrected to remove any errors that may have occurred.

Manual transcripts are regarded to be 100% correct, while ASR transcripts typically contain errors. Correct transcripts are easier to read and hence are more easily consumable. In contrast, the errors in ASR transcripts can make reading more difficult or even impractical if too many errors occur. In addition, manual transcripts may contain punctuation to assist in their interpretation. Automatically inferring punctuation in ASR is a difficult task, and one that is generally avoided. Tools which utilise transcripts for speech retrieval must use transcripts with few errors or else they may present illegible information to the user. Fortunately, properly configured ASR systems can produce accurate transcripts which can be understood without manual correction.

Manual transcripts are, however, more time consuming and more expensive to produce because they require a person to perform the transcription. Manual transcripts may be more suitable for frequently accessed recordings while ASR transcripts, which are faster and cheaper to produce, could be applied to a large collection of existing audio recordings.

The following chapter will go on to describe how audio lectures are used and perceived by students before discussing the target users of my visualisation in Section 3.3, and the requirements of my visualisation in Section 3.4.

Chapter 3

System Requirements

I describe the requirements for SpEx and the motivation behind the requirements in this chapter. First, I discuss my research methodology in Section 3.1. I follow by looking to the literature to gain an understanding of how students perceive and use recorded lectures at university in Section 3.2. I then build a model of target users in the form of personas to guide the design of SpEx in Section 3.3. Finally, I build the requirements for SpEx in Section 3.4.

3.1 Research Methodology

My research methodology started with a definition of requirements. I began with an investigation of students, the target users for my system. The investigation took the form of a short literature review and informal interviews with colleagues. The target users were modelled as *personas*, fictional characters with the goals and motivations of real users. I followed the personas with research into common audio retrieval tasks. The tasks and the personas guided the design of SpEx.

The actual system was developed with an iterative cycle. Each cycle involved design and implementation before showing the result to my colleagues for feedback. The feedback I received influenced the design of the

following iteration. The user interface was additionally investigated by a user study where lecture and presentation audio was displayed to a group of mostly university students — the target audience. The user study was designed to measure how users perceived the system and by extension how well the system was suited for everyday use.

I discuss the investigation of students in the following section.

3.2 How Students use Lecture Audio

It is important to discuss how lecture audio has been used in educational settings. An understanding of students' experience when using lecture audio recordings and their expectations of the role of the lecturer helped me to form the models of target users in Section 3.3. The models of target users served to provide reasoning to the design choices I made in the development of SpEx.

A survey of undergraduate student opinions regarding recorded lectures found that students seek to gather as much lecture material as possible, be it lecture slides, notes, or related material, and would make use of lecture recordings given their availability [69]. In a similar study, university students ranked easy technical handling with high importance after being exposed to lecture audio in a course [31]. The audio *Podcast* has been used to easily distribute recorded lectures to students. Podcasts are a push-based medium which allow users to subscribe and automatically download the latest content. Push-based is a quality of podcast distribution which differs from normal, pull-based, web traffic [8] which may be used to distribute digital lecture resources. Thus, audio and video podcasts have been an effective method to consume lecture recordings with little effort on the part of the podcast subscriber (the student in this case).

Once procured, students more often listen to lecture recordings on their PC than on a portable device [69, 32]. It is thought that PCs are used to listen to lectures because students like to mimic a lecture setting when

revising from recorded audio. For instance, students have said they follow their notes when listening and perform no parallel tasks [32]. Listening to lectures on a PC can be an effective method of study. Students tend to listen to lecture recordings when preparing for assessments, but prefer not to listen to entire lectures. Students desire to search for specific sections of audio rather than listening to entire lectures [69, 32, 13]. Sections of audio are commonly revisited to recap difficult topics.

When judging the topical relevance of a lecture recording, students place a high premium on the title of a lecture followed by the summary provided. Students would only listen to the audio if they were confident from the title and summary and if they knew the length of audio they must listen to is short [53]. When unsure, the first few sentences of the speaker were usually enough to make a decision. Thus, metadata is shown to be important because students cannot easily skim an audio recording like they can with a text document. Metadata can provide a reliable and concise description of audio.

One frustration students mentioned about lecture recordings was the absence of cues in the audio about which parts of the lecture material were discussed and having little information about any whiteboard/blackboard notes their lecturer created [69]. Students suggested that their lecturer be more explicit about any material being referred to so that following lecture notes would be less difficult. Both the study habits of students as well as the teaching habits of lectures must change to support improved learning form lecture recordings.

Listening to specific sections of audio while following lecture notes is thus a common strategy students use to prepare for assessments when they have access to lecture recordings. The following section describes the models of users built with this understanding of students' usage of lecture recordings.

3.3 Model of Target Audience

I developed a model of target users in the form of personas to ensure SpEx would successfully address the goals of its users. Personas are common practise in software development. Each persona is a fictional character that embodies the goals and motives of a *group* of real target users [23]. Personas are written in the form of fictional characters to provide empathy.

The presence of user understanding and empathy helped to ensure that my design and interaction choices would directly add value to users. Personas reveal the motives and frustrations of real users based on observed behaviour to help guide design and implementation decisions. By helping to select and design features, personas not only saved my development time, but ensured my developed features were focused to solve the goals and needs of real users first and foremost. Furthermore, I could consider my personas when designing my user study for SpEx. I could ensure my participants were similar to my persona demographics and ensure the tasks I asked participants to do corresponded to my persona goals.

My personas were created from the characteristics of audio listening behaviour (prior section) and informal discussions with my colleagues. My colleagues were part of my target audience. Three personas were created: two primary personas of students and one secondary persona of the lecturer. The goals of the primary personas take precedence. The personas were structured with three parts. First, a brief description of each persona which highlighted real characteristics in a fictional scenario. Second, a list of six goals to highlight each persona's ambitions. Two *experience goals* described how each persona liked to feel, three *end goals* described what each persona wanted to accomplish, and one *life goal* described what motivated each persona's end goals. Third, a series of scenarios to describe how each persona would use my system. Scenarios were written in pairs: a context scenario which described the high-level use of a system before any interface design, and a key path scenario which described the interaction with

the design in mind. Goals and scenarios helped me to understand what each persona wanted. The personas I developed are described next. [1]

[1]Persona silhouettes courtesy of user shokunin from `http://openclipart.org/`. Licensed under Public Domain Dedication.

3.3.1 Primary Persona 1: *Jack*, a Student

Jack is a twenty year old university student, originally from Auckland (New Zealand) but moved to Wellington (New Zealand) to study. Now in his second year of university, Jack is accustomed to university life. Jack keeps a timetable of lecture and tutorial sessions to go to each day and spends much of his time doing assignments with friends at university or studying alone at home. Jack loves to spend time with his friends and play video games and he has been found to skip lectures to do so. Although not the brightest student, Jack manages to pass his courses and already has plans for a holiday after university.

Jack loves his technology, often taking his laptop and smart phone with him. Jack also loves listening to his music on his way to and from university. He does his assignments on his laptop, where he also stores all his course material, such as lecture notes and slides, for exam revision. Jack receives lecture notes from his courses and downloads the recorded lectures where available. But he gets frustrated when he receives course content late. Jack is always looking for ways to complete his assignments quickly, so he can have more free time (to which he finds he has less of these days). Jack opts to consume the necessary facts rather than study the subject in detail. Jack wants to make skimming his notes and reviewing unclear or missed portions of recorded lectures easier.

Jack's Goals

Goal 1 Enjoy his time at university (experience goal).

Goal 2 Feel knowledgeable (experience goal).

Goal 3 Quickly find important information in course material (end goal).

Goal 4 Recap missed lectures (end goal).

Goal 5 Pass his courses (end goal).

Goal 6 Get a degree (life goal).

Jack's Scenarios

Context Scenario 1. Jack's at home working on an assignment for one of his courses. The assignment is due soon, and Jack's almost done. There are just a couple questions where he is not familiar with the content. Jack accesses SpEx on his laptop. SpEx loads quickly, which is good for Jack because his laptop is old and in need of replacement. A visual interface is presented, which makes Jack's task easier because information can be consumed visually rather than aurally. Jack performs a search for his content and analyses the results to see which ones are most relevant. Jack listens to candidate results until he finds the information he wants.

Key Path Scenario 1. When SpEx loads, Jack is presented with a screen describing the content of the lecture. Jack looks at his assignment question, identifies some key words, and enters those words as a search query. SpEx marks where the search terms occur in the audio . Jack looks at the results to see which parts of the lecture mention the term most frequently. Jack moves the audio position precisely over the part of the audio he is interested in and listens.

Context Scenario 2. Jack has gone to have some fun with his friends. But he missed one of his lectures to do so. The next day, Jack retrieves the lecture notes online to learn the lecture's content. Unclear about some information, Jack launches SpEx and manually locates where the lecturer is talking about the information. Jack listens to those portions of the lecture.

Key Path Scenario 2. When SpEx loads, Jack is presented with a screen displaying the main topics of the lecture. Jack visually scans the topics for any that might relate to unclear portions of his lecture notes. When Jack finds a relevant topic, he moves the audio position and listens to that portion of the audio. The audio control clearly highlights segment boundaries, which Jack uses for positioning the audio.

3.3.2 Primary Persona 2: *Amy*, a Student

Amy is a twenty-one year old university student born in Canberra (Australia). She moved to Wellington (New Zealand) to start her degree and she's now in her final year of study. Amy has become good at managing her time, and she's able to keep up with going to lectures, attending tutorials, and studying for assignments and exams. A high achiever, Amy attends all her lectures and manages her study time well. But she also has a social life and enjoys spending time with her friends.

Although not an expert with technology, Amy is comfortable using her laptop to write her assignments and to listen to her favourite music. Amy is more than capable of browsing the internet and staying in touch with others online. She abandons her laptop to take lecture notes on pen and paper, so she isn't reliant on when her lecturers release course material online. However, Amy does like to listen to recorded lectures and expects recorded lectures to be available promptly. One of her main strategies while studying for assignments and exams is listening to lectures to revise unclear material. To comprehend the information, Amy prefers to listen to an entire subject within a lecture in detail, rather than find individual facts. But Amy finds it difficult to identify which parts of a lecture recording correspond to unclear portions of her notes.

Amy's Goals

Goal 1 Feel in control of her time when studying (experience goal).

Goal 2 Feel knowledgeable (experience goal).

Goal 3 Quickly find important information in course material (end goal).

Goal 4 Recap unclear portions of lectures (end goal).

Goal 5 Achieve good grades in her courses (end goal).

Goal 6 Get a degree (life goal).

Amy's Scenarios

Context Scenario 1. Amy's in the library, revising for her exam. She finds that there are some parts of her notes which are not clear, some even incomplete. Amy launches SpEx on a university computer at the library and finds the lecture she is interested in. A visual interface is presented, which makes Amy's task easier because information can be consumed visually rather than aurally. She finds the topic she is interested in and listens to that portion of the lecture. She then spots topics she had completely omitted from her notes. Amy then updates her notes.

Key Path Scenario 1. Amy navigates to the web address of her course website and launches SpEx for the appropriative lecture. She does not have to install any software on the computer because SpEx is entirely web-based. SpEx displays topics in segments. Amy scans the topics to find audio segments related to what she is looking for. For each segment she is interested in, Amy will use the audio controls to navigate to the beginning of the segment. The audio control clearly highlights segment boundaries, which Amy uses for positioning the audio.

Context Scenario 2. Amy walked away from a lecture feeling unsure and puzzled, she didn't quite understand what her lecturer was teaching. Nor did her friends. During the weekend, at her friend's house, Amy and her friend launch SpEx to try and understand the lecture. They identify where the major and minor topics are spoken and quickly listen to corresponding short portions of the lecture. Eventually, they build a better understanding

of the lecture. To save time, they skipped the administration details the lecturer gave in the middle of the lecture.

Key Path Scenario 2. Amy and her friend launch SpEx on the appropriate lecture. They are greeted with a display presenting the main topics of the lecture. They can also see the minor topics of the lecture, though not as clearly. After selecting interesting topics, SpEx displays where the topics occur in finer detail. They could see regions where the topics were most prominent, and focus on those regions. They could navigate the audio with the audio controls to the onset of each prominent occurrence of a desired topic.

3.3.3 Secondary Persona 1: *Peter*, a Lecturer

Peter is a forty-one year old university lecturer who lives and teaches in Wellington (New Zealand). Although originally from Ireland, Peter moved to Wellington with his family. As an educator, Peter organises and runs lectures for undergraduate courses. Peter compiles and distributes lecture material such as lecture notes, assignments, and revision material. Peter also supervises his postgraduate students and attends staff administrative meetings. What free time he has, he gives to his family. Committed to providing his students with a quality education, Peter has organised his university's technicians to record his lectures. Peter gets the recording from the technicians after each lecture and uploads it to his course website for students to download as a plain audio file.

Peter is quite knowledgeable with the university software systems, but does not have a lot of spare time to annotate the recordings he puts online. A simple title and description suffices. Peter quickly gets frustrated with software he believes is difficult to use and he is quick to try alternative solutions. To improve the learning of his students, Peter wants a way to let students more easily use lecture recordings as part of their study.

Peter's Goals

Goal 1 Don't be overburdened with work (experience goal).

Goal 2 Proud of his students' quality of work (experience goal).

Goal 3 Allow students to make good use of lecture recordings (end goal).

Goal 4 Make older lecture recordings available to students (end goal).

Goal 5 Plan upcoming lectures (end goal).

Goal 6 Raise his kids (life goal).

Peter's Scenarios

Context Scenario 1. Peter is busy in a lecture theatre preparing his lecture. Peter sets up his slides and makes sure the display is working. Meanwhile, the IT technicians set up the lecture theatre to record and transcribe Peter's voice. Peter ties a small microphone to his collar to do so. The microphone is recording his voice. Peter proceeds to give his lecture when he is ready. At the end of the lecture, Peter puts the recording and transcript onto his computer when he has time. Peter uploads the audio recording and transcript to the university system to be visualised by SpEx and appear on the course website.

Key Path Scenario 1. Peter uploads the audio recording and transcript of his lecture to the university system. The university system uses a separate application to pre-process the audio and transcript into segments and identify the main topics. The audio, segments, and topics are fed to SpEx for consumption via an intermediary file.

In addition to the three personas above, I developed a list of requirements based on audio and text information retrieval in the literature. I discuss my requirements in the following section.

3.4 Requirements of SpEx

Two significant sources influenced my design of SpEx: The personas in Section 3.3 and the following models of user tasks in the literature.

I focused on the primary student personas, Jack and Amy, to influence my design decisions of SpEx. Lecturers, such as Peter, may create the content, but it is the students who would most frequently consume and hence require the aid of SpEx. By targeting students foremost, I could ensure the greater acceptance of SpEx. To this end, I designed SpEx so that Jack and Amy could find relevant information in recorded lectures methodically and without the trial-and-error techniques of finding information in traditional audio controls. I intended to provide random-access to any position in the audio and develop SpEx so that students who hadn't listened to a lecture (like Jack) could understand the content and students who have listened to a lecture could more quickly revise the content. To support these requirements, I utilised and modified an existing audio retrieval task taxonomy to provide formal and fine-grained objectives.

I made use of Whittaker et al.'s audio information retrieval task taxonomy [93]. The task taxonomy consists of three tasks which were developed from their earlier work [94]. The task taxonomy is intended to outline the most common intentions for audio retrieval tasks. The task taxonomy is as follows:

- **Relevance Judgements:** Determining whether an audio recording is relevant to the user's search goal. For example, determining which one of five lectures is most relevant.

- **Fact Finding:** Extracting a specific piece of information from an audio recording. For example, identifying who composed a piece of music from a recorded discussion of musical history.

- **Summarisation:** Producing a summary of the content of an audio recording. For example, describing what a recording was about.

Dufour et al. [29] modified the task taxonomy for their user study. Instead of determining which audio recording is most relevant, the Relevance Judgements task was modified to rank audio recordings from most to least relevant. The change of task provides a more fine grain understanding of usage behaviour, especially when candidate audio recordings are similar or partially relevant to the search goal. For a separate kind of user study on chapter search in computer-based textbooks, not an audio search task but nevertheless equivalent in terms of the search behaviours of users, a similar but separately developed workflow was used consisting of (in order) [28]:

1. **Goal Formulation:** Understanding what is to be achieved.
2. **Category Selection:** Filtering what is relevant from what is not.
3. **Information Extraction:** Extraction of relevant data or determination that data is irrelevant.
4. **Integration:** Assimilation of information with what has already been learned in the process.
5. **Recycling:** If more information is needed, the process repeats from Category selection.

The workflow repeats until the user's goal is achieved. It is not hard to see that the elements of the workflow can be mapped to the audio retrieval task taxonomy. Goal Formulation and Category Selection are equivalent to Relevance Judgements, Information Extraction and Integration are equivalent to Fact Finding and Summarisation, and Recycling highlights the iterative process of information search.

Furthermore, Besser et. al. [13] identified "Quotation" as a common audio podcast search goal by university students. The Quotation search goal was regarded as a search for a specific piece of information and also a search for an audio segment. Quotation searches therefore fall into the Fact Finding category.

Hence, other researchers identified similar tasks to the task taxonomy of Whittaker et. al. I therefore took the task taxonomy as one component of the requirements for SpEx with two modifications. First, I added a Section Selection task where relevant segments of audio were identified. I considered Section Selection as important because I believed students such as Amy would be more likely to listen to entire topics for revision rather than finding specific facts. For instance, Amy may want to recap the content of an entire lecture slide rather than searching for each fact she needed individually. Lectures typically cover multiple, albeit related, subjects and students like to listen to only the relevant segments of lecture recordings. Therefore, it is important to cater for Section Selection in order to increase the efficiency of lecture search tasks. Second, I removed the Relevance Judgements task because I believed that within-lecture navigation was a more sought-after task than finding which lectures are relevant. I believed that Jack and Amy would be familiar with their course syllabus and hence understood which lecture would be most relevant to their needs.

With Section Selection added and Relevance Judgements removed, my final task taxonomy is as follows:

T1 **Section Selection:** Identification of relevant segments of the audio.

T2 **Fact Finding:** Extracting a specific piece of information from the audio.

T3 **Summarisation:** Producing a summary of the content of an audio recording.

Additionally, SpEx should support multiple methods of audio navigation. Two common trends in the literature of speech navigation interfaces have been a lack of visual structure of the audio and a lack of indices into the spoken content. Consequently, information such as when a speaker is talking [53], identifying salient topics [28], and distinguishing relevant segments in the audio [64, 29] have all been mentioned as sources for improvement in prior work.

I required SpEx to support three methods of location: *searching*, *scanning* and *browsing*. Supporting searching, scanning, and browsing would promote SpEx to display adequate structural information and indices for locating topics and specific facts. Searching begins by formulating a search query (candidate words for example), analysing the retrieved results, and, if no result is suitable, refining the search query [39]. I expected Jack and Amy to search when they had a precise understanding of what they wanted. Scanning is the process of looking at each item in detail, one by one, until the desired item is found. In other words, a methodical and careful survey with a known goal [10]. I expected Jack and Amy to scan when they knew what they wanted, if they knew where to look, or if a search query would be difficult to form. On the other hand, when browsing the goal is not fully understood beforehand. Browsing is an iterative process which consists of randomly glimpsing across all available choices, selecting a candidate object, further examining the object, and finally acquiring or abandoning the object [10].

In addition, users should be able to navigate by the textual cues of the spoken word as well as the acoustic cues of the audio. Navigating by both textual and acoustic cues will allow users to locate a wide range of information. Besides using speech to locate informative content, acoustic cues such as background noise may be memorable events that users should be able to exploit for navigation.

I also considered non-functional requirements. That is, requirements which affect the operation rather than the behaviour of a system. The following non-functional (NF) requirements also impacted the design of SpEx:

NF 1 **The feature processing pipeline must be configurable.** It must be possible to replace the feature processing methods required for SpEx. Alternative feature processing methods may provide more accurate features for a specific lecture or task domain.

NF 2 **SpEx must not require users to install additional software.** I consider additional software to include any software that does presumably come pre-installed with a computer purchase. Installing software may be barrier to users with a low level of computer knowledge or whose computers do not meet the requirements of the required software.

NF 3 **SpEx must run on a variety of hardware and operating-systems.** Computers come in different forms (commonly desktop and laptop) and run different operating systems (Windows, Apply OS X, and Linux for instance). Students should not be restricted from SpEx based on their choice of computer.

NF 4 **Lecture recordings should process in under five minutes.** University and personal computer resources are limited. Lecture recordings should be processed in under five minutes to ensure quick consumption by users.

NF 1 was intended to allow universities and content creators to implement their own audio feature extraction system. One more suited for their own domain to produce more accurate results for users. NF 2 and NF 3 were designed to support Jack and Amy. Jack and Amy may have access to multiple computers, each configured differently. It was unreasonable to hinder their learning by restricting access of SpEx to a subset of their computers. Jack and Amy also wanted to access recordings quickly, hence the need for NF 4. Waiting five minutes is tolerable to Jack and Amy: five minutes is equivalent to downloading a file or listening to a song.

In light of the requirements discussed, personas, task-taxonomy, and methods of navigation, I devised a concise table to specify the main functional requirements for each persona. The requirements are displayed in Table 3.1. For clarity, non-functional requirements will be treated separately.

Table 3.1: Persona Functional Requirements.

	Basic	Section Select		Fact Find		Summary	Distribution	
	Listen	Locate	Search	Locate	Search	Locate	Record	Upload
Jack	○			○	○			
Amy	○	○	○	○	○	○		
Peter							○	○

To support these broad requirements, Table 3.2 presents the specific features that must be implemented for each persona. The requirements were split into five categories: Basic, Section Selection, Fact Finding, Summary, and Distribution requirements. The Basic requirement was to listen to lecture recordings. Section Selection, Fact Finding, and Summarisation requirements included 'locate' and 'search' to distinguish between manually locating information (scanning and browsing) and using search queries. The Basic, Section Selection, and Fact Finding requirements characterised Jack and Amy who wished to consume lecture recordings. Jack would favour Fact Finding due to his desire to complete his work quickly. I included finding acoustic events to the requirements so that useful non-spoken cues could be used for information retrieval. Meanwhile, Amy would prefer a more in depth revision by listening to entire topics. Amy would also like to summarise lectures, hence specific requirements were set to support Summarisation. The two Distribution requirements characterised Peter who wished to easily record lectures and upload them online promptly for his students to consume. These features will be evaluated at the end of Chapters 4 and 5 to ensure my developed system meets the requirements of my personas. The following chapter will discuss the design of SpEx.

Table 3.2: Detailed Persona Functional Requirements.

Task	Requirement	Jack	Amy	Peter	Supported by
Basic	Listen to audio	○	○		Not yet supported
Section Selection	Display topic structure	○	○		Not yet supported
	Display topic content	○	○		Not yet supported
	Browse/scan for topics		○		Not yet supported
	Search for topics		○		Not yet supported
	Move audio between segments		○		Not yet supported
Fact Finding	Browse for specific spoken information	○	○		Not yet supported
	Search for specific spoken information	○	○		Not yet supported
	Browse/scan for specific acoustic events	○	○		Not yet supported
	Accurately move the audio position	○	○		Not yet supported
Summarisation	Understand surrounding content		○		Not yet supported
	Show regions listened to		○		Not yet supported
	Manoeuvre and listen to audio		○		Not yet supported
Distribution	Record audio with off-the-shelf hardware and software			○	Not yet supported
	Simple configuration when processing audio			○	Not yet supported

Chapter 4

Audio Visualisation

In this chapter, I present an audio navigation system called *Speech Explorer*, abbreviated to *SpEx*. I designed SpEx to address the requirements in Section 3.4 and to complement the desires of the primary personas in Section 3.3. SpEx is designed to support Section Selection by identifying similarities and differences between parts of the audio, Fact Finding by supporting automatic search and manual browsing by users, and Summarisation by distinguishing relevant and irrelevant parts of the audio. SpEx is supported by *TAFE*, a Text and Audio Feature Extractor, which calculated acoustic and text features used by SpEx. TAFE is discussed in Chapter 5.

I begin by discussing the design of SpEx in Section 4.1. Section 4.2 describes the tools I used to implement SpEx, and Section 4.3 compares my final design against alternative designs.

4.1 Design of SpEx

SpEx is a web-based interactive interface which displays the output produced by TAFE. An image of SpEx is displayed in Figure 4.1. I begin by describing why I utilised Word Clouds to visualise the structure of audio recordings in Section 4.1.1. My decision to layout Word Clouds with a Treemap is discussed in Section 4.1.2, followed by how the Word Clouds

were constructed in Section 4.1.3. Section 4.1.4 describes how users could manipulate SpEx and Section 4.1.5 discusses how the Treemap was able to support audio progression. Finally, Section 4.1.6 discusses the tools used to locate acoustic events in the audio.

4.1.1 Representing Topic Structure

Displaying the topic structure of audio recordings was an important design decision. I wanted Jack and Amy to be able locate regions of interest in lecture audio by understanding what was being said and when. In which case, I decided to segment audio into discrete and coherent segments which could easily be assimilated by users. The segmentation process is performed by TAFE and discussed in Chapter 5. Additionally, I concluded that I could use Word Clouds to visually represent the content of each segment. Word Clouds are a text information visualisation technique used to display the most frequently occurring words in a document. An example Word Cloud to visually describe an audio segment is presented in Figure 4.2. More common words are displayed larger than less common words. Word Clouds are used to provide readers with a brief overview of what a document may contain, and have been found to successfully display the segment-level content of audio podcasts [33]. I opted to use Word Clouds rather than displaying large portions of the text transcript because I believed that Word Clouds would clearly display the topic structure of audio. Displaying too much text would force users to take time to skim the content, which would require more time and

Figure 4.2: Example Word Cloud. Larger words occur more often than smaller words.

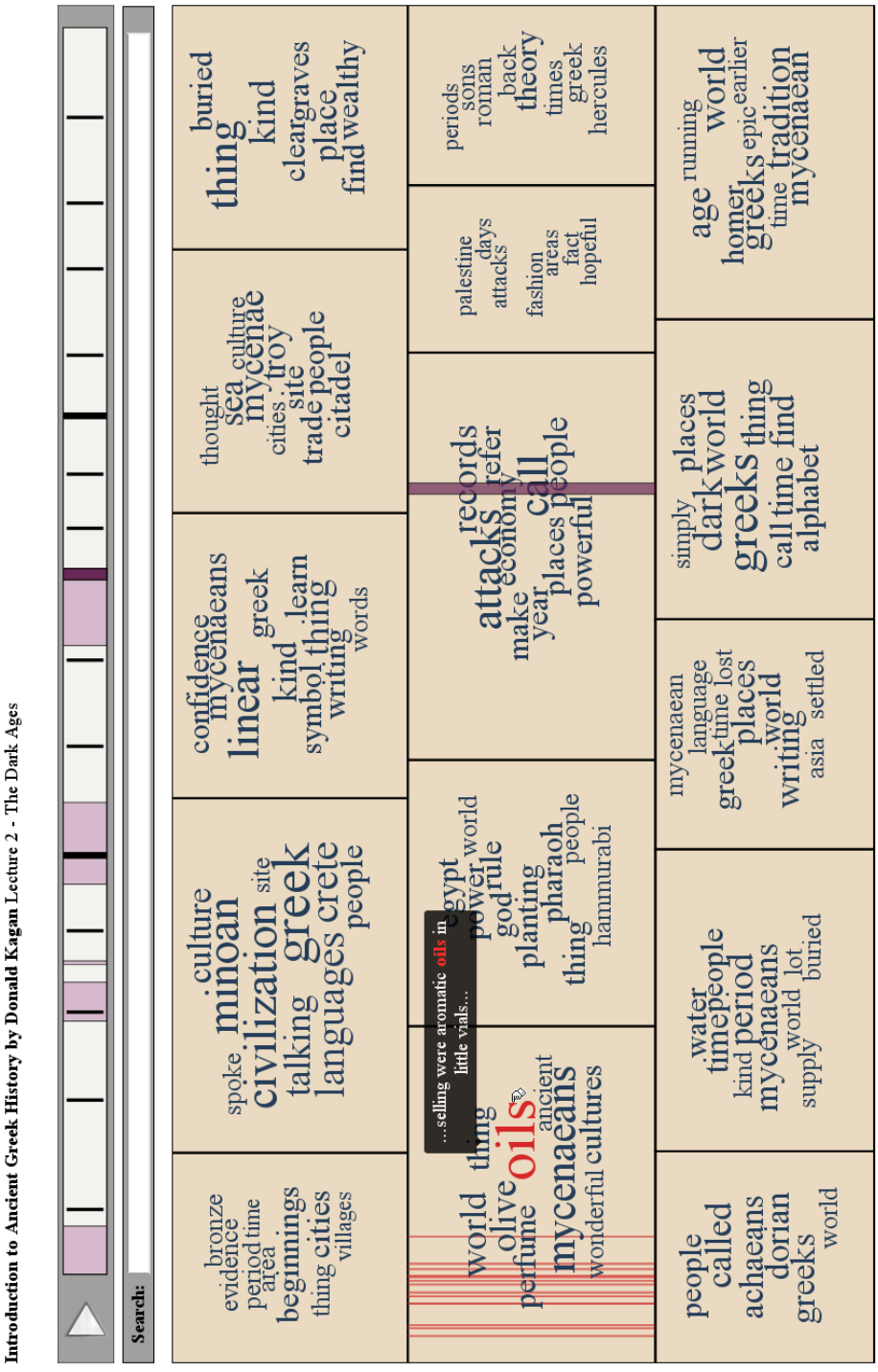


Figure 4.1: SpEx interface for audio retrieval.

effort than identifying Word Cloud words.

Audio recordings could be segmented and Word Clouds could be generated from the text within each segment. Besides presenting users with what a segment may contain, similarities and differences between segments would be visible to guide users away from irrelevant segments and towards alternative and potentially relevant segments. For the Section Selection and Summarisation tasks in particular, the display of relevant segments may guide users to further potentially relevant information.

4.1.2 Layout of Audio Segments

The success of SpEx would depend on the underlying choice of visualisation used to support segment-level Word Clouds. While a good visualisation can present information succinctly, a poor one can hide the important details of the data and make important patterns difficult to see [19]. I identified the *Treemap* visualisation as an ideal candidate to layout segment-level Word Clouds for the visual display of audio structure. Word Clouds fit well with Treemaps because Treemaps are good at utilising screen real estate. Enough so that Word Clouds could be contained with ample room to display multiple words while not appearing cluttered (see Figure 4.1).

A Treemap is a space-filling visualisation which was originally designed to visualise hierarchical information by representing objects as cells (rectangles) [49]. In a Treemap, object quantity maps to cell area and object containment maps to cell nesting. Although cells may be nested, their order is not usually preserved. I used a particular type of Treemap called a *Strip Treemap* [11] which was ordered and therefore displayed segments in the correct temporal order. Figure 4.3 displays an example Strip Treemap in isolation which I utilised for SpEx.

The Strip Treemap algorithm attempted to size and fit all cells such that the aspect ratio of each cell was roughly uniform and hence produced a uniform layout which was easier to comprehend. My Strip Treemap im-

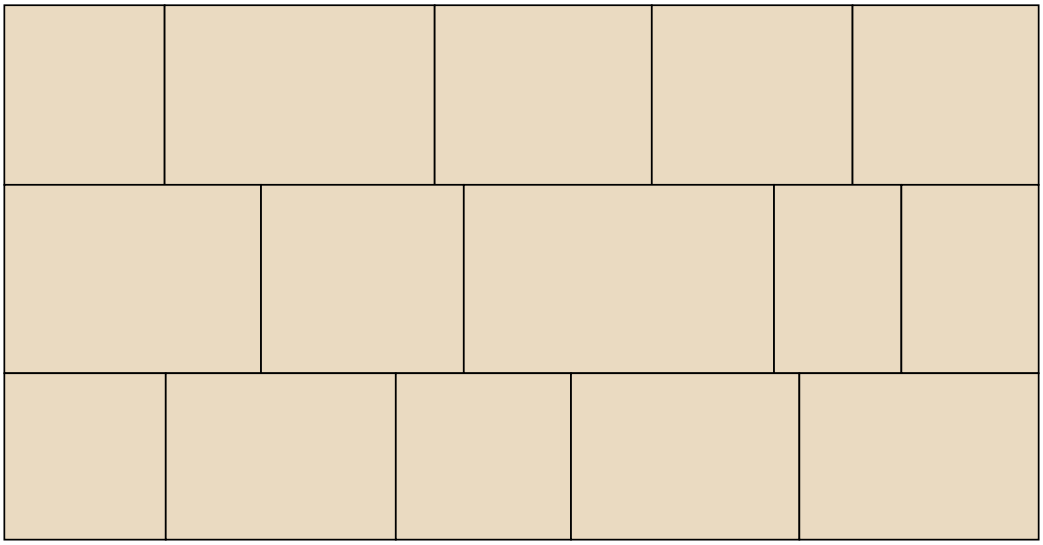


Figure 4.3: Bare Strip Treemap used in SpEx. Each cell is an audio segment. This Strip Treemap is not hierarchical.

plementation did not include forward search [81]. Forward search is used to ensure cells have more uniform aspect ratios. I found that the standard algorithm produced cells that had adequate aspect ratios without the extra computational cost of the forward search method. I mapped audio segments to separate cells in the Treemap where the area of each cell corresponded to the duration of the segment. My Treemap was not hierarchical because my segments were disjoint. Disjoint segments produced a less complicated Treemap at the expensive of revealing less structure about the recorded audio. The segments were ordered in the way English text is read, that is from left to right then top to bottom. I utilised an existing Strip Treemap implementation [81] which I re-wrote in JavaScript.

TAFE segmented audio recordings into exactly fifteen segments. I decided to utilise a Strip Treemap with exactly fifteen segments for two reasons: to ensure consistency between audio recordings which would help users to familiarise themselves with new audio recordings; and because more segments produced a more cluttered interface when segments got

smaller, while fewer segments produced an interface which was too sparse and hence did not make efficient use of space. An average segment size of four minutes (sixty-minute lecture divided into fifteen segments) was roughly the length of a single song that Jack and Amy may listen to. In which case the roughly four minute segment durations are long enough to provide some meaningful information while being short enough to be listened to without becoming impatient.

As a consequence of using a Strip Treemap to layout audio segments, scanning and browsing could be well supported. I expected the Strip Treemap to support scanning by allowing users to methodically analyse each segment in order. Furthermore, I expected the Strip Treemap to support browsing by displaying reasonably descriptive Word Clouds in a clear and unobtrusive manner. A single horizontal visualisation (such as one long row) could not provide each segment with enough space to contain a Word Cloud. Scrolling could be used, but at the expense of only showing a portion of the audio at any one time, reducing searching and browsing efficiency.

One weakness of the Strip Treemap for audio navigation is a lack of time indices. Minor differences in the height and width of cells are difficult to visually compare. I chose to display no time indices to simplify the interface from visual clutter. Consequently, listening to audio at a desired time is not easy because of the difficulty of estimating segment duration. SpEx enforces navigation by the content of audio, in which case mental models of how to find information must be re-learned.

Section 4.1.3, below, will go on to discuss how Treemap segments were overlaid with Word Clouds from the text transcript to assist navigation.

4.1.3 Displaying Segment Topics

I placed a Word Cloud over each segment. Each Word Cloud displayed up to ten of the most frequent words contained in its segment. I decided

that fewer words displayed too little information for a one hour recording while more words made the interface cluttered. When few words reoccurred, only words which appeared more than once were displayed. Words which appeared only once were removed because they were arbitrary choices from the collection of all non-frequent words. Figure 4.4 displays the Strip Treemap overlaid with Word Clouds.

Figure 4.4: Strip Treemap with Word Cloud overlay displaying the most frequent words for each segment.

The size of each word reflected how often the word occurred in the entire audio recording, word frequency. Word frequency was further emphasized by making less frequent words more transparent (and hence less likely to stand out) than more frequent words. Word font size was calculated with Equation 4.1 which produced an almost linear font size progression while slightly reducing the number of very small font sizes.

$$fontSize = \sqrt{\frac{wordFrequency}{maxWordFrequency}} \times maxFontSize \tag{4.1}$$

Where `wordFrequency` was the number of occurrences of the word in the segment, `maxWordFrequency` was the number of occurrences of the most

frequent word across all segments, and `maxFontSize` was a limit on the maximum displayable font size which was set to 50 pixels. Transparency was governed by Equation 4.2 because it only reduced the opacity for the least frequent words. Too much opacity made frequent words difficult to read.

$$opacity = \frac{fontSize}{maxWordFrequency} \tag{4.2}$$

For clarity and aesthetics, words which could not fit into the containing segment were not displayed. The Archimedean spiral layout algorithm was chosen to position the words because it produced Word Clouds which were well centered in each segment. Centered Word Clouds produced gaps between segments which helped to highlight segment boundaries.

Although Word Clouds are generally effective at displaying the main themes of text [77, 33], the meaning of words can often be ambiguous. Words can have different meanings in different contexts. Word Cloud algorithms do exist to place semantically related words nearby [26, 79], though I instead chose to display words in context by presenting the user with the word in a sentence.

Figure 4.5: Pop-up notification displaying the word under the mouse in a sentence to provide context.

Figure 4.5 shows how SpEx presented words in context. When users hovered over a Word Cloud word, a pop-up notification appeared which displayed the first sentence of the segment which contained the word. I selected the first sentence because it may have been the most likely to introduce the meaning of the word. The mouse gesture of hovering over candidate words was selected to allow users to efficiently judge whether or not a word was semantically relevant to their query or not.

The combination of Word Clouds and display of sentences for context served to

aid the Section Selection and Summarisation tasks of my task taxonomy by providing a holistic view of the audio content. Section Selection and Summarisation could be performed by inspecting Word Cloud words and aided by displaying sentences for context. Context sentences could determine whether or not a portion of audio was relevant.

Besides displaying the most frequent words in a segment, the Word Clouds also served to highlight similarities and differences between segments. Users could find similar segments by identifying matching words. For simplicity and ease of use, no extra visual component was added to automatically highlight similar segments. Only when users clicked on candidate words did highlighting occur as discussed in Section 4.1.4.

4.1.4 Navigation by Speech Transcript

With regard to the requirements in Chapter 3, not only did Jack and Amy want to understand the topic structure of audio recordings but access to specific content was required for the Fact Finding and Summarisation tasks. Topic structure was facilitated by the Strip Treemap and Word Clouds mentioned previously. Without more precision when navigating audio, it could be difficult for Jack and Amy to efficiently find their desired information in audio recordings. This inefficiency would only be amplified when several pieces of information were desired or if a particular piece of information was difficult to find. Interfaces in the literature have almost consistently offered a search facility for navigation [91, 74, 78]. But such search queries were useful only for those who fully understood the information they desired. Neither browsing nor scanning were readily catered for. To provide more precise navigation facilities, I designed SpEx to be interactive. I allowed users to click on Word Cloud words to give the display in Figure 4.6.

Two events occurred when a Word Cloud word was selected. First, all matching Word Cloud words would also become selected and, second,

Figure 4.6: State of SpEx when a Word Cloud word ("writing") is selected. When a word is selected, all similar words are also selected and vertical markers indicate where in the audio the word occurs.

vertical Transcript Markers would indicate where the selected word occurred in the audio. Additionally, multiple Word Cloud words could be selected for complex searches and a search facility was available to easily query known words. Transcript Markers and matching Word Cloud words served to help Section Selection by helping users find related information. Each aspect of navigating by words are discussed in turn.

Word Cloud highlighting. When a Word Cloud word was selected, the word would become highlighted by a change of colour. Additionally, every matching word from the other segments would also become highlighted in the same colour. I chose to highlight matching words in every segment to allow users to easily understand which other segments may contain relevant information. Showing related segments supported the Section Selection task by identifying candidate sections, a feature that would be useful for Amy. Selecting a Word Cloud word which was al-

ready highlighted would remove highlighting of the word from all segments. To ensure that words with different suffixes (such as "ing", "ed", and "less" in English) were highlighted together, I made use of the Porter Stemming algorithm [73] to ignore suffixes when matching Word Cloud words. By ignoring suffixes, all words with the same meaning (apart from synonyms) were displayed to the user. I made use of an existing JavaScript implementation rather than creating my own [4].

Multiple highlighted words. Users could highlight multiple words at the same time, each automatically highlighted with a different colour. By highlighting multiple words, users could identify intersections of word occurrences for AND queries or segments which occur with only one highlighted word for AND NOT queries. Users could click on a word again to remove highlighting for the selected word or click on an empty space to clear the Treemap of all highlighted words.

Transcript Markers. Transcript Markers would appear when any Word Cloud words were highlighted and would vanish when their respective Word Cloud words were no longer highlighted. Transcript Markers were vertical lines which indicated the beginning of each sentence which contained a highlighted word (red lines in Figure 4.6). The beginning and end of each sentence was found in the text transcript provided to TAFE. Had I positioned Transcript Markers directly over corresponding words, users may have accidentally listened to the audio after the word occurred or may not have fully comprehended the context the word was spoken in. Transcript Markers facilitated Section Selection by showing which segments contained the desired information and Fact Finding by displaying the position of relevant words. I coloured Transcript Markers in the same colour as the corresponding highlighted Word Cloud word to create a visual link between the marker and the word it marked. I ensured all related words were marked by again applying the Porter Stemming algorithm to

ignore word suffixes.

The use of Transcript Markers had the potential to display too many vertical lines if a word occurred very often or if several words were selected together. Consequently, users may have trouble comprehending the displayed information. I believed that users would deselect words if too many Transcript Markers were present to reduce the impact of information overload.

Search. To support users who fully understood their search queries, I additionally incorporated a search facility. The search facility allowed users to search for words which appeared in Word Clouds. The search facility was located directly above the Treemap and is shown if Figure 4.7.

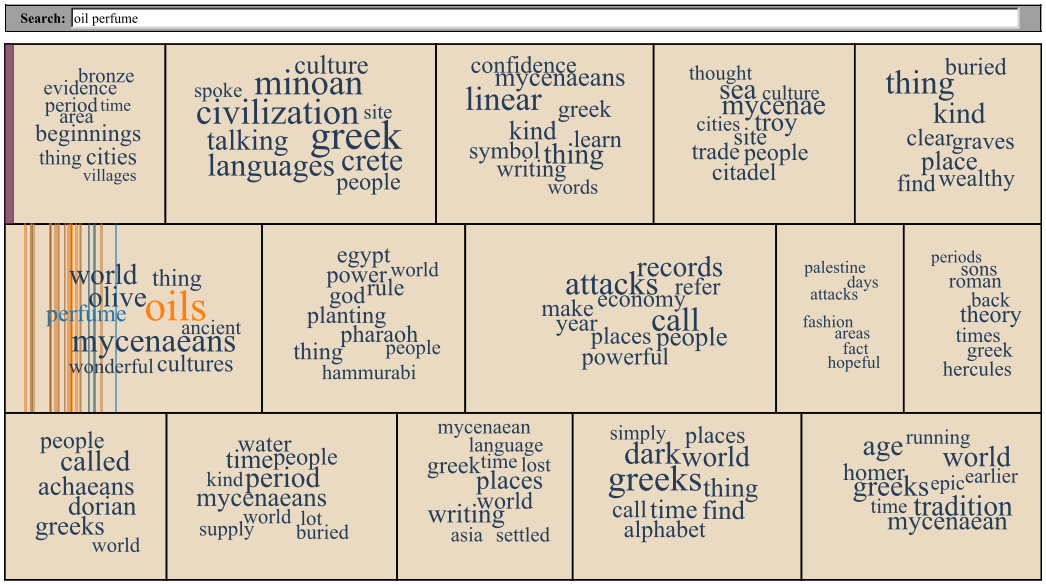


Figure 4.7: Search facility which allowed users to search for words that appeared in Word Clouds. Here, the query "oil perfume" is searched.

Entering a word into the search facility would highlight all matching Word Cloud words and in-turn display the corresponding Transcript Markers. The Porter Stemming algorithm was used to strip word suffixes

so that all relevant words could be retrieved. Similarly, removing a word from the search facility would deselect all corresponding words and Transcript Markers. The search would re-run after any character in the search facility was added or removed. Re-running the search gave users immediate feedback as to whether a search was successful or not. Hence, frustration due to delays could be reduced. A successful search was indicated by highlighting words while an unsuccessful search did not change the state of SpEx. Due to the constantly updating search, warning users of a failed search would introduce numerous warnings which may have been confusing to some.

I made the decision that the search facility could not search for words which did not appear in Word Clouds. Differentiating between Transcript Markers which corresponded to Word Cloud words and Transcript Markers which corresponded to search queries would serve to complicate the interface. For instance, selecting a Word Cloud word which matched a search query would introduce a new visual element to specify that a word was selected twice: once from the Treemap and once from the search. Furthermore, clearing the Treemap of search queries but not selected Word Cloud words could have lead to unexpected behaviour if the two search methods did not adequately distinguish themselves. A simple interface where search and Word Cloud word highlighting were not distinguished reduced cognitive load and hence increased the usability of SpEx. Though there may by lower Fact Finding performance for Jack, I believe Word Clouds and Transcript Markers are adequate for finding most information.

The display of Word Cloud words and Transcript Markers to show where words occurred in the audio, coupled with the pop-up notification to present words in context (Section 4.1.2) worked together to facilitate browsing and scanning by users. By providing navigation at different levels: a rough display of the structure of audio recordings, more precise display of which segments were relevant, and showing users where exactly relevant words occurred would help to guide users to their desired infor-

mation. By displaying information on demand, SpEx could refrain from overburdening users with too much information at once.

Section 4.1.5, below, will go on to discuss how users could manipulate the audio controls as they searched for their information.

4.1.5 Moving the Audio Position

I believed that listening to and navigating the audible audio stream in SpEx was important so that users did not have to synchronise between SpEx and a separate audio player. However, how audio navigation could be integrated with a Treemap is not clear. To my knowledge the literature does not describe how a Treemap can be extended with a progress indicator.

Due to Word Clouds in my Treemap, there was not sufficient space to integrate audio controls onto the Treemap itself. I instead decided to place the audio controls as a separate component above the Treemap with a minimalist and interactive indicator of audio position on the Treemap itself. The audio controls were designed to mimic and behave like a standard audio interface to reduce the learning difficulty associated with SpEx. The audio controls are displayed in Figure 4.8.

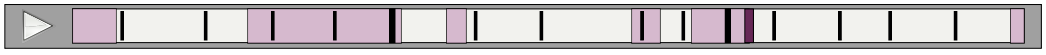


Figure 4.8: Audio Control. Thin black lines indicate segment boundaries. Thick black lines indicate row boundaries. Light purple regions indicate portions listened to. A thick purple line indicates the current play position.

To the utmost left of the control was a Play/Pause button for playing and pausing the audio. To the right of the Play/Pause button was a progress bar which displayed four pieces of information. First, the boundaries between each segment were indicated by thin black lines to help users navigate within and between segments. Second, the boundaries between each row were indicated by thick black lines to warn users where

the audio position would switch rows in the Strip Treemap. Third, the pink regions indicated which parts of the audio had been listened to. Due to the random access nature of the audio, the standard technique of highlighting the entire region up to the current play position is not an accurate indication of how much audio had been listened to. By highlighting only the regions listened to, SpEx assisted users to re-listen or ignore parts of the audio.

Fourth and final, a thick purple bar called the *Thumb* displayed the current play position of the audio track. Users could drag the Thumb to play the audio at different positions or they could click anywhere on the progress bar to instantly reposition the Thumb. Hereafter, the former is referred to as a *drag* operation and the later is referred to a *skip* operation. The Thumb is synchronised with another Thumb which lay on the Treemap itself, as shown in Figure 4.9.

Figure 4.9: Audio Thumb (in purple) over a Treemap segment.

The Thumb on the Treemap could likewise be dragged to move the audio position, but it did not support the skip operation. Clicking on the Treemap itself was reserved for clearing selected Word Cloud words and Transcript Markers. I assumed that clicking on the Treemap to skip the audio would too easily occur accidentally when users attempted to click on Word Cloud words and missed, thus causing annoyance. The Treemap's Thumb could be dragged left and right between segments and up and down across rows. When dragged off the left or right edge of the Treemap, the Thumb would continue its position on the above or below row respectfully. The presence of the Treemap's Thumb allowed users to precisely position the audio to a desired segment or Transcript Marker for efficient information retrieval.

My approach of utilising a standard audio control interface with a minimal Thumb on the Treemap supported audio navigation without cluttering the display. The Treemap's Thumb was synchronised with the Thumb of the audio control to allow users flexibility to adjust the audio play position. Furthermore, the standard audio control interface provided an easy method to learn how to manipulate the audio position while the Treemap's Thumb allowed more experienced users to efficiently manipulate the audio position without leaving the Treemap.

The final feature of SpEx augmented the Treemap with acoustic, rather than textual, features. Section 4.1.6 describes the motivation and design decisions for displaying the loudness and pitch of audio.

4.1.6 Visualising Acoustic Features

So far, SpEx has only visually displayed textual information from the transcript of the audio. The acoustic information of audio can be just as important. For example, prosody can influence the meaning of what has been said, emotion can be conveyed by the tone of ones voice, and background noise can provide cues to important events in the audio. However, capturing information from the tone of voice is a difficult task that is not within the scope of this project. I instead visualised loudness (volume) and pitch (frequency) as an underlay to the Treemap to display tonal and background events. Users could attempt to infer the meaning of the different levels of loudness and pitch depending on the speaker and context of the recording. Figure 4.10 displays the Treemap with the acoustic underlay.

The acoustic underlay was displayed as a bar-graph consisting of fifteen bars for each segment. A bar-graph was chosen because it provided an easily consumable summary of the acoustic conditions of a segment in a well understood format. I chose to display fifteen bars because I believed fewer bars displayed information too coarse to be useful while more bars appeared too complex to be easily comprehended. By creating a sim-

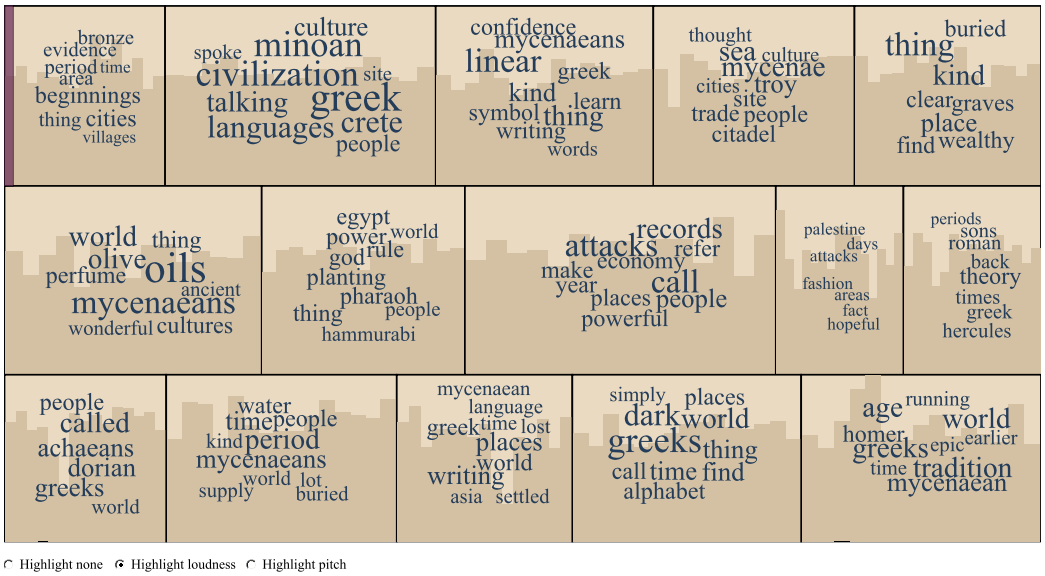


Figure 4.10: SpEx with audio loudness underlay. Three controls at the bottom turn the underlay off, display loudness, and display pitch from left to right respectively.

ple underlay, the information was expected to be easy to understand and not visually detract from the remaining Treemap. Options to change and remove the acoustic underlay appeared at the bottom of the Treemap so they did not become accidentally selected. The underlay was not interactive. Due to the high level of interactivity on the Treemap, making the underlay also interactive could have lead to user error when one operation was performed but an unintended outcome was achieved. I chose a colour which complemented the background colour to ensure that the foreground segment Word Clouds and any Transcript Markers would remain clearly visible. The result of the acoustic underlay was an extension to SpEx which provided an extra dimension of information about the audio to assist navigation when the search query was not textual.

4.2 Implementation of SpEx

I developed SpEx to run in the web browser to satisfy NF 2 (*"SpEx must not require users to install additional software."*). Making SpEx accessible was an important factor of its design. Different students tend to use computers which run different hardware, operating systems, and software. Installing and configuring software is not a familiar process for all students (Amy for example). Running SpEx in the web browser also satisfied NF 3 (*"SpEx must run on a variety of hardware and operating-systems."*) because web browsers were considered ubiquitous. Also, web browsers provide an abstraction from operating system, hardware, and software interfaces.

SpEx was built with HTML5, SVG v1.1, CSS, and JavaScript, all of which run completely in the browser without requirements from other software. Therefore SpEx could be run in a wide variety of computer configurations. The D^3 JavaScript library [15] was used to support the dynamic creation of SpEx in the browser. In which case, Peter could provide Jack and Amy with web address of his lectures. He would not have to provide references to specific software to install or restrict Jack and Amy to their choice of operating system. Consequently SpEx does require an Internet connection which may not always be accessible. I assumed that most students have reliable access to the Internet via their education provider, but I understand that the lack of offline access is a limitation of SpEx.

One important aspect of Information Visualisation and effective user interface design is feedback. When users interact with SpEx, not only should wait time be avoided to reduce user frustration but all interaction must have immediate feedback to confirm to the user that the interaction was successful or unsuccessful. Consequently SpEx pre-processed all of the input from TAFE during start-up. Loaded data are indexed and stored into HashMap data structures with $O(1)$ time cost for retrieval. Therefore SpEx was responsive to user actions despite the amount of data that was processed (which included segments, transcript text, word stemming, and

interface rendering). I noticed no perceivable delay between action and reaction.

Furthermore, besides file hosting, no server-side processing was required to support SpEx. However, to support the user study described in Chapter 6, a server running Apache Server v2.2.15[1] and web.py v0.37[2] (a Python web development framework) was setup to guide users through the tasks required for the user study. SpEx was also built to record all actions performed by users and send those actions to the server for permanent storage. Every invocation of a user interface element (such as clicking a word or searching) was recorded. These recorded interaction logs could be analysed to reveal how users made use of SpEx.

With every component of SpEx described, I believe SpEx has satisfied the audio retrieval requirements of Jack and Amy. Table 4.1 displays the requirements SpEx has addressed. The play/pause button satisfied the Basic requirement to listen to the audio. For Amy (and Jack to a lesser degree), SpEx was able to fully support Section Selection. The Strip Treemap provided an organisation to display discrete audio segments. Word Clouds extended the segments to describe segment topicality. Along with context sentence popups which displayed over Word Cloud words and a Search Facility, SpEx was able to allow the navigation strategies I described in Section 3.4: scanning, browsing, and searching. The vertical markers in the audio control were landmarks to help move the audio position between segments.

Jack's and Amy's Fact Finding tasks were also supported. Browsing and searching for facts were possible. Transcript Markers which appeared on Word Cloud word selection and searches showed precisely where each word occurred. Acoustic filters allowed Jack and Amy to search for acoustic cues. Finally, accurate audio manoeuvrability for Fact Finding was supported by an audio control on the Treemap itself.

[1] Apache Server website: `http://httpd.apache.org/`
[2] web.py website: `http://webpy.org/`

Each Summarisation task was also supported. Understanding the surrounding content was supported by Word Clouds and Transcript markers. Showing regions listened to was supported by highlighting played regions in the audio control. Lastly, manoeuvring the audio was supported by the Audio Control and the Treemap thumb.

However, the Distribution requirements for Peter have not yet been addressed. SpEx is agnostic to how the audio was recorded, only that the audio is of the correct file type. SpEx also does not process any audio. Processing audio is the role of TAFE. TAFE is discussed in Chapter 5.

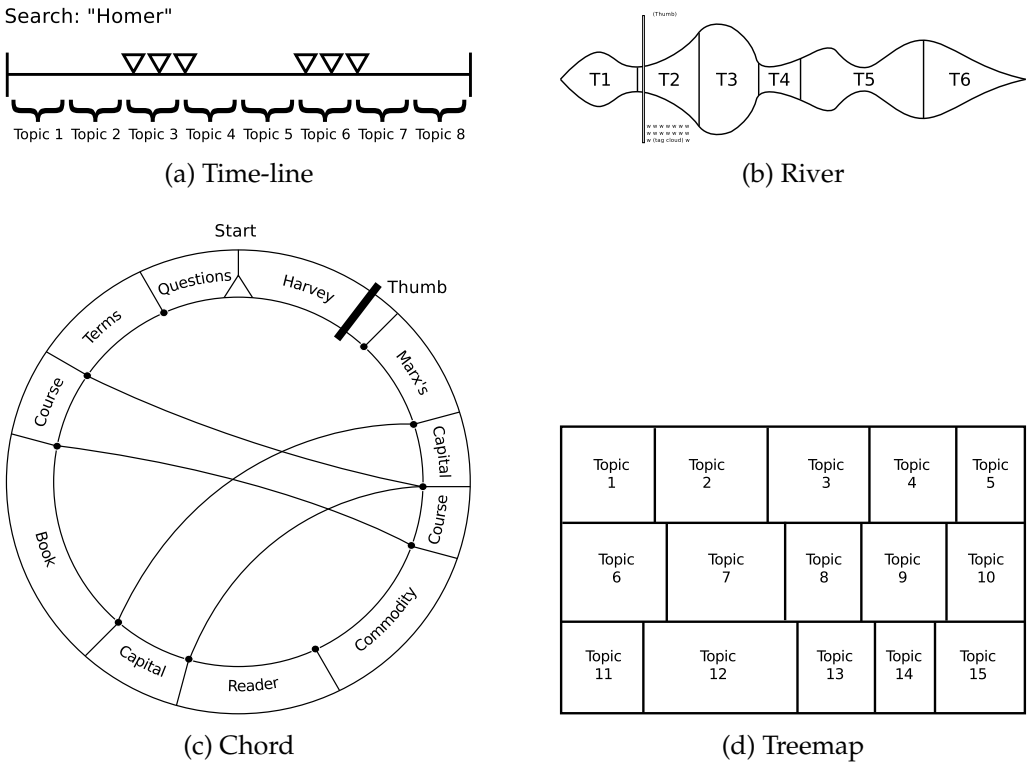
Table 4.1: Persona Functional Requirements satisfied by SpEx.

Task	Requirement	Jack	Amy	Peter	Supported by
Basic	Listen to audio	●	●		Audio Control
Section Selection	Display topic structure	●	●		Strip Treemap (segments)
	Display topic content	●	●		Word Clouds
	Browse/scan for topics		●		Word Clouds and context popups
	Search for topics		●		Search facility
	Move audio between segments		●		Audio Control
Fact Finding	Browse for specific spoken information	●	●		Word Clouds with Transcript Markers
	Search for specific spoken information	●	●		Search facility with Transcript Markers
	Browse/scan for specific acoustic events	●	●		Acoustic filters
	Accurately move the audio position	●	●		Treemap audio thumb
Summarisation	Understand surrounding content		●		Word Clouds with Transcript Markers
	Show regions listened to		●		Audio Control played-region highlight
	Manoeuvre and listen to audio		●		Audio Control and Treemap Thumb
Distribution	Record audio with off-the-shelf hardware and software			○	Not yet supported
	Simple configuration when processing audio			○	Not yet supported

4.3 Design Rationale

Many Information Visualisation designs exist in the literature [19, 20]. Often, designs are generic and unrelated problem domains can be represented with similar models. Hence it is not uncommon for Information Visualisation designs to be applicable to multiple domains. In hindsight, I believe my choice of utilising a Strip Treemap with Word Cloud overlays was a good one, but my design was not an obvious one.

I initially created four candidate designs, each making use of a different Information Visualisation technique. Wireframes of the four designs are shown in Figure 4.11.



(a) Time-line (b) River

(c) Chord (d) Treemap

Figure 4.11: Wireframe designs of my four candidate visualisation methods.

Despite the temporal nature of time, I opted for static visualisations that displayed entire audio recordings. Visualisations that display only a part of a recording at a certain time (one segment for example) would not be adequate. They would not allow users to compare parts of entire recordings when finding related information [1]. The first design (Figure 4.11a) was a time-line, not dissimilar to the designs of VoiceBase [87], BBSearch [78], and Mavis [66] (described in Section 2.2.2). My time-line interface concept consisted of a horizontal line to represent the audio, annotations to display the positions of significant topics, and markers to appear on word search. The time-line could have also been augmented with a line-graph indicating the loudness or pitch along the audio recording. The second design (Figure 4.11b) made use of the river metaphor, much like ThemeRiver [38] and TextFlow [25], to display topics (described in Section 2.3). The design did not overlay co-occurring topics, but only how frequently the topics occurred. I believed displaying co-occurring topics would be difficult to implement correctly. The third design (Figure 4.11c) used a radial layout, sometimes known as a Radial Concordance Diagram (RCD) or a Chord Diagram, much like the visualisations produced by Circos [55]. Textoscope [30] applied a Chord diagram to text visualisation by displaying topics on the circumference of the circular visualisation and topic relationships as edges. I intended to apply the technique to audio recordings by overlaying audio controls and a search facility. I decided that a radial layout would not provide segments with enough space for annotations such as Word Clouds. The fourth design (Figure 4.11d) was the Strip Treemap.

I additionally compared the four designs against my task taxonomy which is described in Section 3.4. I decided that the time-line method would make the poorest choice because Section Selection and Summarisation would not be easy. Summarisation would be difficult because loudness and pitch would not adequately display the structure of a speech recording. Though topics would be visible, the relationships between top-

ics would not be visually displayed. Performing a search would identify relevant regions for Section Selection if users had a well understood search goal. The search facility would also be useful for Fact Finding.

In comparison, the river metaphor would adequately display the topic structure of speech recordings for Summarisation and provide a search facility for Fact Finding. But Section Selection may be difficult because the relationship between segments may be unclear due to a lack of space to properly annotate similarities and differences between segments. I believed that a visualisation based on the Chord diagram would offer an improvement because topics could be placed around the circumference with edges connecting parts of the audio (and hence topics) which were related. The thickness of the edges could correspond to the strength of the relation. Therefore, Section Selection, Fact Finding, and Summarisation could be adequately supported. But I wanted to describe topics with multiple words which was not easily possible with a radial layout, especially if topics were numerous (hence reducing their available space). Further, edges could relate topics, but I did not believe edges would display why topics were related for Section Selection. All visual elements should be clear to users and be easy to learn. Visual elements should improve users' ability to retrieve information.

I thus decided to utilise the Strip Treemap. Treemaps have been a popular visualisation method for online media. Examples include "Their First Words"[3], Newsmap [4], and at least one BBC news article [5] which implies that Treemaps would be the most familiar of the four choices. Audio could be segmented into pieces and each piece given a description to outline what it contained. Being a space-filling visualisation, Treemaps offered ample room to describe segments with Word Clouds, which was useful if multiple topics overlapped. Section Selection and Summarisation could be supported by displaying segment Word Clouds and Fact Finding could

[3] http://www.visitmix.com/labs/descry/theirfirstwords/

[4] http://newsmap.jp/

[5] http://news.bbc.co.uk/2/hi/technology/8562801.stm

be supported by selecting Word Cloud words to display exactly were the selected words occurred in the audio. A search facility could also be added to support Fact Finding.

The cell size and ratio protection offered by Strip Treemaps was particularly effective for displaying Word Clouds. Enough space was provided for a segment description and cells were not produced which were either too wide or too narrow to effectively display legible text.

My use of Word Clouds to display topic descriptions was not my first choice. Topic descriptions initially consisted of significant phrases. That is, phrases which occurred frequently in one segment and infrequently in other segments. But after an informal presentation to my colleagues, I found that the significant phrases were unclear because not enough context was provided (and I must agree). In a second attempt, segment descriptions were represented by the top ten most frequent words and displayed as a list. When the list of frequent words received good feedback I turned the list into a Word Cloud to make the frequencies, and hence the significance, of each word clear to assist browsing.

Context was still lacking from the Word Clouds, so I experimented with two designs to resolve context by using sentences form the transcript. The designs allowed users to highlight either Transcript Markers or Word Cloud words to display the sentence words occurred in. I opted to reject any interaction which involved selecting Transcript Markers and hence only selecting Word Cloud words displayed context. Transcript Markers were too narrow and often too close to each other to accurately select without missing or accidentally selecting a different Transcript Marker.

Chapter 5

Feature Extraction

SpEx was responsible for displaying information about audio for spoken content retrieval. I developed *TAFE*, a Text and Audio Feature Extractor, to perform audio processing and extract the *features* for SpEx to consume. I build TAFE as a separate tool so that audio recordings could be processed in advance with the results cached for SpEx to consume efficiently. TAFE generated features for the loudness and pitch underlays and extracted frequent words from the transcript for SpEx to display over segments. TAFE also segmented the audio using both acoustic and prosodic features, rather than the words spoken. Beyond the spoken words, recorded speech also contains information about what is happening in the background and how something is said which TAFE utilised to provide meaningful segment boundaries.

Below, I discuss the architecture of TAFE and how TAFE analysed text and audio features. The methods used to analyse the features directly correlated to the quality of the features produced and, in turn, the accuracy of SpEx which directly consumed output from TAFE.

5.1 Architecture of TAFE

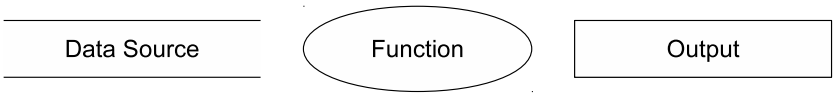
TAFE is a Java application which I created to extract text and audio features from audio recordings to be consumed by SpEx. TAFE extracted loudness and pitch information for the acoustic underlay of SpEx. TAFE also identified audio segments and produced a list of the most frequent words of each sentence for SpEx to display segments and Word Clouds. Finally, TAFE assigned a sentence to each Word Cloud word to be displayed as a context sentence.

More complex audio features such as identifying who was speaking, speaker emotion, and voiced/unvoiced segments were plausible, and potentially useful, additions to TAFE. However, the required machine learning training and a priori knowledge of different conditions [56] may have made configuring TAFE correctly more difficult for Peter. In turn, SpEx may become inaccurate for Jack and Amy. TAFE should be easy to deploy and a trustworthy source of information. In which case, SpEx only visualised text, segments, and raw acoustic features. Although TAFE's limitations influenced SpEx, the visual characteristics of SpEx did influence TAFE as well. For example, the number of segments, Word Cloud words, and granularity of raw acoustic features were dependent on what SpEx could adequately display (as discussed in Chapter 4).

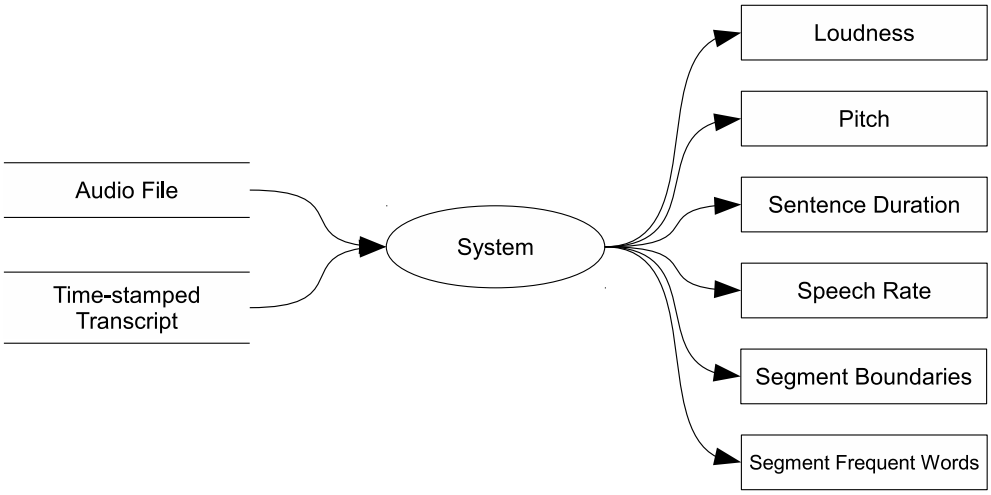
TAFE accepted two files as input: an audio file and it's accompanied transcript. TAFE produced a single file as output which contained five extracted features coupled with the original transcript. Figure 5.1 provides an overview of the inputs and outputs of TAFE.

I designed TAFE as an independent application from SpEx to simplify my development process. Decoupling TAFE simplified the system because TAFE focused on feature extraction alone. Hence, modifications and corrections to TAFE could be made quickly because changes did not propagate to SpEx. Furthermore, my test suite, written in JUnit[1], was

[1]JUnit v4.10.0. `https://github.com/KentBeck/junit`



(a) Data Flow Diagram Legend



(b) 1st-level Data Flow Diagram of TAFE. Inputs are on the left, TAFE is in the center, and outputs are on the right.

Figure 5.1: 1st-level Data Flow Diagram of TAFE

more complete and cohesive which ensured my feature extraction methods were working correctly. As an added bonus, TAFE could potentially be reused with multiple user interfaces (for example, a desktop and a mobile user interface) which would reduce development effort of additional interfaces. Consequently, TAFE could be replaced with a better suited system if need be to satisfy NF 1 (*"The feature processing pipeline must be configurable."*). The ability to pre-process audio recordings was a major benefit of separating TAFE from SpEx because it mitigated the expense of analysing audio recordings. By creating a dependency from the visualisation to the output of TAFE only, lecture recordings could be batch-processed in advance. In which case, SpEx could load the resultant files without notable delay to users and ensure a fluid and responsive experience which did not frustrate.

To keep TAFE as a stand-alone application and promote its reuse, I designed its inputs and outputs to conform to common data formats. By using common data formats, I could ensure that TAFE was compatible with off-the-shelf software that could be easily procured by Peter. As Input, TAFE accepted an MP3 or WAV audio file and a SubRip text transcript file. MP3 and WAV are popular audio formats for recording and storing audio files while SubRip is a simple text-based transcript format which is commonly used for video subtitles. An example of the SubRip format is given in Listing 5.1.

SubRip maps sentence fragments (captions) to the time they occur in the audio. Each caption is represented by an index, a start time, an end time, and its text. Indices start at index one and progress sequentially. Times are represented as a start time followed by an end time with the "`-->`" operator between the times. The time format is:

```
hours:minutes:seconds,milliseconds
```

The text is provided after the time and an empty line separates each caption. Although the example in Listing 5.1 displays punctuation, transcripts, particularly those automatically generated, are not guaranteed to

```
1
00:00:02,050 --> 00:00:06,322
I'm going to talk to

2
00:00:06,322 --> 00:00:11,068
you today about the beginnings
of the Greek experience as far

3
00:00:11,068 --> 00:00:14,470
as we know it,
and I should warn you at once
```

Listing 5.1: Example of the SubRip format. The first three captions of a lecture on ancient Greek history.

contain any punctuation. Inferring punctuation is a non-trivial task that I believe would not contribute much to SpEx because little transcript is displayed to users.

By using common data formats, TAFE's inputs could be procured or generated by existing tools which may already be available for recording lectures. For instance, lecture audio and transcripts could be downloaded directly from online courses[2] and given as input to TAFE.

I decided to have TAFE consume text transcripts rather than produce them automatically from audio. Automatically producing text transcripts from audio is a process called Automatic Speech Recognition (ASR). TAFE did not perform ASR because the resulting transcripts could have been incomprehensible if the ASR system was used outside of its configured environment. To generate accurate, or merely acceptable, transcripts ASR systems must be configured for the specific environments of use. For in-

[2]Both Open Yale (`http://oyc.yale.edu/`) and MIT Open CourseWare (`http://ocw.mit.edu/`) provide captions in SubRip format.

stance, a computer-science lecture spoken in a speaker's second-language may require a different ASR configuration to a political presentation spoken in a speaker's first language.

The quality of the audio recording, accent of the speaker, and the dictionary of words the speaker uses (for instance politics, computer science, or medicine) can significantly impact the accuracy of text transcripts produced and in turn reduce its comprehensibility. A study of ASR errors found nine classes or error [95]: *Critical errors* where a lack of context by the speaker can lead to misinterpretation of a sentence; *Nonsense errors* where a sentence is not understandable; *Addition errors* where erroneous words are inserted; *Deletion errors* where words are missing; *Dictionary errors* where spoken words are not in the ASR dictionary; *Homonym errors* where a correct sounding word is used but not the correct spelling (such as 'to', and 'two'); *Suffix errors* where a word ending is incorrect; *Annunciation errors* where a word is not correct but sounded similar to what was spoken (such as 'air' and 'hair'); and *Spelling errors* where a word was spelled incorrectly. By allowing the use of a properly configured external ASR system, TAFE could operate on accurate transcripts and hence produce more accurate results for SpEx.

With regard to requirements, TAFE met both Distribution requirements as displayed in Table 5.1. Peter could record lectures with off-the-shelf consumer software because TAFE accepted common data formats as input. Audio could be accepted as MP3 or WAV formats and transcripts could be accepted as a SubRip format. Further, TAFE did not contain any configuration settings because advanced machine learning algorithms were avoided. Peter could merely provide the audio and transcript as input and not worry about optimising any settings.

Section 5.2, below, discusses how TAFE extracted features from transcript and audio files and how the extracted features could aid the visualisation of audio recordings.

Table 5.1: Persona Functional Requirements satisfied by SpEx and TAFE.

Task	Requirement	Jack	Amy	Peter	Supported by
Basic	Listen to audio	●	●		Audio Control
Section Selection	Display topic structure	●	●		Strip Treemap (segments)
	Display topic content	●	●		Word Clouds
	Browse/scan for topics		●		Word Clouds and context popups
	Search for topics		●		Search facility
	Move audio between segments		●		Audio Control
Fact Finding	Browse for specific spoken information	●	●		Word Clouds with Transcript Markers
	Search for specific spoken information	●	●		Search facility with Transcript Markers
	Browse/scan for specific acoustic events	●	●		Acoustic filters
	Accurately move the audio position	●	●		Treemap audio thumb
Summarisation	Understand surrounding content		●		Word Clouds with Transcript Markers
	Show regions listened to		●		Audio Control played-region highlight
	Manoeuvre and listen to audio		●		Audio Control and Treemap Thumb
Distribution	Record audio with off-the-shelf hardware and software			●	Common audio format support
	Simple configuration when processing audio			●	Standardised inputs and no configuration

5.2 Feature Extraction Process

TAFE extracted audio and text features by the use of three subsystems:

1. **Audio Feature Extraction:** Produced loudness and pitch features from the audio only.

2. **Text Feature Extraction:** Produced sentence duration and speech rate features from the speech transcript only.

3. **Audio Segmentation:** Produced segment boundaries from the audio loudness, audio pitch, transcript sentence duration, and transcript speech rate. The transcript is also used to create frequent words for describing segments.

I designed TAFE to perform segmentation on acoustic and prosodic features rather than using topic modelling systems (such as Latent Dirichlet Allocation [14]). Acoustic conditions may indicate important background events or vocal cues such as laughter [43] which are understood to indicate topic conclusions. Prosody is equally important because prosody in speech can alter the meaning of an expression: *How* something is said can be just as important as *what* is said. Prosody is a feature not captured in text but carries significant information such as which words were stressed, if a sentence was a question or a statement, and the state of the speaker's emotion [42]. Hence, the output of TAFE included segments based on acoustic and prosodic conditions. My segmentation method was a simplification of the methods described by Maskey et. al. [63] and Jian Zhang et. al. [48] who used acoustic and prosodic information for spoken audio summarisation.

The output of TAFE was a single JSON [24] file which contained the above extracted features accompanied by the original transcript for convenience. JSON is a data-interchange text format built with a subset of the JavaScript syntax. Due to its simple structure, JSON can be consumed by

most popular programming languages. In which case TAFE different visualisations can be developed for TAFE. Indeed, SpEx consumed the output JSON file and the original audio file for display to the user.

Section 5.2.1, below, will give a brief description of digital audio and audio processing concepts before moving on to discuss the individual features and feature extraction processes.

5.2.1 How Digital Audio is Represented

Sound is represented as an *audio signal*. An audio signal is a waveform where the waveform's shape produces volume and tone over time. The vibration of the audio signal on the ear is interpreted as sound. Audio signals can be represented in analog and digital form. Figure 5.2 provides an example of an analog and digital audio signal.

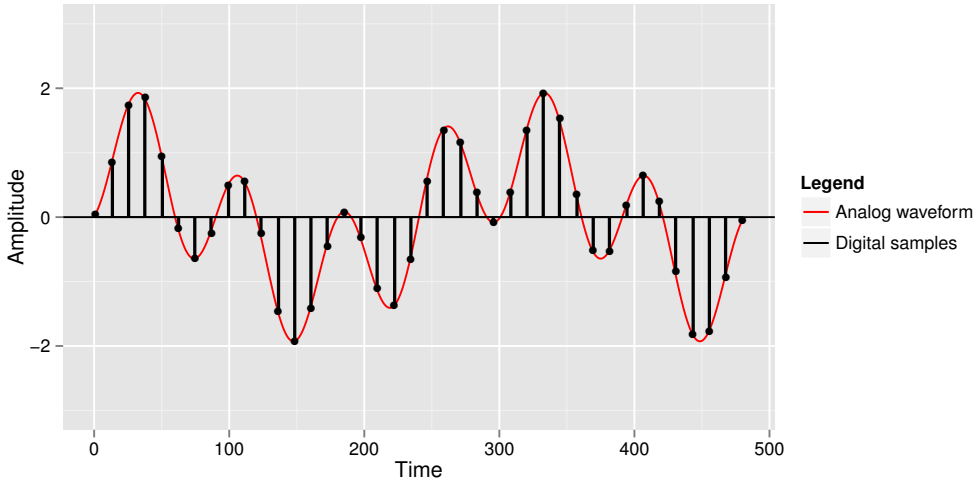


Figure 5.2: An example of the analog and digital representation of audio.

An analog waveform is shown in red. The height, or amplitude, at any point corresponds to volume and the distance between waves corresponds to frequency, or pitch. A digital waveform (represented by computers) is

shown in black. Computers store audio as a series of integers where each integer represents the amplitude of the waveform at regular points in time.

In order to record audio to a computer, the audio signal must be converted from analog to digital form where the continuous analog audio signal is interpreted as a discrete series of integers. To convert analog audio to digital audio the analog signal is *sampled* at regular intervals and each value is stored as an integer. The number of samples taken per second is called the *sample rate*. For instance, CD audio is recorded at 44,100Hz or 44,100 samples per second. The sample rate of an audio file must be chosen to preserve the original waveform. If a low sample rate is chosen, the quality of the audio will be reduced because the shape of the original waveform may not be fully captured by the sampled points. If a high sample rate is chosen, the resultant audio file may become too large for easy storage and transfer.

A process called quantization converts the range of sampled values from an analog signal to a discrete set of values for a digital signal. Each sample is represented as an integer stored in a number of bits. The number of bits used is called the *bit depth* and represents the number of steps between the lowest and highest amplitude value. Quantization converts analog values to the nearest digital values that the bit depth can represent. Figure 5.3 displays the effect of quantization after converting an analog signal to a digital signal.

It is evident that the original analog waveform cannot be fully captured by quantization. Hence, a digital audio signal represents an approximate of the original analog audio signal. Notwithstanding, a large enough bit depth can capture the audio signal in enough detail to provide an accurate representation of the original sound by the human auditory system. A small bit depth gives the appearance of a granular, low-quality, sound. CD audio is typically recorded with a bit depth of 16 bits which can store 65,536 distinct sample values and can represent a large portion of what we can hear. In addition, audio is represented as a number of channels, where

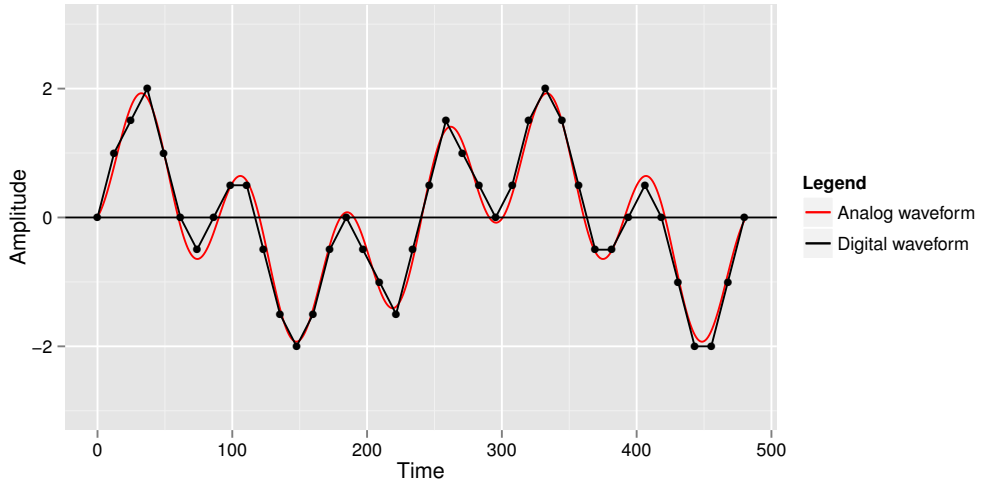


Figure 5.3: Example quantization of a waveform with 8 steps (a bit depth of 3).

each channel is a separate stream of audio. For instance CD audio is two channel, commonly known as stereo, audio. One channel is to be listened on each ear. It is not uncommon to find audio in one channel, known as mono, as well.

When analysing digital audio, the audio is split into a series of contiguous samples called a *window*. A window is a collection of sequential samples that provide a snapshot of an audio signal at a moment in time. Windows are used for calculating features that have a time component (such as frequency). Typically the length of a window is measured in milliseconds to represent a perceivable instant of the audio. An application will start with a window at the beginning of an audio signal and shift (slide) the window until every sample has been observed. Windows are often shifted by a smaller amount than their size, rather than being juxtaposed, to minimise discontinuities at the window boundaries. Sample rate, bit depth, window size, and window shift all effect the accuracy of the features extracted.

5.2.2 Audio Feature Extraction

SpEx made use of loudness and pitch acoustic features to display as an underlay to the Treemap. Loudness and pitch were additionally used by TAFE to identify segment boundaries. Therefore, TAFE extracted the raw loudness and pitch to produce as output and also used the features for later segment boundary detection.

Loudness and pitch were calculated by sliding a window along the audio file and generating a value for each window. The window was 50ms in duration with a shift of 20ms which allowed for overlapping windows to reduce discontinuities at the window boundaries. Root Mean Squared (RMS) Energy was used to calculate loudness. In other words, the square root of the mean of the squared sample values. Assuming n to be the number of samples in a given frame and x_i as an individual sample in the given frame, the formula for RMS Energy is provided in Equation 5.1.

$$rmsEnergy = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} \tag{5.1}$$

Pitch was calculated with the YIN pitch detection algorithm [27] which I ported from the Aubio library [16]. YIN uses autocorrelation to estimate frequency (pitch). A window is taken and shifted forward in time. For each shift a difference is calculated between the original window and the new window. When the difference reaches a minimum, a multiple of the period is found. In other words, when the waveform repeats itself. The time difference is used to estimate frequency. The YIN algorithm additionally takes extra steps to reduce inaccuracies, most notably to reduce unintentionally amplified peaks produced from the difference function and to reduce subharmonic errors.

For each sentence in the corresponding text transcript, the minimum, maximum, mean, and slope of the loudness and pitch were calculated to produce an eight-dimensional acoustic vector to describe the sentence. The acoustic vector was extended with textual features as described below.

5.2.3 Text Feature Extraction

Two text features were extracted: sentence duration and speech rate. Although sentence duration and speech rate were not used by SpEx, sentence duration and speech rate have been found to work alongside loudness and pitch to produce segment boundaries [63, 48]. These segment boundaries take into consideration both the speech and acoustic qualities of the audio.

I chose to denote each individual caption as a sentence rather than finding full-stops because TAFE could not rely on punctuation to be provided. ASR systems are not guaranteed to produce punctuation, and any punctuation produced would be inferred rather than known with certainty (unless the transcript was manually produced or manually corrected). Generally ASR systems will split captions when there is a notable pause in the speech. I assume that these pauses are adequate indicators of sentence boundaries or meaningful sentence fragments for the purpose of segmentation. Pause-based sentence boundary detection is not considered as accurate as other, learning-based, methods such as Statistical Language Models and Support Vector Machines [34]. Spontaneous speech contains too much pause variation for reliable sentence boundary detection. However, such learning-based systems are complex and would require delicate configuration from Peter, who is already a busy person.

For each sentence (or caption), I calculated its duration as the time difference between the first and last words. The times of the first and last words were recorded in the transcript provided to TAFE (Listing 5.1) so no audio processing was required. The calculation follows Equation 5.2.

$$sentenceDuration = endTime - startTime \tag{5.2}$$

Where $endTime$ and $startTime$ denote the time in seconds (as real numbers) from the beginning of the audio recording to the end and start of the sentence respectfully. $sentenceDuration$ was also represented in seconds. Speech rate was calculated per sentence by Equation 5.3.

$$speechRate = \frac{sentenceDuration}{numWords} \tag{5.3}$$

Where $sentenceDuration$ is calculated with Equation 5.2. $numWords$ is the number of words in the sentence (including stop-words such as "a" and "the"). Speech rate could also be calculated by the average syllable duration or the ratio of voiced to unvoiced audio in the sentence. However, like ASR, the accuracy of analysing syllables or voiced and unvoiced frames is affected by audio quality, speaker accent, and background noise. I instead chose the simpler method of using words spoken per second per caption. Consequently, TAFE did not distinguish between long and short words which would impact the speech rate calculation because longer words would take more time to say.

Sentence duration and speech rate were appended to the acoustic features to produce a ten-dimensional feature vector which was used for segmenting the audio.

5.3 Audio Segmentation

Three steps were involved to segment the audio into fifteen pieces. First the ten-dimensional feature vectors which contained acoustic and text features were dimensionality-reduced. Second, the dimensionality-reduced vectors were clustered into a hierarchical dendrogram. Finally, fifteen clusters were chosen from the produced dendrogram.

The ten-dimensional feature vectors were dimensionality-reduced to two dimensions by Principal Component Analysis (PCA). [3] PCA was configured to use the Whitening transformation type. Whitening was used to reduce correlations in the feature vectors. I used the first principal component as the first dimension of the new two-dimensional feature vectors and the order index was used as the second dimension. The order index was used for the clustering step to insure clusters were made of contiguous features.

[3] The `pca_transform` Java library was used: `https://github.com/mkobos/pca_transform`

I used complete-link clustering [62, Ch. 17] with the Euclidean distance measurement to cluster the two-dimensional features into a hierarchical dendrogram. The complete-link clustering was provided by the LingPipe [3] Java library. The hierarchical dendrogram was split into fifteen clusters by starting at the root and breaking links down until fifteen clusters were created. The clusters were composed of contiguous sequences of feature vectors because the Euclidean distance measurement preserved vector ordering.

In the following Section, I analyse the execution time and segmentation accuracy of TAFE.

5.4 Performance Evaluation

I evaluated the ability of TAFE to support SpEx by measuring its execution time in Section 5.4.1 and evaluating the accuracy of its clusters in Section 5.4.2.

5.4.1 Execution Time

Like any computer program, it is desirable for TAFE to execute as quickly as possible. Performance is specified by NF 4 (*"Lecture recordings should process in under five minutes."*). Jack and Amy both expect to access lecture recordings quickly and university hardware to support ASR and TAFE may be limited. ASR systems to capture the speaker's voice as text are computationally expensive, taking real-time or longer to process a recording on a single computer. The long execution time of ASR may be particularly of concern for transcribing an existing database of lecture recordings. Limitations on time and computer resources at universities would govern how many lectures could be processed. TAFE accepts transcripts as input so TAFE must run after ASR, and ideally with minimal additional burden to execution time. Additionally, I would like TAFE to execute within five

minutes on standard personal machines. Peter, Jack, and Amy should be able to process lecture audio away from university infrastructure, when and where they need to.

To measure the execution time of TAFE, I compiled a dataset of lecture recordings. I created my own dataset because, to my knowledge, there was no existing corpus of transcribed lecture recordings which could be used for comparison against other methods in the literature. A detailed description of my dataset is presented in Table 5.2.

My dataset consisted of twenty one-hour lectures: Frontiers and Controversies in Astrophysics (ASTR160) from Yale University [6], lectures one to five; Introduction to Ancient Greek History (CLCV205) from Yale University [50], lectures two to six[4]; The Psychology, Biology, and Politics of Food (PSYC123) from Yale University [17], lectures one to five; and Introduction to Computer Science (CS50) from Harvard College [61], lectures one to five. The mean duration of the lecture recordings was 1h 4min with a standard deviation of 11min. All audio files were converted to the same format to ensure consistency between audio recordings. My audio configuration is described in Table 5.3.

I executed TAFE thirty times for each lecture recording (six hundred runs in total). My test computer had a quad-core Intel Core i5-2400 CPU running at 3.10Ghz, 3Gbyte of RAM, and a HDD speed of 7,200RPM. Java v1.6.0_34 was installed. The `time` Linux/Unix command was used to measure the real time TAFE took to execute and each execution of TAFE was run sequentially.

On average, TAFE took 94.69 seconds to run with a standard deviation of 52.43 seconds. The mean execution time was short and within my tolerable execution time of five minutes (300 seconds). Although the mean time was satisfactory, the standard deviation was large and upon further inspection I found that the execution time of TAFE increased when more captions were present, as shown in Figure 5.4.

[4]CLCV205 Lecture 1 was too short at 33min 2 sec.

Table 5.2: Lecture dataset for evaluating TAFE.

Course	Lecture	Title	Duration (min.)
ASTR160	1	Introduction	46.73
	2	Planetary Orbits	51.38
	3	Our Solar System and the Pluto Problem	45.92
	4	Discovering Exoplanets: Hot Jupiters	46.67
	5	Planetary Transits	49.35
CLCV205	2	Introduction	68.27
	3	The Dark Ages	72.50
	4	The Dark Ages (cont.)	68.00
	5	The Rise of the Polis	66.78
	6	The Rise of the Polis (cont.)	68.52
CS50	1	Week 0: Wednesday	70.13
	2	Week 0: Friday	71.30
	3	Week 1: Wednesday	76.31
	4	Week 1: Friday	53.20
	5	Week 2: Monday	73.70
PSYC123	1	Introduction: What We Eat, Why We Eat and the Key Role of Food in Modern Life	60.57
	2	Food Then, Food Now: Modern Food Conditions and Their Mismatch with Evolution	74.13
	3	Biology, Nutrition and Health I: What We Eat	72.75
	4	Biology, Nutrition and Health II: What Helps Us and Hurts Us	79.18
	5	Biology, Nutrition and Health III: The Psychology of Taste and Addiction	70.30

Table 5.3: Configuration of audio files in dataset.

Property	Value
Sample Rate	22,050Hz
Bit Rate	64kbit/s
Audio Channels	Mono (single channel)

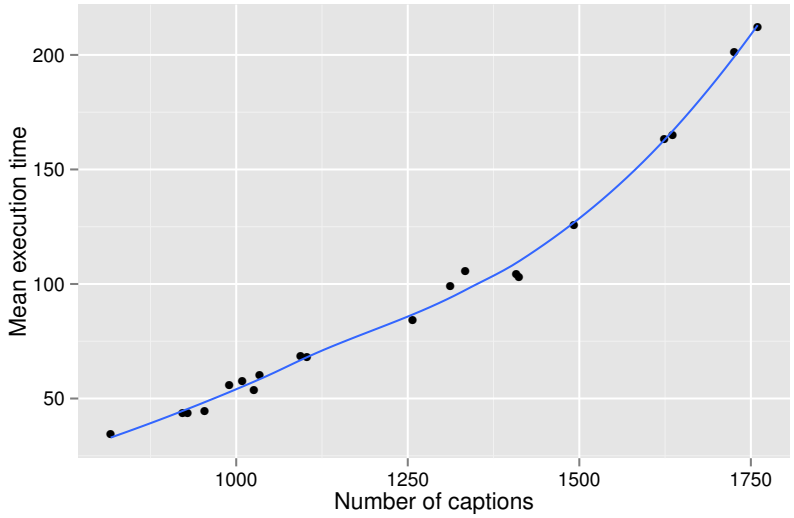


Figure 5.4: Relationship between number of captions and mean TAFE execution time. Each data-point represents one lecture.

The clear increase in execution time with more captions was primarily due to the hierarchical complete-link clustering algorithm. Audio and text feature extraction stages did not not significantly impact execution time because the same amount of audio and text was guaranteed to be processed a finite amount of times. The hierarchical clustering, on the other hand, has $O(n^2)$ time-complexity [60] with the number of captions. Hence, the average number of words an ASR system places per caption, the pace of a speaker's speech (with fewer pauses to separate captions), and particularly the length of an audio recording do impact on the execution time of TAFE. As TAFE has no control over its inputs, the time-complexity of hierarchical clustering is clearly a limitation if execution time is important to Peter, the content creator.

Although an expensive operation, I found that the hierarchical clustering algorithm produced segment boundaries which were, on average, better than my baseline segmentation algorithm. I analyse segment accuracy in the following section.

5.4.2 Segment Accuracy

My design decision for SpEx to display exactly fifteen segments (discussed in Section 4.1.2) not only impacted how much information was displayed to users but also the segmentation performance of TAFE. If I displayed too many segments in SpEx the interface would become cluttered with Word Clouds, while if I displayed too few segments the interface would present too few Word Clouds to give a descriptive structure of an audio recording to support my task taxonomy.

As a result of enforcing a strict number of segments, TAFE must segment audio into fifteen pieces regardless of how many (or how few) segments an audio recording may be expected to have. Therefore, TAFE must generate segments boundaries where they may not be appropriate. I evaluated the segments TAFE produced against a baseline segmentation algorithm which segmented audio into fifteen uniformly sized segments. Segments derived from TAFE and the baseline segmentation algorithm were compared against the official topic locations provided with the online university lectures. The official topic locations were manually generated. My dataset consisted of the same twenty lectures described in Section 5.4.1. The number of official topics ranged between three and nine (inclusive) topics with a mean of 5.35 topics.

Following the evaluation procedures of lecture segmentation in the literature [58, 75, 74], I used Precision, Recall, and F-Measure to evaluate the accuracy of the segments produced by comparing against baseline segments. I compared TAFE to the base-line algorithm and against the official topic locations provided by the respective universities. I considered a segment boundary to be correct if it was +/- thirty seconds from the official topic boundary. The results are displayed in Table 5.4.

It is clear that segments produced by TAFE and the baseline produced poor precision (0.074 and 0.050 respectfully). Most segments were not close to official topic locations because both TAFE and the baseline produced exactly fifteen segments, much greater than the five official topics

	TAFE			Baseline		
Task	Precision	Recall	F-Measure	Precision	Recall	F-Measure
ASTR160 Lec. 1	0.133	0.4	0.2	0.133	0.4	0.2
ASTR160 Lec. 2	0.067	0.333	0.111	0	0	-
ASTR160 Lec. 3	0.133	0.5	0.211	0.067	0.25	0.105
ASTR160 Lec. 4	0.067	0.333	0.111	0.133	0.667	0.222
ASTR160 Lec. 5	0.2	0.75	0.316	0.067	0.25	0.105
CLCV205 Lec. 2	0.067	0.25	0.105	0.067	0.25	0.105
CLCV205 Lec. 3	0.133	0.667	0.222	0	0	-
CLCV205 Lec. 4	0	0	-	0.133	0.667	0.222
CLCV205 Lec. 5	0.067	0.2	0.1	0.133	0.4	0.2
CLCV205 Lec. 6	0.067	0.25	0.105	0	0	-
CS50 Lec. 1	0.067	0.5	0.118	0	0	-
CS50 Lec. 2	0.067	0.25	0.105	0	0	-
CS50 Lec. 3	0.067	0.333	0.111	0	0	-
CS50 Lec. 4	0.067	0.5	0.118	0.067	0.5	0.118
CS50 Lec. 5	0.067	0.2	0.1	0.067	0.2	0.1
PSYC123 Lec. 1	0	0	-	0	0	-
PSYC123 Lec. 2	0.067	0.143	0.091	0.067	0.143	0.091
PSYC123 Lec. 3	0.067	0.2	0.1	0.067	0.2	0.1
PSYC123 Lec. 4	0.067	0.25	0.105	0	0	-
PSYC123 Lec. 5	0	0	-	0	0	-
Mean	0.074	0.303	0.116	0.050	0.196	0.078
Std. Dev	0.048	0.206	0.076	0.052	0.227	0.083

Table 5.4: TAFE and baseline segmentation accuracy compared to official topic locations. Values rounded to three decimal places.

present on average. However, the mean precision and recall produced by TAFE was greater than that produced by the baseline which indicates that the clustering of acoustic and text features can help with identifying distinct regions in the audio. However, the results need not be very accurate because SpEx visually depicted important topics to be found by users in the form of Word Clouds and offered interaction mechanisms to identify exactly where key topics occurred in the audio. The segments produced by TAFE merely broke audio into easily consumable segments which, when visualised by SpEx served to provide a high-level structure of audio to assist navigation.

The effectiveness of SpEx and in-part the effectiveness of the segments TAFE produced were evaluated in a user study which is described in the following two chapters.

Chapter 6

User Study

I undertook a user study to understand how well users were able to use SpEx to navigate audio and the strategies users employed during the process. My user study was designed to analyse user performance against the context of the primary personas, Jack and Amy.

To replicate the conditions of university education, lecture and presentation audio were used and undergraduate university students made up the majority of participants. In total, twenty participants took part in my user study. I asked each participant to perform a predetermined series of tasks for each audio recording. My tasks were designed to characterise the tasks in my task taxonomy and the scenarios of Jack and Amy.

I recorded user actions to produce a set of quantitative data for statistical analysis of usage patterns while user opinions were obtained to gain insight into user thoughts and perceptions. The data from the user study will be analysed to verify or disprove experimental hypothesis and provide answers to open questions about SpEx.

6.1 Type of User Study

There were two categories of user study I could create, a laboratory study where participants are given artificial tasks, and a field study where SpEx

is deployed for a real university course. I opted for a laboratory study. While analysing the genuine usage of SpEx in the field to fulfil real goals is useful, laboratory study could offer a more controlled environment. A controlled environment would allow me to carefully tailor the tasks users performed to directly correspond to my task taxonomy and I could additionally gather observational data for each participant. Consequently, I would not expose SpEx to untested environments where I could not guarantee a quality of experience at the prototype stage that it was in. Field studies are known to discover usage scenarios and behaviours not found in laboratory studies [44], so I leave a field study for further work.

Additionally, I did not design a comparative study. I believed existing audio retrieval interfaces have few comparable features to SpEx. Further, no user study of an audio retrieval system has before used my task taxonomy to structure its tasks, making comparison difficult. Analysing SpEx alone still allows me to gain insight into key usability issues that may hinder its use.

6.2 User Study Goals

Information Visualisation prescribes the use of visual display and interaction to filter and manipulate information. Therefore, it is important to understand how users make use of the different visual elements of SpEx to accomplish their goals. Observing how a visualisation is used for real-world tasks is key to understanding the worth of individual visual elements and how these elements work together to serve the goals of the user. An understanding of how interaction components benefit or hinder users would extend to all target users characterised by Jack and Amy. The results would help to demonstrate the utility of SpEx and may help others to create more effective audio navigation tools in the future.

With this in mind, I wished to understand how users interacted with and perceived SpEx. The user study was thus designed to resolve the

following hypotheses:

H1 Segments and Word Clouds will help display the structure of speech audio for Section Selection.

H2 Transcript Markers and the search facility will help users to locate specific regions in speech audio for Fact Finding.

H3 SpEx will help users to comprehend the content of speech audio for Summarisation.

H4 SpEx will allow users to find changes in acoustic conditions.

H5 Users will respond positively to SpEx.

H6 Users will want to use SpEx in the future.

H2, H4, and H5 were created with regard to Jack's and Amy's scenarios to find specific information. Jack and Amy both searched for facts and expected the interface to be intuitive. H1 and H3 were additionally related to Amy's scenarios to recap entire topics. H6 was related to Peter's goal to support students' learning.

In addition to the above experimental hypotheses, the user study was also designed to answer some open questions about how users interacted with SpEx:

Q1 How do users perceive the workload involved?

Q2 Which navigation strategies lead to the effective use of SpEx?

Q3 Which elements of SpEx are used most often?

These open questions served to provide insight into the search and browsing strategies of users. The open questions also exposed users' subjective measures of the workload employed to complete their tasks. To prove or disprove the hypothesis and answer the open questions above, a group of university students were recruited to take part in the user study. A description of the user study participants is given below.

6.3 Participants and Laboratory Setup

University students were recruited to match the characteristics and needs of the student primary personas, Jack and Amy. The results of the user study will more closely match those if deployed in the real world because SpEx is targeted towards the needs of students. In total, twenty participants took part in the user study where eighteen participants were male and two were female. The majority of participants were young with nine participants aged between eighteen and twenty-one, nine aged between twenty-two and twenty-five, and two older than twenty-five.

Two participants cited listening to recorded lectures frequently while the remaining were at least familiar with the concept. Fifteen participants were undergraduate university students and the remaining five were post-graduate students. All participants except three were studying in the field of computer science and engineering. Although the majority of the participants had a high level of computer knowledge, no participant had prior experience with SpEx or any similar audio navigation tool. Therefore, I did not consider any participant to be at an advantage to any of the others.

Each participant sat at a standard desktop computer with a mouse, a keyboard, and speakers. A monitor resolution of 1920×1080 was used. Participants were asked to use SpEx on the computer while providing answers to tasks on pen and paper. All participants remain confidential and were labelled with a unique ID from P1 – P20 (P for "participant").

My user study was approved by Victoria University of Wellington's Human Ethics Committee (HEC). All participants were provided with an information sheet concerning their expectations and rights before giving consent. The application provided for HEC approval and documents given to participants are found in Appendix A.

The following section will discuss the design of the user study including the tasks participants completed and how the results were taken from

participants.

6.4 User Study Design

I designed my user study to collect the usage characteristics and personal opinions of participants when doing common audio navigation tasks. I based the tasks users performed on my personas (Section 3.3) and audio navigation task taxonomy (Section 3.4).

The user study was structured to produce quantitative data for the statistical analysis of usage patterns and qualitative data for the interpretation of participant opinions. My aim was to gain a holistic understanding of user behaviour and experience when using SpEx. My user study consisted of a set of audio information retrieval tasks for participants to complete which I measured and a questionnaire for participant opinions which I interpreted. The total length of my study was approximately forty-five to sixty minutes per participant.

Only transcripts of audio recordings with 100% accuracy were used in my user study. The effect of transcript accuracy on audio navigation was not within the scope of the thesis and has already been documented [46, 68]. Transcripts with a word error rate (WER) [1] of 45% are marginally useful, while it is desirable to have a WER less than 25%. Transcripts with errors would also add another variable to my user study, making my results more difficult to interpret.

The specific tasks users performed and experimental procedure are described in Section 6.4.1. My procedure for marking the quality of participant answers is given in Section 6.4.2. Section 6.4.3 describes the questionnaire that was provided to participants upon completing all tasks.

[1] Word Error Rate (WER) is defined as the number of word substitutions, deletions, and insertions divided by the number of words actually spoken.

6.4.1 User Study Tasks

I presented participants with four audio recordings, one at a time. Each audio recording presented a different subject matter to ensure that participants could not apply existing knowledge when completing their tasks. By not applying existing knowledge, participants were forced to make use of SpEx for every task. Results were not biased towards those who understood a specific subject matter well. The four audio recordings included three introductory university lectures on astrophysics, ancient Greek history, and politics of Food. The remaining audio recording was a political speech. Lectures and formal speeches were assumed to be indicative of the content Jack and Amy would listen to during their course. I noted that seventeen out of twenty participants studied computer science topics. Therefore, I did not include any computer science lecture recordings. I assigned one recording for a practise session to train the participants for the user study. Two other recordings contained a set of five content-based tasks regarding what the speaker talked about. The last recording contained a single acoustic-based task asking to identify when a change of speaker occurred.

Participants completed all tasks for one recording before moving on to the next recording. Each task was in the form of a question for participants to answer by using SpEx. The questions were produced based on lecture notes for lectures and news articles for the political speech. The content-based tasks were designed to correspond with the audio navigation task taxonomy in order to replicate how SpEx may be used in everyday situations. Therefore, tasks mirrored the tasks Jack and Amy would perform. The acoustic-based task was designed to exploit the extra dimension of acoustic data not found in the text transcript. One benefit of listening to audio is the preservation of acoustic cues such as physical events or prosody in the speaker's voice. These cues may be significant events in an audio recording, and therefore the effectiveness of SpEx to identify such cues is important.

Each content-based recording had the same ratio of task types: two questions to identify regions where a particular topic was mentioned (Section Selection), two questions to extract specific facts from the audio (Fact Finding), and one question to summarise a roughly ten minute portion of the audio (Summarisation). The Summarisation question appeared only once because Pilot studies revealed that it was the most difficult and time consuming of the questions. The intent of the user study was not to strain participants, but to gather meaningful information about participant behaviour under normal conditions.

The questions were not only designed to correspond to my task taxonomy, but also to correspond to the goals and scenarios of my primary personas. The Section Selection and Fact Finding questions were designed to elicit Amy's and Jack's persona scenarios respectfully as well as their Goal 3 (*"Quickly find important information in course material."*). That is, to recap unclear portions of lectures and finish assignments quickly. The Summarisation question addressed Jack's and Amy's Goal 4 (*" Recap unclear potions of lectures."*) as well as Amy's scenario 2.

As an example, Figure 6.1 displays the tasks for the introductory lecture on ancient Greek history (a recording used in the user study).

1. When defining civilisation, what is the difference between a city and a village according to the speaker?
2. In which segments would you look to find information on writing symbols?
3. Name one attacker of Egypt the speaker mentions.
4. In which segments would you look to find information on Homer's epics?
5. Summarise the content of the two right-most segments of the top row.

Figure 6.1: Tasks given to participants for the lecture recording on ancient Greek history.

Tasks 1 and 3 correspond to extracting specific facts. Jack's scenario 1

is thus catered for. Tasks 2 and 4 correspond to identifying a region of a topic. Jack's scenario 2 and Amy's scenario 1 are thus catered for. Finally, Task 5 corresponds to summarising a portion of the audio. Amy's scenario 2 is thus catered for.

As you may have noticed, no question could be properly answered with prior knowledge of the subject area alone. My questions asked for content specific to each lecture to ensure that participants made proper use of SpEx for every question. Thus, as well as providing variation between the different topics in the audio recordings, I have another means to mitigate the effect of prior knowledge on results to ensure my data is reliable. Table 6.1 provides an outline of the user study. The full user study material, with answers, can be found in Appendix B.

Order	Description
1	**Round 1: Practise**
1.1	Introduction to Astrophysics, "Introduction" [6] (46 minutes)
1.2	NASA-TLX rating scale
2	**Round 2: Real**
2.1 — 2.3	Part A: Introduction to Ancient Greek History, "The Dark Ages" [50] (1 hour 8 minutes)
2.1 — 2.3	Part B: State of the Union 2012, "An America Built to Last" [71] (1 hour 5 minutes)
2.1 — 2.3	Part C: The Psychology, Biology and Politics of Food, "Introduction: What We Eat, Why We Eat and the Key Role of Food in Modern Life" [17] (1 hour)
2.4	NASA-TLX rating scale
2.5	NASA-TLX workload comparison
2.6	Feedback

Table 6.1: User Study Outline. The user study was split into two rounds. The first round for practise and the second round was measured. Participants would interact with four audio recordings and complete a NASA Task Load Index to measure workload.

I split the user study into two rounds, the first round was the practise session and the second was where user performance was measured. In total, four speech audio recordings were presented to participants. Three audio recordings were of lectures and one was of a political speech. Round 1 contained one lecture recording only. Round 2 contained two lecture recordings and one political speech recording of which one lecture recording had an acoustic-based question. For convenience, I refer to the lecture on ancient greek history as "Part A" and the State of the Union address as "Part B" (content-based tasks). I refer to the remaining lecture on The Psychology, Biology and Politics of Food as "Part C" (acoustic-based task).

In the practice round, I would walk the participant through using SpEx and during this time the participant was free to experiment with SpEx and ask questions. The participant was given an audio recording and a set of three questions to attempt at their own pace to become familiar with the content and format of Round 2.

Round 2 was recorded and measured. Participants were presented with three audio recordings given in random order to mitigate any learning effects. As well as recording the actions of participants when using SpEx, participant workload was measured using the NASA Task Load Index (NASA-TLX) [70] and participants could provide their opinions at the end of the user study. The following section describes these subjective forms of feedback in detail.

6.4.2 Measure of Participant Performance

I measured participant task performance in order to distinguish those participants who performed well from those who performed less well. My intention was to understand how SpEx was used when contrasting high and low performing participants to gain insight into the effective usage of SpEx. I also intended to compare participant performance against NASA-TLX workload ratings to reveal why certain sources of workload were

high or low in order to suggest possible design improvements.

I defined performance as the quality of answers participants provided for each task. I gave the results provided by each participant a mark and the overall performance score for each participant was created by summing all of his or her marks. The ideal answers for each task can be found in Appendix B. I weighted the marks to ensure that each task type equally influenced the overall score.

For Parts A and B, the Fact Finding tasks were given a Boolean correct/incorrect mark worth one point. As there were four Fact Finding tasks, a total of four marks were available (two questions for both Part A and B). The Summarisation tasks were given a score of one to four (inclusive) where one was considered a very poor answer and four was considered a very good answer. I added the marks from the two Summarisation questions and divided the sum by two to maintain the same weighting as the Fact Finding questions (a total of four marks). Both question types were marked by experts to mitigate the marking bias of a single marker. Two colleagues and I each marked all Fact Finding and Section Selection tasks individually and an average was taken. The average was rounded to the nearest valid integer.

Unlike the Fact Finding and Summarisation tasks, I marked the Section Selection tasks by measuring Precision and Recall of segment accuracy. Precision was defined as the number of segments correctly indicated divided by the total number of segments indicated by the participant. Recall was defined as the number of segments correctly indicated divided by the total number of correct segments. Each mark (Precision and Recall for each question across both content-based recordings) was divided by two to give a total weighting of four — a weighting equal to the other task types. The task to find an acoustic event, in Part C, was marked as either Correct or Incorrect. A correct answer gave four marks.

6.4.3 User Study Questionnaires

Two questionnaires were used to gain an understanding of participants' perception of SpEx. Participant perceptions may extend to target users characterised by Jack and Amy. The first was the NASA-TLX [70]. The NASA-TLX attempts to measure the perceived workload of participants through *workload rankings*. Workload rankings have participants rank each workload's influence to a task on a 21-point scale. The six sources of workload can be found in Table 6.2.

By analysing workload, I could understand the effort and difficulty associated with participant performance when using SpEx. A low perceived workload is an indication of an easier, less strenuous, and more enjoyable task. The lower the perceived workload, the more useful SpEx may be for everyday use. Although initially designed for the aviation industry, the NASA-TLX has been commonly applied to visual and auditory displays [36] which makes it applicable for my user study.

Despite the popularity of the NASA-TLX, it is important to understand that sources of workload are the subjective opinions of participants. Different people will attribute different qualities to their perception of workload. The NASA-TLX helps mitigate workload subjectivity by performing *workload comparisons*. Workload comparisons have participants view combinations of pairings between all six workloads and for each pair select the workload that contributed more to the task. The final workload rankings can then be weighted based on which sources of workload were more important to each participant.

I conducted the NASA-TLX workload rankings once after each round while the workload comparison was performed once after Round 2 only. Only the ranking after Round 2 was analysed. The ranking performed after Round 1 was considered a practise to help participants settle on a ranking strategy and to make participants aware of the sources of workload while performing Round 2.

The second questionnaire was used to ask participants how they felt

Title	Endpoints	Description
Mental Demand	Low / High	How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
Physical Demand	Low / High	How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
Temporal Demand	Low / High	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
Performance	Good / Poor	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
Effort	Low / High	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Frustration Level	Low / High	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

Table 6.2: NASA Task Load Index Sources of Workload.

about the visualisation. Four open questions were provided to gain an insight into the participants' opinions of the different elements of SpEx. The questions asked for positive feedback, negative feedback, further comments, and whether or not the participant would consider using SpEx again in the future.

6.4.4 Recorded Information

Participant actions with SpEx and answers for tasks as well as both NASA-TLX and open questions questionnaires were recorded for analysis. SpEx recorded user actions such as which visual elements were used and how much audio was listened to. Screen capture software was also used to create a video of the screen and record the audio that was played. It was assumed the use of screen capture would help participants feel more at ease than a camera looking over their shoulder and hence may have a negligible impact on performance and NASA-TLX results.

All answers to tasks while using SpEx, open questions, and NASA-TLX rankings and comparisons were provided on paper. Paper was used instead of a digital form as any difficulty switching between task and visualisation could be avoided. The visualisation was a full-screen application, and hence I did not want to further clutter the screen. Keeping the screen space focused on just SpEx was intended to avoid confusing participants with visual elements unrelated to SpEx itself.

Chapter 7

User Study Results

In the following sections I discuss both the quantitative and qualitative results produced from my user study. My results provide insight into how well suited SpEx was for common audio navigation tasks. By understanding which elements were more useful and how participants perceived SpEx, I may prove or disprove my experimental hypothesis and open questions described in Section 6.2.

In Section 7.1, I begin by describing two issues which arose in the user study and how the issues were handled in my results. I follow by discussing participant performance in Section 7.2. Participant workload and user comments are discussed in Section 7.3 before comparing how SpEx was used between the highest and lowest performing participants in Section 7.4. A discussion of the results is found in Section 7.5. Finally, In Section 7.6 I discuss feature enhancements for SpEx in light of lessons learned form my user study.

7.1 User Study Issues

Two issues arose from my user study. The first issue involved an error in Part C which nine of the twenty participants encountered (P1 – P9). Two items of metadata indicating when a student and lecturer ("Student"

and "Prof" tags respectfully) were speaking in the transcript were not removed and were visible in the Word Clouds. The words were not large and not all participants noticed the words, but I discounted the affected participants' results for Part C. The remaining eleven participants (P10 – P20) were given the corrected task. Because the remaining participants all completed the task successfully and with little difficulty, I marked all participants as having completed the task correctly.

The second issue involved P19 who did not complete every task in the allotted time of one hour. P19 completed all tasks except for Part B. P19 was excluded from all analysis which made use of Part B.

7.2 Task performance

The success of any audio retrieval interface depends on its ease of use and the relevance of the information users can retrieve from it. Although computers have been effective at searching through text such as documents, web pages, and meta-data stored in audio and video content, the effectiveness of audio search has not been often investigated. Audio retrieval interfaces in the literature have either omitted a formal user study [74, 78] or have not looked at the relevance or quality of non-fact-based results by participants [85, 45]. By analysing the performance of users on a range of task types, I built a holistic understanding of the performance of SpEx. Each of the four task types (Section Selection, Fact Finding, Summarisation, and acoustic event detection) were marked as described in Section 6.4.2. Marks for each participant are found in Appendix C. Figure 7.1 displays the average marks achieved by each participant (except P19) for each task type in Part A and B. For clarity, all marks are scaled to be percentages in the range of 0% to 100%.

My box and whiskers graphs consist of a box, whiskers, and outliers. The box contains three lines to mark the 1st quartile (Q_1, value splitting 25% of data), the median, and the 3rd quartile (Q_3, value splitting 75%

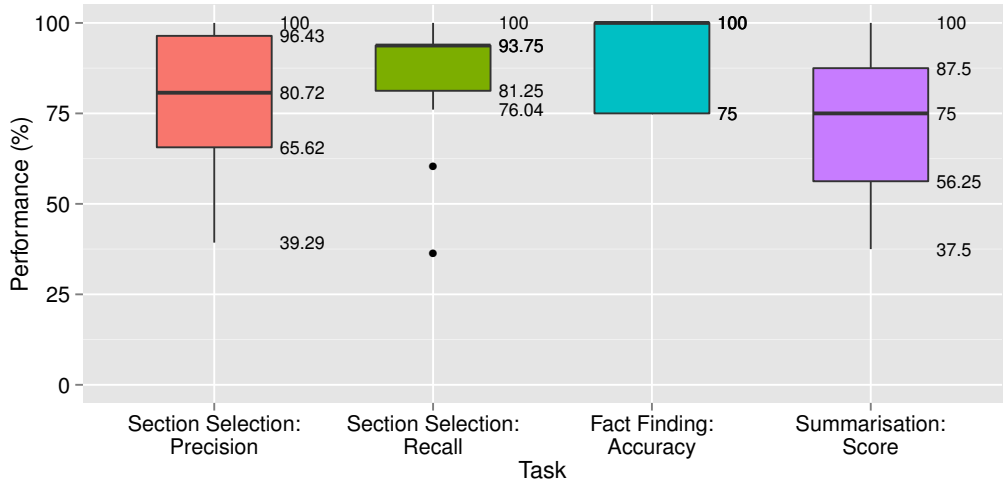


Figure 7.1: Marks for Part A and Part B of the user study. Marks are scaled to be percentages and are rounded to two decimal places. Constructed from mean score for each task type for 19 participants.

of data) from top to bottom respectively. Whiskers mark the largest (or smallest) values that are not outliers. Outliers are any values less than $Q_1 - (1.5 \times IQR)$ or greater than $Q_3 + (1.5 \times IQR)$ where IQR is the inter-quartile range ($Q_3 - Q_1$).

A Shapiro-Wilk test for each task found that Section Selection: Recall could not be considered as normally distributed ($W = 0.763, p < 0.05$) and so too Fact Finding ($W = 0.591, p < 0.05$). Section Selection: Recall and Fact Finding performance distributions were skewed towards the higher performance due to an upper performance limit. In contrast Section Selection: Precision could be more closely regarded as normal ($W = 0.922, p = 0.123$), so too can Summarisation ($W = 0.935, p = 0.214$). These tasks were more difficult and hence performance was not as limited by an upper bound like the other tasks.

The results show that for the Section Selection tasks, participants were able to locate a respectable median of 93.75% of all relevant segments (recall measure). Conversely a median 80.72% of segments participants se-

lected were actually correct (precision measure) which indicates that participants over-selected segments. Participants understood well which segments contained topical words, but had difficulty being selective about which segments contained content significantly related to those topical words.

Comparatively, the Fact Finding tasks were easier than the Section Selection tasks. The results show that for the four Fact Finding tasks, participants had a median accuracy of 100.00% (mean of 92.10%). With a maximum of 25.00% (one of four) of answers incorrect, it is safe to say that SpEx is effective for Fact Finding. Thus, users may be attracted to SpEx because they can be confident that they can find the correct information.

Participants also performed well on the final task type, Summarisation. A median score of 75.00% indicated that most participants were capable of using SpEx to comprehend a portion of the audio. A common strategy participants employed for the Summarisation task was to listen to the audio while, at the same time, highlight words and revealing context sentences of nearby words. By multitasking, participants could comprehend a part of the audio efficiently by selectively inspecting interesting parts and ignoring parts thought to be minor or irrelevant. Hence, SpEx offers an effective method to summarise the content of a portion of audio.

7.3 Perception of Workload

I calculated a NASA-TLX adjusted workload rating for each participant by multiplying the rating given to each source of workload by the importance of the workload as indicated by each participant. Figure 7.2 displays the distribution of each workload.

It is clear in the figure that Mental Demand was perceived by participants as the largest source of workload. After Mental Demand, Temporal Demand, Performance, and Effort were perceived as the second largest sources of workload. Physical Demand and Frustration were considered

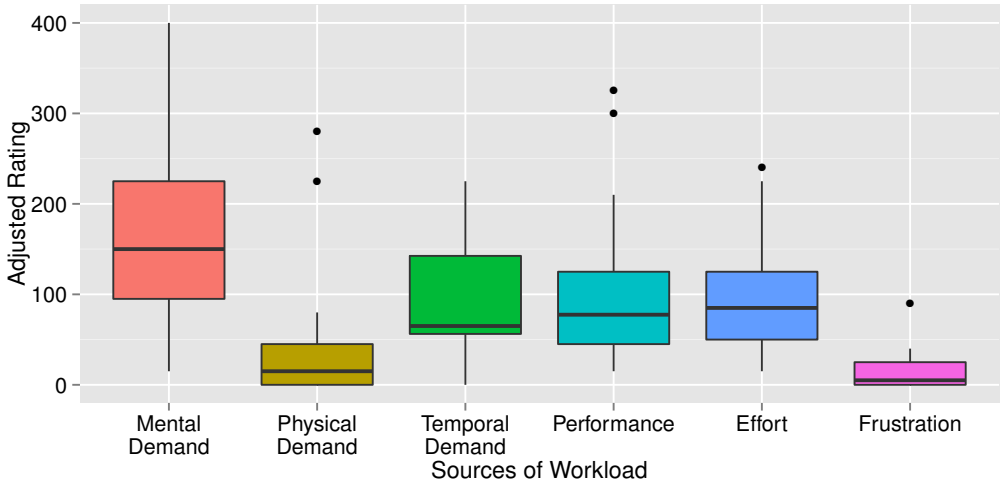


Figure 7.2: NASA-TLX adjusted workload ratings.

to have the least influence.

As Mental Demand was the highest perceived workload, using SpEx must require much thought on the part of the user. For instance, when provided with a task, participants must either identify useful search queries to enter into the search facility (which may include synonyms and related topical words) and/or either browse or scan [10] Word Clouds for words that correspond to the task at hand. Such a process may lead to multiple candidate regions of the audio which may be relevant. Participants must then narrow their search to identify the most relevant regions to listen to.

Although searching is a non-trivial task, participants rated Frustration as contributing little to the tasks. Low frustration indicated that, although much thinking was required, participants did not have difficulty when performing their tasks. Low frustration is a good indicator that SpEx could be used effectively for audio retrieval tasks. Sufficient information was provided to participants, information was displayed clearly, and interaction controls were intuitive. Comments from participants attest to the ease of use of SpEx:

> "Search feature, can find the right section really fast. Highlighting, makes it very easy to see the most likely place to look." (P7)

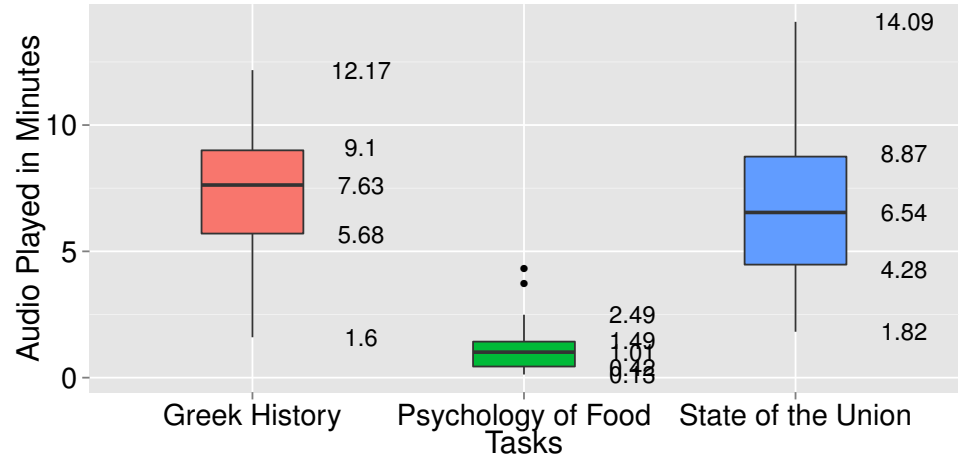
> "It's easy to tell what parts of the talk are about which things. You can actually get an idea of the content of the sections without fully listening to them." (P9)

With regard to performance, Section 7.2 describes the good performance participants were able to achieve. Median Section Selection precision of 80.72%, Section Selection recall of 93.75%, Fact Finding accuracy of 100% (mean of 92.10%), and Summarisation quality of 75.00% indicates that participants were able to use SpEx successfully. But despite the results, participants' perception of performance was not rated low. During the user study, I observed participants frequently re-listening to portions of audio that were believed to contain the answer. Re-listening was used to double-check an answer before settling on the answer. But double-checking was not always performed. The extra time-cost of re-adjusting the audio play position back to the start of a sentence coupled with the inability to quickly perform an in-depth scan of the surrounding content (which is possible with text documents) was an issue. The issue contributed to participants settling on an answer before being fully comfortable with the answer obtained. As an example, here is one such feedback regarding replaying a sentence that was just listened to:

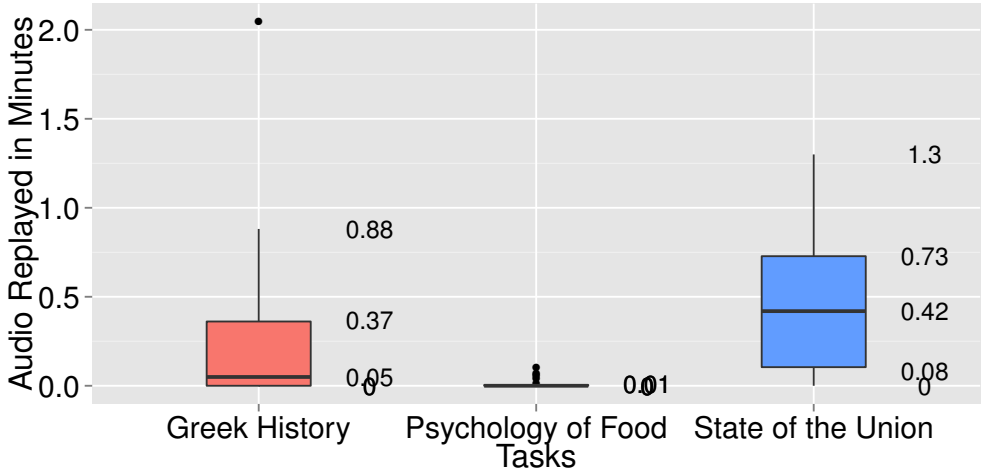
> "Something to show the start and end of each sentence. I.e. when sentence A meets sentence B." (P8)

The amount of audio listened to by participants is presented in Figure 7.3a. As a comparison, the amount of audio listened to more than once is displayed in Figure 7.3b.

It is evident that while Parts A and B contained a median play duration of 7.63 minutes and 6.54 minutes respectfully, the median amount of



(a) Audio Played



(b) Audio Replayed

Figure 7.3: Duration of audio played and re-played in Parts A, B, and C.

audio re-played for Parts A and B were 0.05 minutes and 0.42 minutes respectfully. Most participants replayed some audio which indicates that the need to replay audio is important for audio retrieval. Re-playing audio allows users to double-check what was said. Had SpEx incorporated mechanisms to support efficient audio re-play, participants may have been more confident of their results.

The desire to re-listen to what was just played coupled with searching for audio to listen to had lead to a notable amount of perceived effort on participants. As previously mentioned, searching for places in the audio to listen to is a non-trivial process. Participants frequently attempted multiple search queries, compared multiple segments, and listened carefully to the playing audio. A high perceived effort may mean SpEx is not suitable for use when multitasking. Users may need to pay full attention to the search task at hand. It seems like the inability to multitask may be an issue, but research has found that students like to mimic lecture settings and refrain from other activities when listening to lecture recordings [32].

Common issues by participants that may have increased the effort involved include not providing sufficient information about the transcript and accidentally deselecting all highlighted words. In particular, participants mentioned that some context sentences were ambiguous and there was a desire to produce context sentences over Transcript Markers, not just over Word Cloud words. As a consequence, SpEx may have been more difficult to use. Accidentally deselecting all highlighted words occurred when participants would click on the background of SpEx with the intent to skip the audio to the selected position. Instead, all selected words would deselect and the participant would manually reselect all relevant words again. As one participant, of several, noted:

> "I kept accidentally removing word markers when trying to change when I was listening to — would be better if clicks on blank space moved the time bar if audio was playing (only cleared the screen if paused)." (P15)

An alternative enhancement would be to allow skipping of the audio by clicking the desired location in the background but create a separate button for clearing all highlighted words. In which case the extra interaction to pause the audio would not be necessary to clear highlighted words.

I did not expect Temporal Demand to be rated as high as it was. Participants had sufficient time to complete their tasks and I did not instruct them to go about their tasks quickly. Upon looking at the feedback from the five participants who rated Temporal Demand as highest (P5, P7, P8, P18, and P20), I found a consistent trend. Each participant commented on improving the efficiency of navigating the audio, a comment scarcely raised by the other participants. These participants felt that dragging the audio by locating and repositioning the audio thumb to a precise location on the Treemap took long enough to reduce the speed they expected to use the interface. Hence, clicking the Treemap to seek the audio would be a worthwhile enhancement.

The final source of workload, Physical Demand, was rated by eighteen participants as low. The visualisation did not require too much typing or clicking for the tasks given to participants. Often, bouts of searching the audio were broken with a break to listen to the audio. As such, the level of interaction was not perceived as an issue for participants.

7.4 Usage Strategies

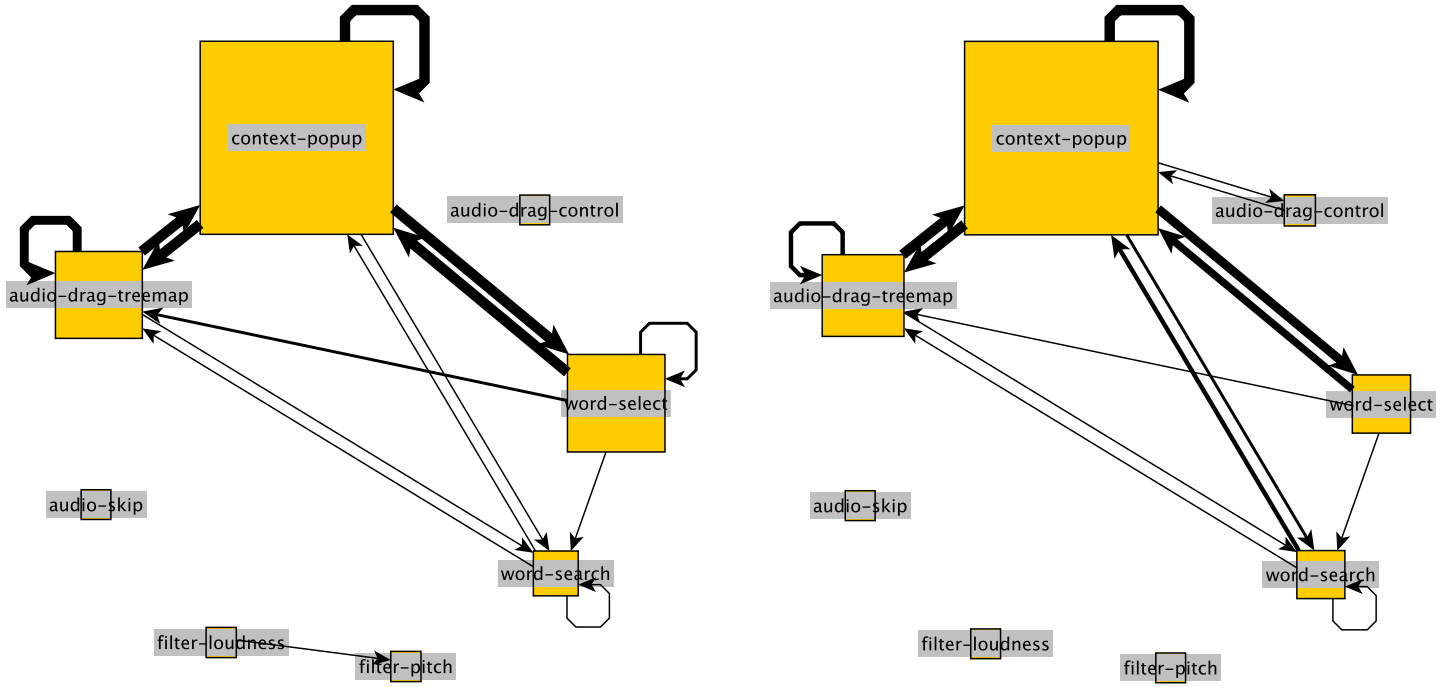
In this section, I discuss the usage strategies of participants for Parts A and B (Section 7.4.1) and Part C (Section 7.4.2) of my user study. Usage strategies were represented by measured tool switching frequency and tool use frequency and were compared between the highest and lowest performing participants. I aimed to provide insight into which tools of SpEx were more useful than others and how favoured tools were made more effective when used together.

7.4.1 Comparing Strategies by Performance for Part A and Part B

I desired to gain insight into how participants made effective use of SpEx for common audio retrieval tasks. Previous studies have looked at tool switching behaviour [29], tool frequency [47, 93], surveys [91], and observations [85] across all participants. However, I wanted to compare high and low performing participants so I could better understand which combinations of tool use lead to better retrieval performance and which did not. By understanding how participants used SpEx effectively, I could make recommendations for its improvement and offer insight into how future audio navigation interfaces should be designed. I chose to analyse tool switching behaviour and observational notes to compare strategies between the highest and lowest performing participants.

I separated the highest and lowest performing participants by ordering participants by overall score (discussed in Section 6.4.2). The top scoring 25% were assigned as the *highest performing participants* and the bottom scoring 25% were assigned as the *lowest performing participants*. Figure 7.4 compares the tool switching behaviour between the highest and lowest performing participants for Parts A and B of my user study. Tool switching frequency was calculated as the average number of times a tool was used across all participants. Only the nineteen participants who completed Parts A and B were included in the measurement.

I found an overall strategy by comparing Figures 7.4a and 7.4b. The most common actions were navigating the audio via dragging, displaying context sentence popups, selecting words, and performing word searches. The audio filters and skipping the audio using the audio control were not found to be useful for finding spoken content. Acoustic filters did not identify relevant spoken information and skipping was not as efficient as dragging because participants found it difficult to skip to precisely manoeuvre the audio due to the disconnect between the audio controls and



(a) Highest performing

(b) Lowest performing

Figure 7.4: Comparison of tool switching between the highest 25% and lowest 25% performing participants in Parts A and B. Nodes represent tools and node size represents tool use frequency. Edges represent tool switching and edge thickness represents tool switching frequency. Edges with tool switching frequency less than 0.1 are omitted for clarity.

the Treemap. Strong back-and-forth relationships existed between dragging the audio position and displaying context sentence popups, and displaying context sentence popups and selecting Word Cloud words. Users would select Word Cloud words to see where certain words occurred, utilise the context sentence popup to display sentence fragments to reveal the transcript, and then drag the audio to listen to regions where the sentence fragments seemed relevant. Observations revealed that users used context sentence popups to build a quick gist of a region of audio by displaying the popup over nearby words, despite the intention of the popups to disambiguate Word Cloud words. When Word Clouds were used to disambiguate words, some participants complained that some context sentence popups did not disambiguate a term because there was too little text.

There were three main differences between the tool switching strategies of the highest performing participants in Figure 7.4a and the lowest performing participants in Figure 7.4b. First, the highest performing participants performed fewer word searches as indicated by a smaller `word-search` node. Searching was commonly the first action taken, and hence a good search could help participants to quickly identify where a word occurred without spending extra time to visually scan Word Clouds for the word. The search space could be narrowed quickly so less time would be spent determining if a region of audio was relevant or not.

Secondly, the highest performing participants clicked on Word Cloud words more often to reveal Transcript Markers as indicated by a larger `word-select` node. Third, the `audio-drag-treemap` node is accompanied by a stronger cyclic arrow and a stronger arrow originating from the `word-select` node. There was a back-and-forth strategy of revealing Transcript Markers and adjusting the audio play position to listen to the audio at the markers to determine relevance. Selecting more Word Cloud words to reveal more Transcript Markers coupled with more audio navigation indicates an effective local search strategy around searched words.

Participants would better understand the structure of a particular region of audio to find their desired information. The local search was possible because the narrow search-space produced from the effective search query beforehand allowed participants to ignore irrelevant regions.

On the other hand, the lowest performing participants made slightly more word searches and selected Word Cloud words significantly less than the highest performing participants. The lowest performing participants instead spent much of their effort displaying the context sentence pop-ups rather than selecting Word Cloud words. Hence, a simpler understanding of the audio structure would be learned and a larger search-space would be searched because word positions and word relationships would not have been properly understood. Consequently, sub-optimal and incorrect answers would be produced.

My analysis would not be complete without considering usage strategies for Part C. Usage strategies for Part C were evaluated separately and are discussed in the following section.

7.4.2 Comparing Strategies by Time for Part C

Usage strategies for Part C of my user study were evaluated by comparing participants who completed the task quickly against participants who completed the task slowly. Time, as opposed to score, was used for splitting participants because the result for Part C was Boolean and all participants were considered to complete Part C correctly. Participants were asked to complete the task in their own time. Therefore, I believed that task completion times in my user study were indicative of task completion times of a real world scenario and were valid for this experiment. Of the twenty participants, I took the eleven participants who completed the corrected Part C task. From the eleven participants, the 25% who completed the task in the shortest time were assigned as the *fastest participants* and the 25% who completed the task in the longest time were assigned as

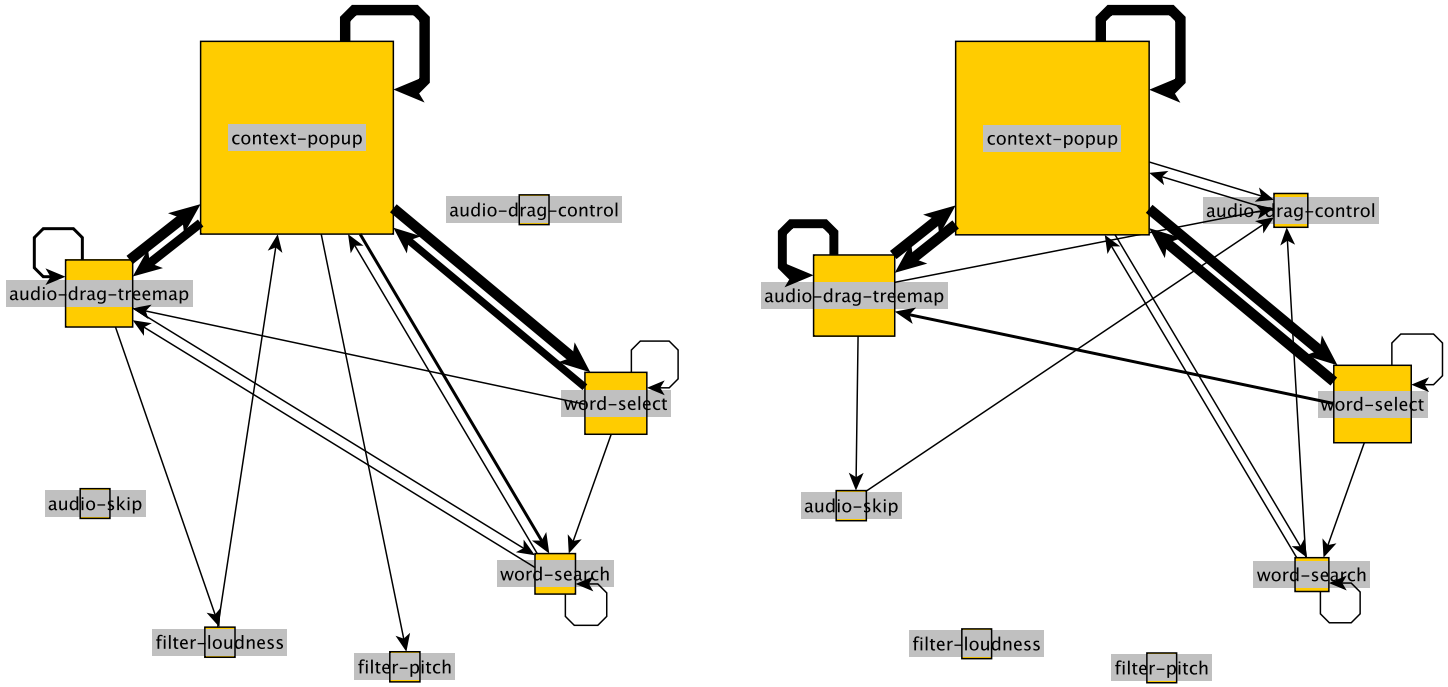
the *slowest participants*. Figure 7.5 compares the tool switching behaviour between the fastest and slowest participants for Part C.

The usage strategies were not entirely dissimilar to those of the highest and lowest performers for Parts A and B displayed in Figure 7.4. The reason for the similarities is because when users were faced with the task, their initial attempts were to make use of the transcript to identify when a student was speaking, when the lecturer was posing a question, or when the lecturer was acknowledging a student. One reason may be the design of SpEx. SpEx was focused on navigating audio by the transcript and so audio filter options were not easily visible. Making the audio filter options more visible may have lead participants to use them earlier rather than later. Some participants used the transcript successfully for the task. For instance, the lecturer would say "yes" when a student wanted to speak. Therefore, selecting "yes" in a Word Cloud would highlight positions where the lecturer was signalling a student to speak. However, the transcript did not make turn-taking behaviour obvious: too few words were visible to decide who was talking based on speech style. Displaying the raw transcript text may have mitigated the issue somewhat by revealing changes in speech style.

I found that by comparing Figures 7.5a and 7.5b it was apparent that the fastest participants made more use of the audio filters to locate when a student was speaking. In contrast, the slowest participants, those who relied more on the transcript, did not. A general pattern of revealing the audio filters and using the context sentence popups to reveal the transcript allowed participants to identify when a student was speaking in short time. Participant feedback supports the use of audio filters as an effective navigation tool for Part C:

> "I found highlighting pitch useful to find where someone might be asking a question." (P11)

> "The volume and pitch highlighting features were interesting and helpful." (P17)



(a) Fastest performing

(b) Slowest performing

Figure 7.5: Comparison of tool switching between the fastest 25% and slowest 25% performing participants in the corrected Part C. Nodes represent tools and node size represents tool use frequency. Edges represent tool switching and edge thickness represents tool switching frequency. Edges with tool switching frequency less than 0.1 are omitted for clarity.

I did observe that participants did not know what to look for when displaying the loudness and pitch acoustic underlays. Participants would look for outlying patterns, usually when the loudness fell to indicate a student, distant from the microphone, was speaking. In which case, the acoustic underlay was successful for locating when a student was speaking, but its utility to search for other acoustic events (such as emotion or background noise) is yet to be understood.

7.5 Discussion

SpEx performed well in my user study. My user study contained questions designed to test each task in my task taxonomy and participants were able to produce satisfactory results for each. Fact Finding was the easiest task, followed by Section Selection, and then Summarisation. Insights from each task are described below in light of my hypothesis and open questions described in Section 6.2. Additionally, participant feedback, workload, usage strategy, and perception of audio segments are also discussed.

Section Selection. Participants were successfully able to identify relevant audio segments in my user study. In most cases, correct segments were found and participants were able to highlight Word Cloud words to reveal other potentially relevant segments. Additionally, the use of Transcript Markers allowed participants to locate segments where glancing references to the desired topic were made and to identify sub-segment regions which were relevant. I found that a common issue of Section Selection was selecting too many segments. Selecting more segments than necessary wasted time because more effort would be needed to analyse more segments. Evidently, SpEx needed to improve the way it displayed segments to allow users to more accurately gauge the relevance of a segment to a topic.

One improvement may be to visually indicate how relevant two segments were based on the number of shared words. Displaying the degree of relevance of each segment to another may discourage users from selecting segments with little relevance to the current segment, saving time and effort. An alternative improvement would be to display more of the underlying transcript to give users a more accurate depiction of the contents of a segment. Notwithstanding, SpEx adequately described the structure of audio for Section Selection so hypothesis H1 is satisfied (*"Segments and Word Clouds will help display the structure of speech audio for Section Selection."*).

Fact Finding. For the more precise Fact Finding tasks, the display of Word Clouds and Transcript Markers proved to be effective. Participants used Word Clouds to identify relevant segments which may contain the desired information. Participants could then select relevant Word Cloud words to display Transcript Markers. Transcript Markers would show exactly where the words occurred in the audio. The highest performing participants made good use of the search facility by forming good search queries. The audio thumb which lay on the Treemap could then be dragged with precision over Transcript Markers to accept or reject regions of audio as containing the desired information. For Fact Finding, the ability of SpEx to initially display the topic structure of audio recordings before letting participants drill-down to more precise information allowed participants to skip irrelevant audio and find their information efficiently. I therefore believe that hypothesis H2 is satisfied (*"Transcript Markers and the search facility will help users to locate specific regions in speech audio for Fact Finding."*).

Summarisation. The Summarisation task was found to be more difficult than the Section Selection and Fact Finding tasks. Participants still performed well however. A common strategy of using context sentence

popups coupled with displaying Transcript Markers was utilised. Transcript Markers gave an indication of where certain words commonly reoccurred, a perceived sign of significance of the words. Context sentence popups were misused to access the transcript to pull more information about what was said. A facility to display a readable portion of the transcript may have benefited the Summarisation task by allowing participants to skim the text. Skimming text is a skill more common than skimming audio, and when used with a text transcript could have lead to efficient Summarisation. Although SpEx required improvement to support Summarisation, SpEx was able to assist users to summarise a portion of audio. Hence, hypothesis H3 is satisfied (*"SpEx will help users to comprehend the content of speech audio for Summarisation."*).

Acoustic Event Detection. For Part C, where participants were tasked to find where a student was speaking, the fastest participants made more use of the loudness and pitch acoustic filters than the slowest participants. But participants did not know which cues indicated a speaker change. Participants would explore different visual anomalies of the acoustic underlay until a speaker change was found. While speaker change was identified, I am unclear whether different acoustic events could also be identified. Hence, Hypothesis H4 is yet to be satisfied (*"SpEx will allow users to find changes in acoustic conditions."*).

Feedback. Hypothesis H5 (*"Users will respond positively to SpEx."*) and H6 (*"Users will want to use SpEx in the future."*) are also satisfied because every participant had something nice to say about SpEx. Further, 90% of participants stated they would use SpEx in the future, given the opportunity.

Workload and Usage Strategies. With regard to open question Q1 (*"How do users perceive the workload involved?"*), overall SpEx was found to require

much effort and mental work to use. In turn, SpEx may be unattractive to some and difficult to use when multitasking. However, users found SpEx easy to learn and use because low frustration was reported. Students such as Jack and Amy would prefer a low learning curve to their software and in turn may not be discouraged to use SpEx. Additionally, with little physical effort involved and a medium level of perceived workload, I believe students would not be hesitant to try a new tool such as SpEx when revising from lecture recordings.

Open questions Q2 (*"Which navigation strategies lead to the effective use of SpEx?"*) and Q3 (*"Which elements of SpEx are used most often?"*) remain. Dragging the audio, selecting Word Cloud words, performing searches, and, mostly, viewing context sentence popups were the most frequently used visual elements. Acoustic features were only found useful for locating specific acoustic cues. For locating spoken content, a feature to read the transcript at any point in the audio was desired. Context sentence popups were inappropriately used. Coupled with high tool switching frequencies between context sentence popups, positioning the audio, and selecting Word Cloud words, a tool to view the transcript from any position would greatly reduce the number of interactions required when deciding the relevance of a region of audio.

Segmentation. With regard to the fifteen segments produced by TAFE, participants gave no feedback about the quality or accuracy of the segmentation. The segmentation was therefore adequate for audio retrieval and, while not particularly accurate, broke audio into discrete units which could be consumed by participants. Participants were able to focus their attention on the overall structure of an audio recording, before analysing in more detail particular segments of interest. Word Clouds served to mitigate segmentation inaccuracy by visually indicating whether two neighbouring segments were related or not. Related segments could be analysed together.

I could have attempted to implement a text-based topic modelling segmentation algorithm and augment the algorithm with acoustic features. Then I could have captured text and prosodic events to generate more accurate segments. But, the number of segments produced could significantly impact SpEx. A segmentation algorithm which produced a non-fixed number of segments may hinder SpEx. If too few segments were produced, segments would be too large and wont be able to display enough information about the audio. If too many segments were produced, either there would be too many Word Clouds to be easily assimilated by users or segments would be so small that no Word Clouds could be displayed. By enforcing fifteen segments I could achieve an effective balance between too much and too little information.

7.6 Improvements and Lessons Learned

In light of lessons learned from my user study, I made three improvements to SpEx. First, I increased the amount of text visible when a context sentence popup appeared to make disambiguating words easier. Additionally, participants who used context sentence popups to create a gist of a portion of audio could view more text. By displaying more words, audio should be easier to navigate because a better understanding of the underlying transcript could be gained.

Second, responding to P15's comment, clicking the background now re-positions the audio position. I believed that users could be more efficient by clicking to seek rather than dragging due to the simplified and more accurate interaction. Furthermore, clicking the background seemed to be a natural response for participants. In which case, the learning difficulty of using SpEx could be reduced because fewer accidental mistakes would be made. I added a button to de-select all highlighted Word Cloud words to allow users to still clear the Treemap quickly.

Third and final, I added a new feature called *Transcript Anywhere*. Tran-

script Anywhere was a display which allowed users to read the underlying transcript at any point in the audio. Transcript Anywhere is displayed in Figure 7.6.

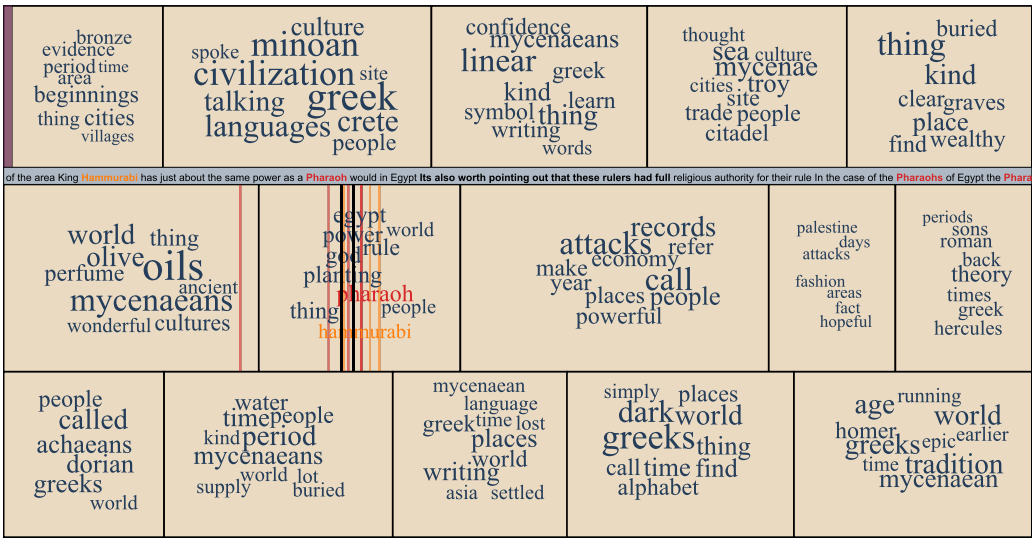


Figure 7.6: Transcript Anywhere. An unobtrusive display to view the transcript at any point in the audio.

By clicking and holding on a position in the Treemap, a display with the text underneath the cursor was displayed. The current sentence was highlighted in bold to provide positional awareness, with neighbouring sentences to the left and right. Moving the mouse to the left and right scrolled the text in a manner similar to stock-tickers. Scroll speed could be increased or decreased by moving the mouse further away or closer to the original mouse point respectively. Clicking produced a black vertical line at the current position and moving the mouse left or right displayed a second black vertical line to show the offset from the current position. Transcript Anywhere also coloured highlighted words to allow users to find the text beneath Transcript Markers.

Transcript Anywhere supported Summarisation by giving users direct access to the underlying transcript in an unobtrusive manner. The remain-

der of SpEx was still visible. With Transcript Anywhere, Summarisation should be an easier task and produce more accurate Summarisation results. Transcript Anywhere would also allow users to quickly read sentences that were just listened to, supporting recap and in-turn increasing user confidence of information that was just listened to.

As a final note, though my user study was able to evaluate the effectiveness of SpEx against my task taxonomy and help to understand the mental workload involved, the effectiveness of SpEx for transferring knowledge was not measured. One important question is whether SpEx promotes 'shallow' or 'robust' learning [7]. A student who exhibits shallow learning is only capable of applying their learning to similar circumstances, while a student who exhibits robust learning is able to adapt their learning to new and more difficult circumstances. Hence, robust learning is preferred. Transcript Markers in SpEx may support shallow learning by making it too easy for students to skip potentially relevant portions of audio that may supplement their learning. In contrast, Word Clouds display which segments may be related and hence may support more robust learning by guiding students to further related information. Performing a question-answer test about the topics presented in the audio before and sometime after my user study may have provided insight into the ability of SpEx to support effective student learning.

Chapter 8

Conclusion

This thesis describes the design, development, and analysis of Speech Explorer (SpEx), a tool to support navigation within spoken audio.

SpEx was developed for the university domain, to support students to find information in audio lecture recordings for revision. SpEx was developed with regard to my developed task taxonomy, a set of common audio retrieval tasks based on the literature: Section Selection, Fact Finding, and Summarisation. A model of target users and personas were also used to guide the design decisions of SpEx. Requirements are discussed in Chapter 3.

The underlying structure for the design of SpEx was a Strip Treemap. Each cell corresponded to a segment of the audio and segments were populated with Word Clouds to describe their content. A search facility to locate Word Cloud words and Transcript Markers to locate specific word occurrences were implemented to allow users to find the information they wanted. Audio controls were also integrated within SpEx for intuitive manipulation and playback of the audio. The design of SpEx is discussed in detail in Chapter 4.

I developed an accompanying application to process audio recordings for SpEx called TAFE, a Text and Audio Feature Extractor. TAFE accepted an audio recording and its transcript as input to produce the required seg-

ments, Word Cloud words, and acoustic features necessary for SpEx to visualise. TAFE was evaluated to measure both execution time and segmentation accuracy. Execution time fell below my five minute threshold. Segment accuracy was not great, but did outperform my baseline segmentation algorithm and was not found to be a hindrance in the user study of SpEx. TAFE is discussed in Chapter 5.

I designed a user study whereby participants (mostly undergraduate students) used SpEx to retrieve information from a series of audio recordings. The tasks corresponded to my task taxonomy. I could measure the performance, mental workload, and usage strategies of participants. My user study was described in Chapter 6.

The user study revealed good results for each task in my task taxonomy. Section Selection (locating regions which correspond to a particular topic) achieved a median precision of 80.72% and a median recall of 93.75%. Fact Finding (locating a specific piece of information) achieved a median accuracy of 100.00% (mean of 92.10%). Summarisation achieved a median quality score of 75.00%. These results indicated that SpEx was effective for audio retrieval, particularly for locating where specific information was found. A high mental workload was identified by participants, but participants did not find SpEx frustrating, which indicated that SpEx was not difficult to use, only that it required notable focus to use. Analysing usage strategies revealed that forming good search queries when finding spoken information and making use of acoustic filters when finding non-spoken information lead to the best performance. Finally, feedback from users were positive and indicated a desire to use SpEx in the future. The user study results are discussed in Chapter 7.

The specific contributions of my work are described below.

8.1 Contributions

My study contributes to the field of speech navigation, particularly in universities, as well as to the field of Information Visualisation. Specific contributions of my thesis are described below:

Task Taxonomy

I developed a task taxonomy of common audio retrieval tasks targeted towards lecture audio. The task taxonomy was modified from existing work by Whittaker et. al. [93] and Dufour et. al. [29]. The task taxonomy consists of three tasks:

- **Section Selection:** Identification of relevant segments of the audio.
- **Fact Finding:** Extracting a specific piece of information from the audio.
- **Summarisation:** Producing a summary of the content of an audio recording.

TAFE

TAFE was a system to extract textual features from transcripts and acoustic features from audio recordings to be used for Information Visualisation purposes. TAFE contained an algorithm to segment audio into a predefined number of segments based on how something was said, rather than what was said. The algorithm clusters loudness, pitch, speech rate, and sentence duration features to generate discrete segments. Features were dimension-reduced by PCA before being clustered into segments by complete-link clustering.

SpEx

SpEx was an interactive Information Visualisation prototype capable of displaying textual and acoustic features about audio recordings. SpEx was designed to adhere to the requirements of my task taxonomy and be respectful of my personas. Indeed, SpEx supported Section Selection, Fact Finding and Summarisation with only the audio and transcript.

SpEx consisted of a novel Strip Treemap organisation of spoken audio. Cells corresponded to audio segments and cell size corresponded to segment duration. Cells were populated with Word Clouds to describe the content of their segments. A search facility was provided to search for Word Cloud words. Users could highlight Word Clouds words to display words in a sentence for context. Users could select Word Cloud words to reveal Transcript Marks. Transcript Marks were vertical colour-coded bars which highlighted precisely where word occurrences were found. Audio controls at the top of the interface coupled with an audio thumb on the Treemap itself allowed users to efficiently navigate to desired locations in the audio. Audio filters were also available to allow users to search for non-spoken cues such as background noise and speaker change.

User Study

I developed a user study to measure the performance of a spoken content retrieval system against my task taxonomy. Participants would attempt to extract information from audio recordings while following a series of questions. Each task in my task taxonomy corresponded to a question type: Section Selection asked where a topic would be found, Fact Finding asked for a specific piece of information, and Summarisation asked to summarise a portion of the audio. Performance, mental demand, and usage strategies could be recorded via marked answers, NASA-TLX, and interaction logs respectfully.

I discuss future work in the following section.

8.2 Future Work

This thesis presented the design, development, and analysis of SpEx, but there were areas that needed improvement and ways SpEx could have been extended. Here, I suggest future work for SpEx:

More accurate segments. With a mean segment accuracy recall of 30.3%, TAFE leaves much room for improvement. A method of identifying more accurate segments may allow users to more effectively browse SpEx to find information more quickly. Possible methods may include topic modelling by analysing what was said and allowing users to adjust or correct the segment boundaries manually.

Better selection of words for Word Clouds. TAFE selected up to ten of the most frequent words in a segment to display as Word Clouds in SpEx. However, alternative methods of picking words may have produced Word Clouds with words more representative of each segment. For example, term frequency-inverse document frequency ($tf.idf$), considers words that occur frequently in the current document (segment in this case) and infrequently in others. Hence subsiding non-representative words.

Better display of segment relevance. Participants tended to over-select segments for Section Selection tasks. A visual cue for SpEx to signify how relevant two segments are may reduce the number of segments incorrectly thought to be related.

Evaluate SpEx when transcripts have errors. The results of my user study were based on the use of SpEx with perfect transcripts. The performance of SpEx when transcripts with errors are used can only be left to speculation. Another user study must be conducted to evaluate SpEx with non-perfect transcripts.

Evaluate SpEx in a field study. SpEx was only evaluated in a laboratory setting, rather than provided for students to use in a real course. While a laboratory study did provide insights into the effectiveness of SpEx, important questions remain unanswered. For instance: Can SpEx effectively be used for course revision? Are their frustrations that occur only after prolonged usage? What additional software and hardware infrastructure is required to deploy SpEx for use in a university?

Support mobile and gestural interfaces. With the popularity of portable technology, web browsers can be found in devises such as smart phones and tablets. Catering for portable devices may increase the utility of SpEx. Further, multi-touch and gestural interfaces are becoming more common. The interaction mechanisms of SpEx should cater for devices without a keyboard and mouse.

Incorporate Video. Audio can frequently be accompanied by a video stream. Incorporating a video into SpEx would allow users to see presentation slides, whiteboard/blackboard diagrams, and physical examples that are otherwise hidden from the audio stream.

I enjoy listening to spoken audio such as podcasts and lectures myself. I hope that SpEx, along with TAFE, can contribute to the wider area of speech navigation. We must disregard spoken audio as being a single, unbreakable, file that must be listened to from end to end and start treating spoken audio as a document that can be skimmed, searched, and browsed.

Appendix A

Human Ethics and Consent Forms

Appendix D

TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI
VICTORIA
UNIVERSITY OF WELLINGTON

HUMAN ETHICS COMMITTEE

Application for Approval of Research Projects

Please write legibly or type if possible. Applications must be signed by supervisor (for student projects) and Head of School

Note: The Human Ethics Committee attempts to have all applications approved within three weeks but a longer period may be necessary if applications require substantial revision.

1. NATURE OF PROPOSED RESEARCH:

(a) Student Research

(b) If Student Research .....

Degree: Master of Engineering (ME): Software Engineering (SWEN)

Course Code: ENGR 591

(c) Project Title: Visualisation and Navigation of Speech Audio

2. INVESTIGATORS:

(a) Principal Investigator

Name: Fahmi Abdulhamid

Email address: fahmi.abdulhamid@ecs.vuw.ac.nz

School/Dept/Group: School of Engineering and Computer Science (ECS)

(b) Other Researchers Name Position

None

Supervisor (in the case of student research projects)

Dr. Stuart Marshall

3. DURATION OF RESEARCH

(a) Proposed starting date for data collection: 17th of September 2012

(**Note:** that NO part of the research requiring ethical approval may commence prior to approval being given)

(b) Proposed date of completion of project as a whole: 31st of March 2013

4. PROPOSED SOURCE/S OF FUNDING AND OTHER ETHICAL CONSIDERATIONS

(a) Sources of funding for the project

Please indicate any ethical issues or conflicts of interest that may arise because of sources of funding e.g. restrictions on publication of results

Vouchers from ECS may be given to participants.

(b) Is any professional code of ethics to be followed **Y ☐ N ✓**

If yes, name

.....

(c) Is ethical approval required from any other body **Y ☐ N ✓**

If yes, name and indicate when/if approval will be given

.....

5. DETAILS OF PROJECT

Briefly Outline:

(a) The objectives of the project

To understand how existing information visualisation techniques may be used to help people better comprehend the structure of a voice recording to locate the information they want.

(b) Method of data collection

User testing. Users will be recorded via screen capture as they complete a series of common audio-related search tasks. A questionnaire will also be filled out to record user thoughts.

(c) The benefits and scientific value of the project

To improve the consumption of recorded speech audio by visually displaying the audio and thus reducing the time it takes to find the information that is desired.

(d) Characteristics of the participants

About 20 participants, largely university students from the School of Engineering and Computer Science. Participants will be between the ages of 20 and 30 years and will be able to give consent to participate in the user study.

(e) Method of recruitment

Email notification and word of mouth.

(f) Payments that are to be made/expenses to be reimbursed to participants

Vouchers from ECS may be given to participants.

(g) Other assistance (e.g. meals, transport) that is to be given to participants

None

(h) Any special hazards and/or inconvenience (including deception) that participants will encounter

None

(i) State whether consent is for (delete where not applicable):

(i) the collection of data

(iii) release of data to others

(iv) use for a conference report or a publication

Attach a copy of any questionnaire or interview schedule to the application

See the provided Task Sheet.

(j) How is informed consent to be obtained (see sections 4.1, 4.5(d) and 4.8(g) of the Human Ethics Policy)

(i) the research is strictly anonymous, an information sheet is supplied and informed consent is implied by voluntary participation in filling out a questionnaire for example (include a copy of the information sheet) Y □ N ✓

(ii) the research is not anonymous but is confidential and informed consent will be obtained through a signed consent form (include a copy of the consent form and information sheet) Y ✓ N □

(iii) the research is neither anonymous or confidential and informed consent will be obtained through a signed consent form (include a copy of the consent form and information sheet) Y □ N ✓

(iv) informed consent will be obtained by some other method (please specify and provide details) Y □ N ✓

.....

.....

With the exception of anonymous research as in (i), if it is proposed that written consent will not be obtained, please explain why

.....

.....

(k) If the research will not be conducted on a strictly anonymous basis state how issues of confidentiality of participants are to be ensured if this is intended. (See section 4..1(e) of the Human Ethics Policy). (E.g. who will listen to tapes, see questionnaires or have access to data). Please ensure that you distinguish clearly between anonymity and confidentiality. Indicate which of these are applicable.

(i) access to the research data will be restricted to the investigator Y □ N ✓

(ii) access to the research data will be restricted to the investigator and their supervisor (student research) Y □ N ✓

(iii) all opinions and data will be reported in aggregated form in such a way that individual persons or organisations are not Y ✓ N □

identifiable

(iv) Other (please specify)

For the purpose of reproducibility of my work, I intend to provide the data publicly with all features identifying participants removed. Personal details of the participants will be omitted. Each participant will be identified by a unique ID such that the ID could not be used to discover the individual. Any quotes from participants in publications will be referenced to the respective unique ID.

(l) Procedure for the storage of, access to and disposal of data, both during and at the conclusion of the research. (see section 4.12 of the Human Ethics Policy). Indicate which are applicable:

(i) all written material (questionnaires, interview notes, etc) will be kept in a locked file and access is restricted to the investigator Y □ N ✓

(ii) all electronic information will be kept in a password-protected file and access will be restricted to the investigator Y □ N ✓

(iii) all questionnaires, interview notes and similar materials will be destroyed:

(a) at the conclusion of the research Y □ N ✓

(b).....years after the conclusion of the research; or Y □ N ✓

(iv) any audio or video recordings will be returned to participants and/or electronically wiped Y □ N ✓

(v) other procedures (please specify):

I will keep an anonymised database of participant results to allow others to reproduce and repeat the science. I will provide public access to the database.

If data and material are not to be destroyed please indicate why and the procedures envisaged for ongoing storage and security

I cannot guarantee all material will be destroyed as the material will be shared publicly. All material may be made public via a website maintained by my primary supervisor such that participants will be unidentifiable.

(m) Feedback procedures (See section 7 of Appendix 1 of the Human Ethics Policy). You should indicate whether feedback will be provided to participants and in what form. If feedback will not be given, indicate the reasons why.

The consent form will give participants the option to be notified of any publications made containing the results. Participants are also free to contact me to view their individual results.

(n) Reporting and publication of results. Please indicate which of the following are appropriate. The proposed form of publications should be indicated on the information sheet and/or consent form.

(i) publication in academic or professional journals Y ✓ N □

(ii) dissemination at academic or professional conferences Y ✓ N □

(iii) deposit of the research paper or thesis in the University Library (student research) Y ✓ N □

(iv) other (please specify) Y ✓ N □

Participant quotes may be used in publications and attributed to the respective unique IDs to keep the participants anonymous. Anonymised results will be made public via a website maintained by the primary supervisor.

Signature of investigators as listed on page 1 **(including supervisors) and Head of School**.

NB: All investigators and the Head of School must sign before an application is submitted for approval

..... Date.....

..... Date.....

..... Date.....

Head of School:

..... Date.....

APPLICATIONS FOR HUMAN ETHICS APPROVAL

CHECKLIST

- Have you read the Human Ethics Policy?
- Is ethical approval required for your project?
- Have you established whether informed consent needs to be obtained for your project?
- In the case of student projects, have you consulted your supervisor about any human ethics implications of your research?
- Has your supervisor read and signed the application?
- Have you included an information sheet for participants which explains the nature and purpose of your research, the proposed use of the material collected, who will have access to it, whether the data will be kept confidential to you, how anonymity or confidentiality is to be guaranteed?
- Have you included a written consent form?
- If not, have you explained on the application form why you do not need to get written consent?
- Are you asking participants to give consent to:
  - collect data from them
  - attribute information to them
  - release that information to others
  - use the data for particular purposes
- Have you indicated clearly to participants on the information sheet or consent form how they will be able to get feedback on the research from you (e.g. they may tick a box on the consent form indicating that they would like to be sent a summary), and how the data will be stored or disposed of at the conclusion of the research?
- Have you included a copy of any questionnaire or interview checklist you propose using?
- Has your application been seen by the head of your school or department (or the person given responsibility to consider applications on behalf of the head (see section 4.5(b) of the Human Ethics Policy).

PLEASE FORWARD YOUR COMPLETED APPLICATION FORM TO THE SECRETARY, HUMAN ETHICS COMMITTEE OR, IN THE CASE OF APPLICATIONS FROM SCHOOLS OR DEPARTMENTS WITH AN APPROVED ETHICS SUB-COMMITTEE, TO THE CONVENER OF THAT SUB-COMMITTEE

TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI
VICTORIA
UNIVERSITY OF WELLINGTON

School of Engineering and Computer Science

Visualisation and Navigation of Speech Audio

User Study Consent Form

ID:

Project Details

This user study is part of a project that aims to understand how to better navigate audio recordings.

The project and user study are carried out by Fahmi Abdulhamid (the researcher) and supervised by Dr. Stuart Marshall, both from the School of Engineering and Computer Science at Victoria University of Wellington.

Consent

I have been provided with and understand the information relating to the nature and objectives of this research project, and I have been given the opportunity to seek further clarification and explanation to my satisfaction.

I understand that any information I provide or is recorded via screen capture will be anonymised. The information will not contain my name and no opinions will be attributed to me in any way that will identify me.

I understand that the anonymised results of this user study may be published in conference and journal publications, will contribute to a Masters Thesis, and will be made available on a public website.

I understand that I may at any time view and discuss the information relating to my participation in the user study.

I understand that I may withdraw my results from the user study by giving notice before 1st November 2012.

I agree to complete the provided questions and post-study questionnaire truthfully.

I have read and understood the accompanying information sheet and I agree to all of the above.

Name: Date:

Signed:

By providing my email address I choose to be notified of publications which have made use of information from this user study:

Email address:

TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI
VICTORIA
UNIVERSITY OF WELLINGTON

School of Engineering and Computer Science

Visualisation and Navigation of Speech Audio

Information Sheet

About the Project

The objective of the project is to make the information in recorded speeches more accessible by understanding how existing information visualisation techniques can be applied to audio. The project is developed by Fahmi Abdulhamid in fulfilment of a Masters by Thesis. Fahmi is a student in the School of Engineering and Computer Science at Victoria University of Wellington.

By making audio easier to consume, we can provide better access to lectures, podcasts, and presentations as well as make existing digital audio archives more accessible.

Participation in the User Study

This user study will have you use a visualisation tool to find information in audio recordings. You will be given a question sheet to work through during the study and a post-study questionnaire after the study. Your questionnaire answers will be stored and your actions with the visualisation tool will be recorded via a screen recorder and by the visualisation tool itself. The user study will take approximately 40 – 50 minutes.

The user study has been approved by the Human Ethics Committee. Your participation will remain confidential. You will be given a unique ID to which your answers and recordings will be attributed to. The results may appear in aggregate form in conference and journal publications as well as a Masters Thesis with the exception of free-form answers which may be quoted to your unique ID. The Masters thesis will be deposited in the university library. If you like, you may at any time view and discuss information relating to *your* participation and may opt to be notified of any publications which have included material from this user study. Further, having completed the user study, you may choose to withdraw and have your results and record of participation deleted provided you give notice before the 1st of November 2012.

All data from this user study will be made available on a public website for the purpose of reproducing the results of my research. The data will have all identifying features removed before being made public. You will not be identifiable by the data.

Contact Details

Student
Fahmi Abdulhamid
Masters Student
School of Engineering and Computer Science
Victoria University of Wellington
Phone: 021 296 5698
Email: fahmi.abdulhamid@ecs.vuw.ac.nz

Supervisor
Dr. Stuart Marshall
Senior Lecturer
School of Engineering and Computer Science
Victoria University of Wellington
Phone: +64 4 463 6730
Email: stuart.marshall@ecs.vuw.ac.nz

Appendix B

User Study Question Sheet and Answers

TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI
VICTORIA
UNIVERSITY OF WELLINGTON

School of Engineering and Computer Science

Visualisation and Navigation of Speech Audio

Task Sheet

ID: **Date:**

Age (circle one): <18 18-21 22-25 26-30 31-40 41-50 51-60 61+

Male / Female (circle one)

Occupation or degree:

Instructions

The user study is split into two rounds: the **Practice round** and the **User Study round**.

The **Practice round** will give you a chance to learn how to use the visualisation and provide some sample tasks to complete if you wish.

When you are comfortable using the visualisation, you may begin the **User Study round** where your performance is measured against set tasks. The User Study round consists of two parts (A and B) and may be presented to you in a random order.

After each round, you will fill out a performance rating scale. After all rounds are complete, you will have the chance to provide feedback.

Practice Round

A chance to learn how to use the visualisation. Answer the following questions in the lecture recording:

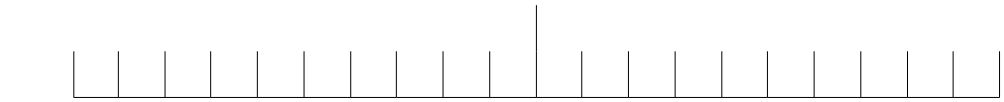
Frontiers and Controversies in Astrophysics

1. How many mid-term exams are there in the class?
2. In which segments would you find information on epicycles?
3. Summarise the content of the two left-most segments of the top row.

Please fill in the rating scale below after the Practice round

MENTAL DEMAND

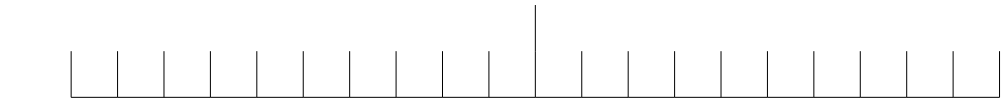
How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?



Low **High**

PHYSICAL DEMAND

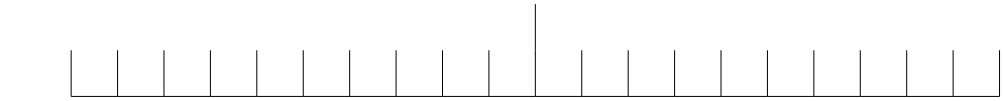
How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?



Low **High**

TEMPORAL DEMAND

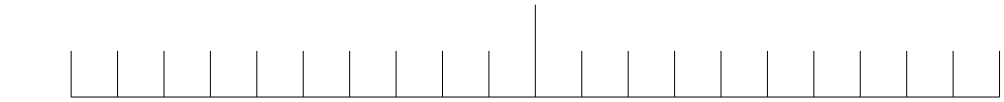
How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?



Low **High**

PERFORMANCE

How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?

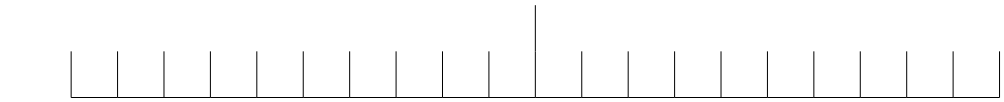


Good **Poor**

EFFORT

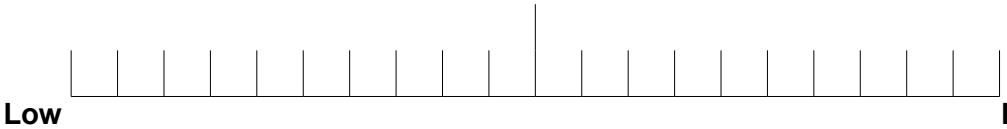
How hard did you have to work (mentally and physically) to accomplish your level of performance?

Low **High**



FRUSTRATION

How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

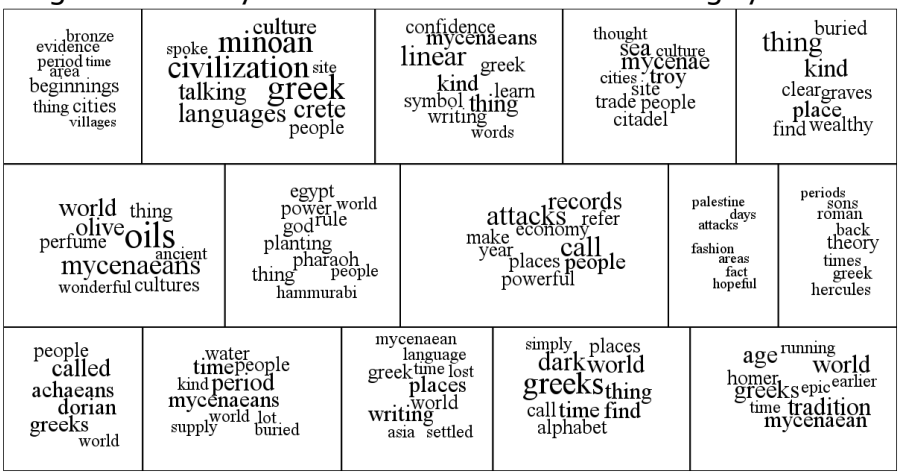


Low **High**

User Study Part A: Introduction to Ancient Greek History

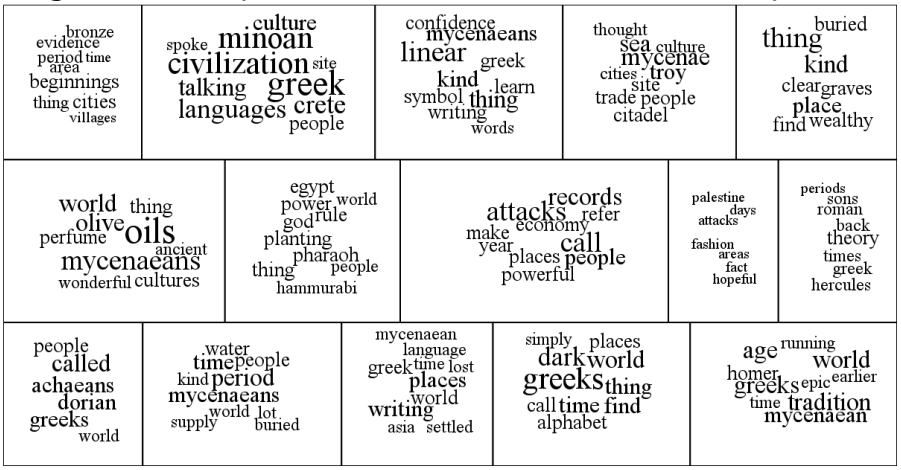
1. When defining civilisation, what is the difference between a city and a village according to the speaker?

2. In which segments would you look to find information on writing symbols?



3. Name one attacker of Egypt the speaker mentions.

4. In which segments would you look to find information on Homer's epics?

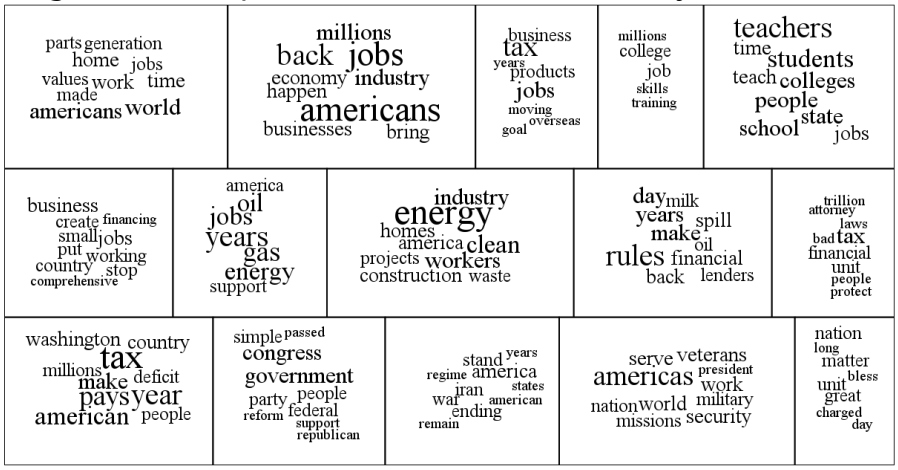


5. Summarise the content of the two right-most segments of the top row.

User Study Part B: State of the Union 2012

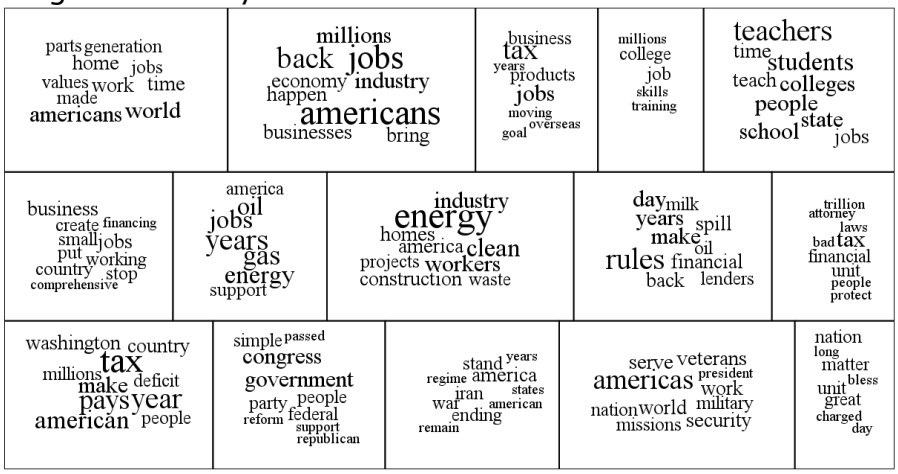
1. According to Obama, why do teachers matter?

2. In which segments would you look to find information about jobs?



3. What does Obama want his administration to do with offshore oil and gas resources?

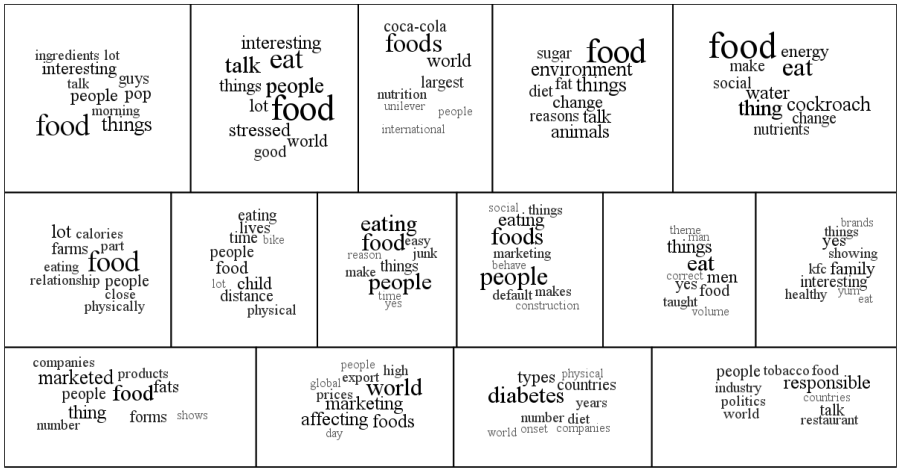
4. In which segments would you look to find information on tax?



5. Summarise the content of the two right-most segments of the bottom row.

User Study Part C: The Psychology, Biology and Politics of Food

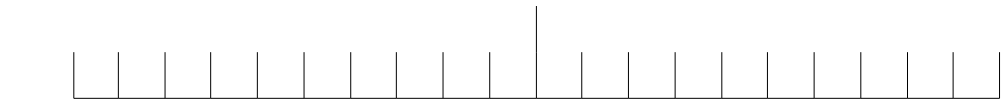
1. Please mark as closely as possible one location where a student from the audience is talking.



Please fill in the rating scale below after the User Study Parts A and B

MENTAL DEMAND

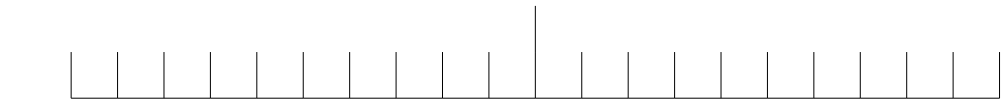
How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?



Low **High**

PHYSICAL DEMAND

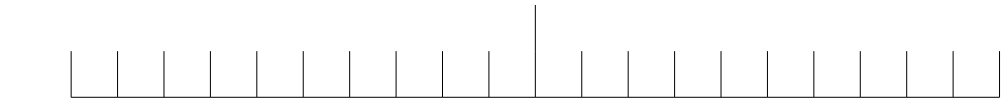
How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?



Low **High**

TEMPORAL DEMAND

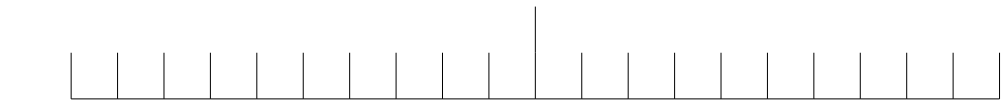
How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?



Low **High**

PERFORMANCE

How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?

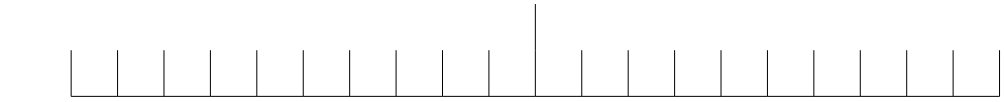


Good **Poor**

EFFORT

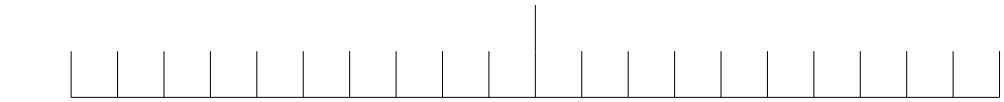
How hard did you have to work (mentally and physically) to accomplish your level of performance?

Low **High**



FRUSTRATION

How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?



Low **High**

Please fill in the workload comparisons below after the User Study round

For each pair below, circle the workload which you consider to have contributed more to the tasks.

Effort	or	Performance
Temporal Demand	or	Frustration
Temporal Demand	or	Effort
Physical Demand	or	Frustration
Performance	or	Frustration
Physical Demand	or	Temporal Demand
Physical Demand	or	Performance
Temporal Demand	or	Mental Demand
Frustration	or	Effort
Performance	or	Mental Demand
Performance	or	Temporal Demand
Mental Demand	or	Effort
Mental Demand	or	Physical Demand
Effort	or	Physical Demand
Frustration	or	Mental Demand

Mental Demand: How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?

Physical Demand: How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?

Temporal Demand: How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?

Performance: How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?

Effort: How hard did you have to work (mentally and physically) to accomplish your level of performance?

Frustration: How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

Please provide additional feedback after the User Study round

Which parts of the visualisation did you **like**, and why?

Which parts of the visualisation **need improvement**, and why?

Please provide any **further comments** you may have:

Would you use this visualisation in the future?

TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI
VICTORIA
UNIVERSITY OF WELLINGTON

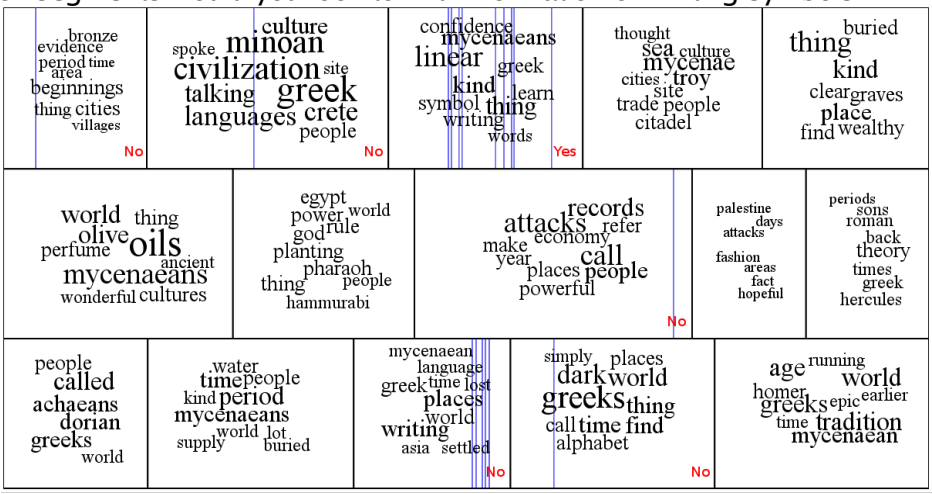
School of Engineering and Computer Science

Visualisation and Navigation of Speech Audio

Answer Sheet

User Study Part A: Introduction to Ancient Greek History

1. When defining civilisation, what is the difference between a city and a village according to the speaker?
   "...a city contains a number of people who do not provide for their own support. That is to say, they don't produce food. They need to acquire it from somebody else. Instead, they do various things like govern and are priests, and are bureaucrats, and are engaged in other non-productive activities that depend upon others to feed them. That's the narrowest definition of cities. Of course, with cities we typically find a whole association of cultural characteristics, which we deem civilization."
2. In which segments would you look to find information on writing symbols?



(Blue lines indicate positions of "writing" and "symbol". Red words indicate relevance of segments)

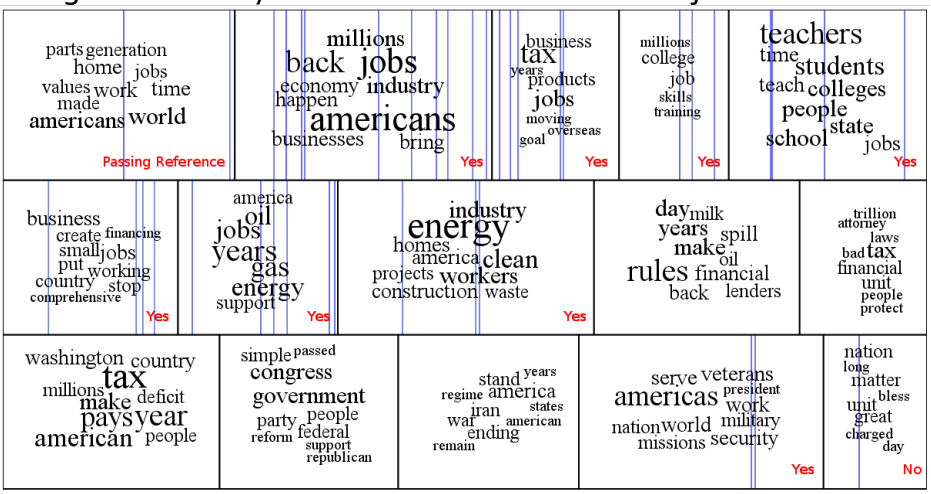
3. Name one attacker of Egypt the speaker mentions.
   "...there are attackers from Libya, we hear, but there are also attackers that are simply called from the sea; the sea people attack."
4. In which segments would you look to find information on Homer's epics?

(Blue lines indicate positions of "homer" and "epic". Red words indicate relevance of segments)

5. Summarise the content of the two right-most segments of the top row.
   Donald Kagan discusses the culture and structure of the city of Mycenae (and other such citadels), a city once thought to be fictional but was discovered in the 1800s. Mycenae was built on a formidable hill which was close, but not on, the water and surrounded by farmland. Keeping distance form the water insured security from sea attackers while providing access to sea trade. Donald Kagan goes on to discuss the evidence for Mycenae's power and wealth: Kings were buried in large underground tombs built at an enormous expense. Furthermore, the bodies of royalty were surrounded by expensive burial items.

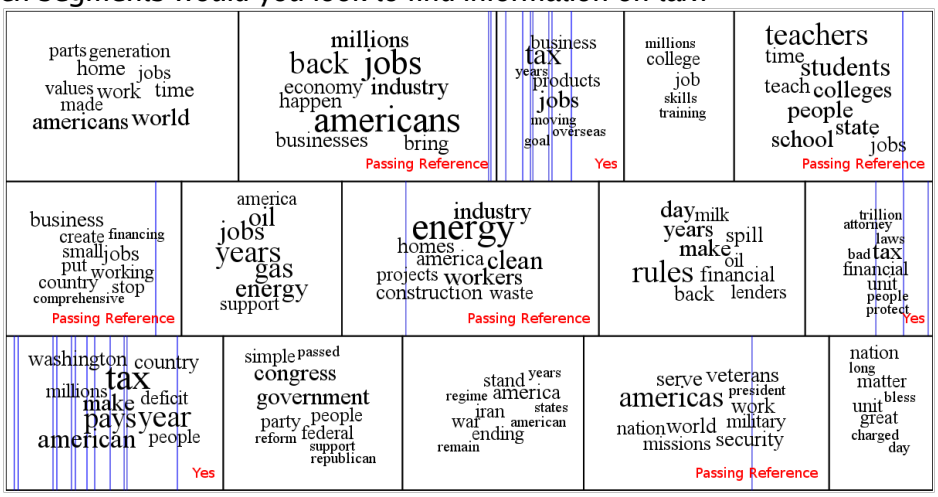
User Study Part B: State of the Union 2012

1. According to Obama, why do teachers matter?
   "We know a good teacher can increase the lifetime income of a classroom by over $250,000. A great teacher can offer an escape from poverty to the child who dreams beyond his circumstance. Every person in this chamber can point to a teacher who changed the trajectory of their lives."
2. In which segments would you look to find information about jobs?



(Blue lines indicate positions of "job". Red words indicate relevance of segments)

3. What does Obama want his administration to do with offshore oil and gas resources?
   "...I'm directing my administration to open more than 75% of our potential offshore oil and gas resources."
4. In which segments would you look to find information on tax?

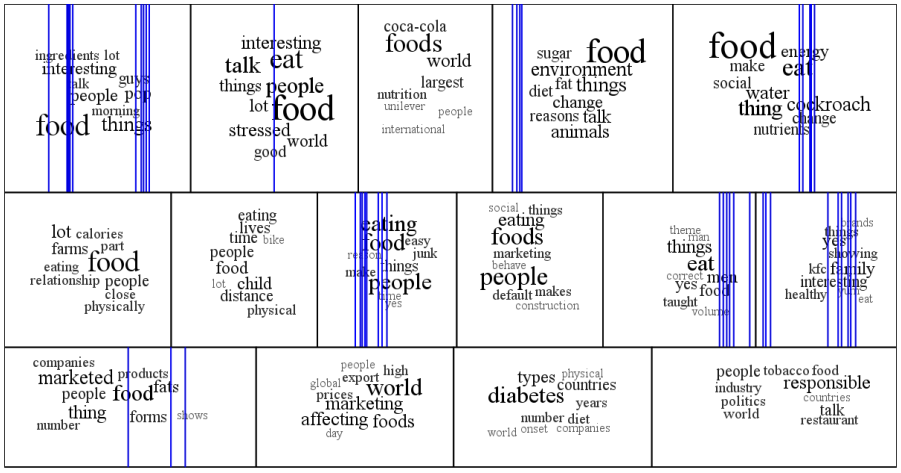


(Blue lines indicate positions of "tax" Red words indicate relevance of segments)

5. Summarise the content of the two right-most segments of the bottom row.
   Barack Obama highlights the international alliances and influence of America. From America's commitment to Israel's security, to missions against hunger and disease, and the moral example America has set for others. This is followed by America's commitment to its military and its support to help veterans find new jobs in America. Barack Obama concludes with an example of an anti-terrorism mission to highlight the unity and trust solders give to each other as an example for a prosperous America.

User Study Part C: The Psychology, Biology and Politics of Food

1. Please mark as closely as possible one location where a student from the audience is talking.



(Blue lines indicate positions where a student is speaking)

Appendix C

User Study Raw Results

Table C.1: User study participant results for Parts A, B, and C. Legend: Yes/Correct (Y), No/Incorrect (N), Not Available (N/A), Precision (Pre.) , and Recall (Rec.).

	Part A							Part B							Part C
	Q1	Q2 (%)		Q3	Q4 (%)		Q5	Q1	Q2 (%)		Q3	Q4 (%)		Q5	Q1
ID	Correct	Pre.	Rec.	Correct?	Pre.	Rec.	Score	Correct?	Pre.	Rec.	Correct?	Pre.	Rec.	Score	Correct?
P1	Y	100	100	Y	100	100	3	Y	100	75	Y	100	66.67	4	N/A
P2	Y	16.67	100	Y	100	100	4	Y	80	100	Y	42.86	100	3	N/A
P3	Y	50	100	Y	20	100	2	Y	100	50	Y	75	100	2	N/A
P4	Y	100	100	Y	100	100	1	Y	100	75	N	100	100	2	N/A
P5	Y	33.33	100	Y	50	100	2	Y	88.89	100	Y	33.33	100	3	N/A
P6	Y	50	100	N	100	100	2	Y	100	37.5	Y	100	66.67	3	N/A
P7	Y	0	0	Y	100	100	2	N	100	75	Y	100	66.67	2	N/A
P8	Y	100	100	Y	100	100	2	Y	100	75	Y	100	33.33	2	N/A
P9	Y	100	100	Y	100	100	3	Y	80	100	Y	42.86	100	3	N/A
P10	Y	50	100	Y	100	100	3	Y	100	75	Y	100	100	2	Y
P11	Y	100	100	Y	100	100	3	Y	100	37.5	Y	100	66.67	3	Y
P12	Y	50	100	N	100	100	2	Y	100	75	Y	100	100	3	Y
P13	Y	0	0	Y	100	100	4	Y	100	12.5	Y	100	33.33	4	Y
P14	Y	50	100	Y	25	100	3	Y	100	75	Y	100	100	3	Y
P15	Y	20	100	Y	14.29	100	3	Y	80	50	Y	42.86	100	4	Y
P16	N	100	100	Y	100	100	3	Y	85.71	75	Y	100	100	3	Y
P17	Y	100	100	Y	100	100	2	Y	85.71	75	Y	100	100	2	Y
P18	Y	33.33	100	Y	100	100	4	Y	100	75	Y	60	100	3	Y
P19	Y	100	100	N	100	100	1	Y	N/A	N/A	N	N/A	N/A	1	Y
P20	Y	100	100	Y	20	100	4	Y	80	100	N	50	100	3	Y

Table C.2: User study participant results for NASA Task Load Index (NASA-TLX). Legend: Mental Demand (MD), Physical Demand (PD), Temporal Demand (TD), Performance (P), Effort (E), and Frustration (F).

	Rating Scale (out of 100)						Workload Comparisons (out of 5)						Calculated Weighted Workloads						
ID	MD	PD	TD	P	E	F	MD	PD	TD	P	E	F	MD	PD	TD	P	E	F	Overall
P1	30	40	15	15	35	10	2	2	4	5	2	0	60	80	60	75	70	0	23.000
P2	80	15	15	35	60	15	5	1	4	3	2	0	400	15	60	105	120	0	46.667
P3	65	75	40	20	60	30	5	3	1	4	2	0	325	225	40	80	120	0	52.667
P4	45	5	40	65	30	15	4	0	3	5	2	1	135	0	120	325	60	15	43.667
P5	90	20	55	25	75	45	4	1	3	5	2	0	360	20	165	125	150	0	54.667
P6	50	15	15	20	30	5	2	1	4	5	3	0	100	15	60	100	90	0	24.333
P7	15	15	40	25	15	15	1	3	4	5	2	0	15	45	160	125	30	0	25.000
P8	50	30	50	60	50	20	4	0	3	5	1	2	200	0	150	300	50	40	49.333
P9	75	5	70	15	75	10	5	0	2	4	3	1	375	0	140	60	225	10	54.000
P10	60	10	55	25	45	30	4	0	2	5	1	3	240	0	110	125	45	90	40.667
P11	35	5	20	15	40	10	4	0	2	3	5	1	140	0	40	45	200	10	29.000
P12	15	10	45	10	25	10	4	0	1	4	2	4	60	0	45	40	50	40	15.667
P13	35	10	20	15	15	15	5	2	0	3	4	1	175	20	0	45	60	15	21.000
P14	35	70	35	5	35	20	2	4	2	3	4	0	70	280	70	15	140	0	38.333
P15	45	15	20	15	20	20	3	0	4	2	4	2	135	0	80	30	80	40	24.333
P16	55	45	20	20	60	10	4	1	3	3	4	0	220	45	60	60	240	0	41.667
P17	25	15	25	25	25	25	5	1	1	3	4	1	125	15	25	75	100	25	24.333
P18	45	5	55	50	30	25	4	0	3	4	3	1	180	0	165	200	90	25	44.000
P19	20	10	30	5	5	5	4	1	2	5	3	0	80	10	60	25	15	0	12.667
P20	40	25	45	70	35	15	4	2	5	3	1	0	160	50	225	210	35	0	45.333

Bibliography

[1] AIGNER, W., MIKSCH, S., MÜLLER, W., SCHUMANN, H., AND TOMINSKI, C. Visualizing time-oriented data – a systematic view. *Comput. Graph. 31*, 3 (June 2007), 401 – 409.

[2] ALBERTI, C., BACCHIANI, M., BEZMAN, A., CHELBA, C., DROFA, A., LIAO, H., MORENO, P., POWER, T., SAHUGUET, A., SHUGRINA, M., AND SIOHAN, O. An audio indexing system for election video material. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on* (Apr. 2009), pp. 4873 – 4876.

[3] ALIAS-I. Lingpipe 4.1.0, 2011. `http://alias-i.com/lingpipe` (Accessed December 7, 2012).

[4] ANDARGOR. Javascript implementation of the porter stemming algorithm, 2009. (The Porter Stemming Algorithm), `http://tartarus.org/martin/PorterStemmer/` (Accessed November 13, 2012). Revised by Christopher McKenzie.

[5] ARONS, B. SpeechSkimmer: Interactively skimming recorded speech. In *Proceedings of the 6th annual ACM symposium on User interface software and technology* (New York, NY, USA, 1993), UIST '93, ACM, pp. 187 – 196.

[6] BAILYN, C. ASTR 160: Frontiers and controversies in astrophysics., 2007. (Yale University: Open Yale Courses), `http://oyc.yale.`

edu (Accessed August 28, 2012). License: Creative Commons BY-NC-SA.

[7] BAKER, R. S., GOWDA, S. M., CORBETT, A. T., AND OCUMPAUGH, J. Towards automatically detecting whether student learning is shallow. In *Intelligent Tutoring Systems*, S. Cerri, W. Clancey, G. Papadourakis, and K. Panourgia, Eds., vol. 7315 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2012, pp. 444 – 453.

[8] BANERJEE, A., FALOUTSOS, M., AND BHUYAN, L. Profiling podcast-based content distribution. In *INFOCOM Workshops 2008, IEEE* (Apr. 2008), pp. 1 – 6.

[9] BARNES, C., GOLDMAN, D. B., SHECHTMAN, E., AND FINKELSTEIN, A. Video tapestries with continuous temporal zoom. *ACM Trans. Graph. 29* (July 2010), 89:1 – 89:9.

[10] BATES, M. J. What is browsing — really? a model drawing from behavioural science research. *Information Research 12*, 4 (2007). paper 330.

[11] BEDERSON, B. B., SHNEIDERMAN, B., AND WATTENBERG, M. Ordered and quantum treemaps: Making effective use of 2D space to display hierarchies. *ACM Trans. Graph. 21*, 4 (Oct. 2002), 833 – 854.

[12] BERGERVOET, E. J. Visualization of speech content in search results. In *Proceedings of the 6th Twente Student Conference on IT* (Feb. 2007). http://referaat.cs.utwente.nl/conference/6/paper/6795/visualization-of-speech-content-in-search-results.pdf.

[13] BESSER, J., LARSON, M., AND HOFMANN, K. Podcast search: User goals and retrieval technologies. *Online Information Review 34*, 3 (2010), 395 – 419.

[14] BLEI, D. M., NG, A. Y., AND JORDAN, M. I. Latent dirichlet allocation. *Journal of Machine Learning Research 3* (Mar. 2003), 993 – 1022.

[15] BOSTOCK, M., OGIEVETSKY, V., AND HEER, J. D^3; data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on 17*, 12 (Dec. 2011), 2301 – 2309.

[16] BROSSIER, P. The Aubio library at MIREX 2006. `http://www.music-ir.org/evaluation/MIREX/2006_abstracts/AME_BT_OD_TE_brossier.pdf`, Oct. 2006.

[17] BROWNELL, K. D. PSYC 123: The psychology, biology and politics of food., 2008. (Yale University: Open Yale Courses), `http://oyc.yale.edu` (Accessed October 19, 2012). License: Creative Commons BY-NC-SA.

[18] BUXTON, W. *Sketching User Experiences: Getting the Design Right and the Right Design*. Interactive Technologies. Elsevier/Morgan Kaufmann, 2007.

[19] CARD, S., MACKINLAY, J., AND SHNEIDERMAN, B. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Series in Interactive Technologies. Morgan Kaufmann Publishers, 1999.

[20] CHEN, C. *Information Visualization: Beyond the Horizon*. Springer, 2006.

[21] CHINCHOR, N., THOMAS, J., WONG, P., CHRISTEL, M., AND RIBARSKY, W. Multimedia analysis + visual analytics = multimedia analytics. *Computer Graphics and Applications, IEEE 30*, 5 (Sept. – Oct. 2010), 52 – 60.

[22] COLLINS, C., CARPENDALE, S., AND PENN, G. DocuBurst: Visualizing document content using language structure. *Computer Graphics Forum 28*, 3 (2009), 1039 – 1046.

[23] COOPER, A., REIMANN, R., AND CRONIN, D. *About Face 3: The Essentials of Interaction Design*. Wiley, 2007.

[24] CROCKFORD, D. The application/json media type for javascript object notation (JSON). *RFC4627*, 4627 (2006), 1 – 10.

[25] CUI, W., LIU, S., TAN, L., SHI, C., SONG, Y., GAO, Z., TONG, X., AND QU, H. TextFlow: Towards better understanding of evolving topics in text. *Visualization and Computer Graphics, IEEE Transactions on 17*, 12 (Dec. 2011), 2412 – 2421.

[26] CUI, W., WU, Y., LIU, S., WEI, F., ZHOU, M., AND QU, H. Context-preserving, dynamic word cloud visualization. *Computer Graphics and Applications, IEEE 30*, 6 (Nov. – Dec. 2010), 42 – 53.

[27] DE CHEVEIGNÉ, A., AND KAWAHARA, H. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America 111*, 4 (2002), 1917 – 1930.

[28] DREHER, M. J., AND GUTHRIE, J. T. Cognitive processes in textbook chapter search tasks. *Reading Research Quarterly 25*, 4 (1990), 323 – 339.

[29] DUFOUR, C., BARTLETT, J. C., AND TOMS, E. G. Understanding how webcasts are used as sources of information. *Journal of the American Society for Information Science and Technology 62*, 2 (2011), 343 – 362.

[30] FELDMAN, R., AUMANN, Y., FRESKO, M., LIPHSTAT, O., ROSENFELD, B., AND SCHLER, Y. Text mining via information extraction. In *Principles of Data Mining and Knowledge Discovery*, J. M. Żytkow and J. Rauch, Eds., vol. 1704 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 1999, pp. 165 – 173.

[31] FIETZE, S. Podcast in higher education: Students usage behaviour. In *Same places, different spaces. Proceedings ascilite Auckland 2009* (2009), pp. 314 – 318.

[32] FIETZE, S. Podcast in higher education: Students' experience and assessment. In *e-Education, e-Business, e-Management, and e-Learning, 2010. IC4E '10. International Conference on* (Jan. 2010), pp. 12 – 16.

[33] FULLER, M., TSAGKIAS, M., NEWMAN, E., BESSER, J., LARSON, M., JONES, G., AND DE RIJKE, M. Using term clouds to represent segment-level semantic content of podcasts. In *2nd SIGIR Workshop on Searching Spontaneous Conversational Speech (SSCS 2008)* (Singapore, July 2008).

[34] FURUI, S., AND KAWAHARA, T. Transcription and distillation of spontaneous speech. In *Springer Handbook of Speech Processing*, J. Benesty, M. Sondhi, and Y. Huang, Eds. Springer Berlin Heidelberg, 2008, pp. 627 – 652.

[35] GAUGHAN, G., SMEATON, A. F., GURRIN, C., LEE, H., AND MCDONALD, K. Design, implementation and testing of an interactive video retrieval system. In *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval* (New York, NY, USA, 2003), MIR '03, ACM, pp. 23 – 30.

[36] HART, S. G. NASA-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 50*, 9 (2006), 904 – 908.

[37] HAUBOLD, A., AND KENDER, J. R. Augmented segmentation and visualization for presentation videos. In *Proceedings of the 13th annual ACM international conference on Multimedia* (New York, NY, USA, 2005), MULTIMEDIA '05, ACM, pp. 51 – 60.

[38] HAVRE, S., HETZLER, B., AND NOWELL, L. ThemeRiver: Visualizing theme changes over time. In *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on* (2000), pp. 115 – 123.

[39] HEARST, M. *Search User Interfaces*. Search User Interfaces. Cambridge University Press, 2009.

[40] HEARST, M. A. TileBars: Visualization of term distribution information in full text information access. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 1995), CHI '95, ACM Press/Addison-Wesley Publishing Co., pp. 59 – 66.

[41] HINDUS, D., SCHMANDT, C., AND HORNER, C. Capturing, structuring, and representing ubiquitous audio. *ACM Trans. Inf. Syst. 11*, 4 (Oct. 1993), 376 – 400.

[42] HIRSCHBERG, J. Communication and prosody: Functional aspects of prosody. *Speech Communication 36*, 1 – 2 (2002), 31 – 43. ESCA Workshop on Dialogue and Prosody, September 1999.

[43] HOLT, E. The last laugh: Shared laughter and topic termination. *Journal of Pragmatics 42*, 6 (2010), 1513 – 1525.

[44] HORNECKER, E., AND NICOL, E. What do lab-based user studies tell us about in-the-wild behavior? Insights from a study of museum interactives. In *Proceedings of the Designing Interactive Systems Conference* (New York, NY, USA, 2012), DIS '12, ACM, pp. 358 – 367.

[45] HÜRST, W. User interfaces for speech-based retrieval of lecture recordings. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2004* (Lugano, Switzerland, 2004), L. Cantoni and C. McLoughlin, Eds., AACE, pp. 4470 – 4477.

[46] HÜRST, W., KREUZER, T., AND WIESENHÜTTER, M. A qualitative study towards using large vocabulary automatic speech recognition to index recorded presentations for search and access over the web. In *ICWI* (2002), pp. 135 – 143.

[47] HÜRST, W., LAUER, T., AND GÖTZ, G. An elastic audio slider for interactive speech skimming. In *Proceedings of the third Nordic conference on Human-computer interaction* (New York, NY, USA, 2004), NordiCHI '04, ACM, pp. 277 – 280.

[48] JIAN ZHANG, J., CHAN, H. Y., AND FUNG, P. Improving lecture speech summarization using rhetorical information. In *Automatic Speech Recognition Understanding, 2007. ASRU. IEEE Workshop on* (Dec. 2007), pp. 195 – 200.

[49] JOHNSON, B., AND SHNEIDERMAN, B. Tree-maps: A space-filling approach to the visualization of hierarchical information structures. In *Proceedings of the 2nd conference on Visualization '91* (Los Alamitos, CA, USA, 1991), VIS '91, IEEE Computer Society Press, pp. 284–291.

[50] KAGAN, D. CLCV 205: Introduction to ancient greek history., 2007. (Yale University: Open Yale Courses), `http://oyc.yale.edu` (Accessed August 28, 2012). License: Creative Commons BY-NC-SA.

[51] KAMABATHULA, V., AND IYER, S. Automated tagging to enable fine-grained browsing of lecture videos. In *Technology for Education (T4E), 2011 IEEE International Conference on* (July 2011), pp. 96 – 102.

[52] KEIM, D. A., MANSMANN, F., AND THOMAS, J. Visual analytics: How much visualization and how much analytics? *SIGKDD Explor. Newsl. 11*, 2 (May 2010), 5–8.

[53] KIM, J., OARD, D. W., AND SOERGEL, D. Searching large collections of recorded speech: A preliminary study. *Proceedings of the American Society for Information Science and Technology 40*, 1 (2003), 330 – 339.

[54] KIM, W., AND HANSEN, J. H. *Handbook of Research on Digital Libraries: Design, Development, and Impact*. Hershey, PA: Information Science Reference, 2009, ch. Speechfind: Advances in Rich Content Based Spoken Document Retrieval, pp. 173 –187.

[55] KRZYWINSKI, M., SCHEIN, J., BIROL, I., CONNORS, J., GASCOYNE, R., HORSMAN, D., JONES, S. J., AND MARRA, M. A. Circos: An information aesthetic for comparative genomics. *Genome Research 19*, 9 (2009), 1639 – 1645.

[56] LAHTI, T., HELÉN, M., VUORINEN, O., VÄYRYNEN, E., PARTALA, J., PELTOLA, J., AND MÄKELÄ, S.-M. On enabling techniques for personal audio content management. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval* (New York, NY, USA, 2008), MIR '08, ACM, pp. 113 – 120.

[57] LARSON, M., AND JONES, G. J. F. Spoken content retrieval: A survey of techniques and technologies. *Foundations & Trends in Information Retrieval 5*, 4/5 (2011), 237 – 422.

[58] LIN, M., CHAU, M., CAO, J., AND NUNAMAKER JR., J. F. Automated video segmentation for lecture videos: A linguistics-based approach. *International Journal of Technology and Human Interaction (IJTHI) 1*, 2 (2005), 27 – 45.

[59] LUZ, S., AND MASOODIAN, M. A mobile system for non-linear access to time-based data. In *Proceedings of the working conference on Advanced visual interfaces* (New York, NY, USA, 2004), AVI '04, ACM, pp. 454 – 457.

[60] MAIMON, O., AND ROKACH, L. *The Data Mining and Knowledge Discovery Handbook: A Complete Guide for Researchers and Practitioners*. The Kluwer International Series in Engineering and Computer Science. Springer Science+Business Media, 2005.

[61] MALAN, D. J. Computer science 50: Introduction to computer science I, Fall 2011. (Harvard College), `http://cs50.tv/2011/fall/` (Accessed December 11, 2012).

[62] MANNING, C. D., RAGHAVAN, P., AND SCHÜTZE, H. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[63] MASKEY, S., AND HIRSCHBERG, J. Summarizing speech without text using hidden markov models. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers* (Stroudsburg, PA, USA, 2006), NAACL-Short '06, Association for Computational Linguistics, pp. 89 – 92.

[64] MASON, A., EVANS, M. J., AND SHEIKH, A. Music information retrieval in broadcasting: Some visual applications. In *Audio Engineering Society Convention 123* (Oct. 2007).

[65] MATLIN, M. W. *Cognition*, fourth ed. Harcourt Brace College Publishers, Fort Worth, 1998.

[66] MICROSOFT. Microsoft audio video indexing service (MAVIS). `http://research.microsoft.com/en-us/projects/mavis/` (Accessed January 11, 2012).

[67] MOERE, A. V., AND PURCHASE, H. C. On the role of design in information visualization. *Information Visualization 10*, 4 (2011), 356–371.

[68] MUNTEANU, C., BAECKER, R., PENN, G., TOMS, E., AND JAMES, D. 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* (New York, NY, USA, 2006), CHI '06, ACM, pp. 493 – 502.

[69] MUPPALA, J. K., AND KONG, C. K. Podcasting and its use in enhancing course content. In *Proceedings of the 10th IASTED International Conference on Computers and Advanced Technology in Education* (Anaheim, CA, USA, 2007), CATE '07, ACTA Press, pp. 492 – 495.

[70] NASA. *NASA Task Load Index (TLX) v1.0 Manual*, 1986.

[71] OBAMA, B. The 2012 state of the union: An america built to last, Jan. 2012. `http://www.whitehouse.gov/state-of-the-union-2012` (Accessed August 28, 2012).

[72] PALEY, W. TextArc: Showing word frequency and distribution in text. In *Poster presented at IEEE Symposium on Information Visualization* (2002).

[73] PORTER, M. F. An algorithm for suffix stripping. In *Readings in Information Retrieval*, K. Sparck Jones and P. Willett, Eds. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997, pp. 313 – 316.

[74] REPP, S., GROSS, A., AND MEINEL, C. Browsing within lecture videos based on the chain index of speech transcription. *Learning Technologies, IEEE Transactions on 1*, 3 (July – Sept. 2008), 145 – 156.

[75] REPP, S., AND MEINEL, C. Segmenting of recorded lecture videos — the algorithm VoiceSeg. In *SIGMAP* (Aug. 2006), pp. 317 – 322.

[76] RICE, S. V. Frequency-based coloring of the waveform display to facilitate audio editing and retrieval. In *Audio Engineering Society Convention 119* (Oct. 2005).

[77] RIVADENEIRA, A. W., GRUEN, D. M., MULLER, M. J., AND MILLEN, D. R. Getting our head in the clouds: Toward evaluation studies of tagclouds. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 2007), CHI '07, ACM, pp. 995 – 998.

[78] ROŞU, M.-C. Accessing speech documents on smartphones. In *Proceedings of the 5th Annual International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services* (ICST, Brussels, Belgium, Belgium, 2008), Mobiquitous '08, ICST (Institute for Computer

Sciences, Social-Informatics and Telecommunications Engineering), pp. 51:1 – 51:10.

[79] SCHRAMMEL, J., LEITNER, M., AND TSCHELIGI, M. Semantically structured tag clouds: An empirical evaluation of clustered presentation approaches. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2009), CHI '09, ACM, pp. 2037 – 2040.

[80] SEGARAN, T., AND HAMMERBACHER, J. *Beautiful Data: The Stories Behind Elegant Data Solutions*. Theory in practice. O'Reilly Media, Incorporated, 2009.

[81] SHNEIDERMAN, B. Treemaps for space-constrained visualization of hierarchies, 2009. Available from `http://www.cs.umd.edu/hcil/treemap-history/` (Accessed 15 September, 2012). Revised by Catherine Plaisant.

[82] STONE, M. C. *A Field Guide to Digital Color*. Ak Peters Series. Peters, 2003.

[83] TUCKER, R. C. F., HICKEY, M., AND HADDOCK, N. Speech-as-data technologies for personal information devices. *Personal Ubiquitous Comput. 7*, 1 (May 2003), 22 – 29.

[84] VAN THONG, J.-M., MORENO, P., LOGAN, B., FIDLER, B., MAFFEY, K., AND MOORES, M. Speechbot: An experimental speech-based search engine for multimedia content on the web. *Multimedia, IEEE Transactions on 4*, 1 (Mar. 2002), 88 – 96.

[85] VEMURI, S., AND BENDER, W. Next-generation personal memory aids. *BT Technology Journal 22* (2004), 125 – 138. 10.1023/B:BTTJ.0000047591.29175.89.

[86] VIÉGAS, F., WATTENBERG, M., AND FEINBERG, J. Participatory visualization with wordle. *Visualization and Computer Graphics, IEEE Transactions on 15*, 6 (Nov. – Dec. 2009), 1137 – 1144.

[87] VOICEBASE. Voicebase: Store, search & share recordings. `http://www.voicebase.com/` (Accessed January 11, 2012).

[88] WATTENBERG, M., AND VIEGAS, F. The word tree, an interactive visual concordance. *Visualization and Computer Graphics, IEEE Transactions on 14*, 6 (Nov. – Dec. 2008), 1221 – 1228.

[89] WELLNER, P., FLYNN, M., AND GUILLEMOT, M. Browsing recordings of multi-party interactions in ambient intelligent environments. In *Proc. CHI Workshop Lost in Ambient Intelligence* (Vienna, Austria, 2004).

[90] WHITTAKER, S., DAVIS, R., HIRSCHBERG, J., AND MULLER, U. Jotmail: A voicemail interface that enables you to see what was said. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (New York, NY, USA, 2000), CHI '00, ACM, pp. 89 – 96.

[91] WHITTAKER, S., AND HIRSCHBERG, J. Accessing speech data using strategic fixation. *Computer Speech &; Language 21*, 2 (2007), 296 – 324.

[92] WHITTAKER, S., HIRSCHBERG, J., AMENTO, B., STARK, L., BACCHIANI, M., ISENHOUR, P., STEAD, L., ZAMCHICK, G., AND ROSENBERG, A. SCANMail: A voicemail interface that makes speech browsable, readable and searchable. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2002), CHI '02, ACM, pp. 275 – 282.

[93] WHITTAKER, S., HIRSCHBERG, J., CHOI, J., HINDLE, D., PEREIRA, F., AND SINGHAL, A. SCAN: Designing and evaluating user interfaces to support retrieval from speech archives. In *Proceedings of the*

22nd annual international ACM SIGIR conference on Research and development in information retrieval (New York, NY, USA, 1999), SIGIR '99, ACM, pp. 26 – 33.

[94] WHITTAKER, S., HIRSCHBERG, J., AND NAKATANI, C. H. Play it again: A study of the factors underlying speech browsing behavior. In *CHI 98 conference summary on Human factors in computing systems* (New York, NY, USA, 1998), CHI '98, ACM, pp. 247 – 248.

[95] ZAFAR, A., MAMLIN, B., PERKINS, S., BELSITO, A. M., OVERHAGE, J., AND MCDONALD, C. J. A simple error classification system for understanding sources of error in automatic speech recognition and human transcription. *International Journal of Medical Informatics 73*, 9 – 10 (2004), 719 – 730.

[96] ZINMAN, A., AND DONATH, J. Navigating persistent audio. In *CHI '06 extended abstracts on Human factors in computing systems* (New York, NY, USA, 2006), CHI EA '06, ACM, pp. 1607 – 1612.