

Time Is Of The Essence: An Evaluation Of Temporal Compression Algorithms

Simon Tucker and Steve Whittaker

Department of Information Studies

University of Sheffield

{s.tucker,s.whittaker}@shef.ac.uk

ABSTRACT

Although speech is a potentially rich information source, a major barrier to exploiting speech archives is the lack of useful tools for efficiently accessing lengthy speech recordings. This paper develops and evaluates techniques for *temporal compression* - reducing the time people take to listen to a recording while still extracting critical information. We first describe an exploratory study that identifies novel *excision* techniques that remove unimportant words or utterances from the recording. We then develop a new method for evaluating how well temporal compression supports users in forming a general understanding of a recording. Applying this method, we demonstrate that excision techniques are generally more effective than standard compression techniques that simply speed up the entire recording.

Categories & Subject Descriptors

H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems - audio input/output, evaluation/methodology; H.5.2 [Information Interfaces and Presentation]: User Interfaces - evaluation/methodology, prototyping, voice I/O; H.1.2 [Models and Principles]: User/Machine Systems - human factors, human information processing; H.5.5 [Information Interfaces and Presentation]: Sound and Music Computing - methodologies and techniques.

Author Keywords

Temporal Compression, Speech Summary, Summarization, Speech Manipulation, Evaluation Methods, Audio Interfaces.

INTRODUCTION

Speech is a ubiquitous and expressive medium [5]. Furthermore, as the cost of digital storage decreases, large speech archives are becoming available for different speech genres, including meetings [16], news [23], voicemail [26]

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CHI 2006, April 22-28, 2006, Montréal, Québec, Canada.

Copyright 2006 ACM 1-59593-178-3/06/0004...\$5.00.

and conference presentations [15]. Until recently, however, the lack of good end user tools for searching and browsing speech made it tedious to extract information from these archives.

Recent research has begun to develop such end user tools. For example, numerous projects have developed visual interfaces that allow users to browse meeting records using various indices such as speaker, topic, visual scene changes, user notes or slide changes [2,7,16,19,20]. Other research has developed methods that allow users to browse and search transcript-centric representations derived by applying automatic speech recognition (ASR) to the recording [22,26].

One important limitation of these tools is that they make use of feature-rich visual displays to show complex representations of speakers, ASR transcripts, documents, whiteboard, video and slides. But people are increasingly using simpler devices such as phones or PDAs for accessing data, and these have so far received less attention from browser designers. This paper focuses on tools to help users derive an understanding of archival recordings using interfaces that lack complex display characteristics. Specifically, we examine techniques that support *temporal compression* - exploring simple interfaces that reduce the amount of time needed to listen to an entire speech recording. Temporal compression is intended to support the extraction of *gist*, i.e. a general understanding of a recording, rather than providing access to specific facts within the recording.

There are two obvious ways to reduce the amount of time spent listening to a recording. We can either *excise* portions of the recording, or alter the playback rate to *speedup* the recording. In the speedup case, playback rate is typically altered without affecting speaker pitch, and recent implementations attempt to mimic ways that human speakers increase their speech rate [6]. Studies show speedup to be effective. Users can comprehend information played at twice the normal rate and, after exposure to sped up speech, they prefer it to normal speech rate [4,18].

Speedup means that the user hears the entire recording, albeit at a faster rate. The alternative approach, *excision*, saves user processing time by removing unimportant information. Various methods are used to identify

unimportant information, exploiting both acoustic and semantic cues. For example, one acoustically-based excision technique identifies and removes silences or inter-word pauses from the recording [2]. Intonation can also be used to identify important speech segments allowing users to focus on these [2,19]. More recent approaches to excision have used semantic summarization techniques. They apply ASR to the recording to derive transcripts, and then use text processing techniques to identify important parts of the recording, excising unimportant ones [12,14].

One limitation of this prior work is that we lack information about the comparative effectiveness of temporal compression techniques, making it difficult to provide detailed guidelines for browser designers. Previous studies have evaluated different techniques in isolation [2,12,18], but not compared them directly.

One goal of this paper is therefore to explore and evaluate different temporal compression techniques. We are particularly interested in comparing excision with more standard speedup techniques, to explore tradeoffs between hearing a complete recording at high speed versus hearing selected extracts played at normal speed.

We also examine *how* people use these different techniques to process speech, in particular what strategies they use to deal with processing difficulties. Do they uncompress excised or sped-up speech when they fail to understand it, or do they replay the parts they misunderstood? And do these repair strategies differ for different compression techniques and for different amounts of compression?

We first review an initial study comparing 8 different temporal compression techniques (consisting of both excision and speed up techniques). We compare their perceived effectiveness, when users are presented with short continuous clips illustrating the techniques. On the basis of that study we identify two promising excision techniques, insignificant utterance- and insignificant word-excision. We then carry out a follow up study which addresses the limitations of the initial study by developing a new method for comparing different temporal compression techniques objectively. We apply this method to compare excision techniques with a more standard compression technique, speedup, and an uncompressed baseline. We also explore how users exploit different temporal compression techniques and the processing problems they experience with those techniques.

INITIAL QUALITATIVE STUDY

To identify promising temporal compression techniques, we carried out an exploratory subjective user evaluation of eight different temporal compression techniques [21]. These techniques include different types of *speedup* (linear, speech rate normalized, and increasing within a phrase), different types of *excision* (removing either silence, insignificant words or insignificant phrases), as well as *hybrid* techniques where we identified insignificant phrases

or silence but played these highly sped up. The speedup component of the hybrid techniques was intended to give users information about how much unimportant material had been excised. We used short (two minute) compressed clips of semi-scripted meetings presented continuously, i.e. listeners could not pause or replay material. We asked them to subjectively rate the intelligibility and effort involved in understanding the clips.

Contrary to our expectations, users preferred excision techniques to speedup, preferring information removal over faster presentation of complete recordings. They also preferred excision techniques based on semantic, rather than acoustic cues. Hybrid techniques were rated no better than simple speedup or excision. Users also reported problems with speedup, indicating it demanded increased concentration to follow the recording.

The study led us to identify two specific techniques for further analysis. They both used semantic excision: one removed unimportant words; the other removed unimportant utterances. We wanted to know whether these would overcome the reported problems with speedup.

Our study collected rapid subjective information about a large number of compression techniques, and identified problems with existing techniques. It had several limitations, however, which prevent us from drawing concrete conclusions about temporal compression techniques. We now discuss these limitations and how we addressed them in a follow up study.

LIMITATIONS OF THE INITIAL STUDY

The initial qualitative study had three major limitations.

1. We used short excerpts, whereas we expect the main benefits of compression will occur for much longer recordings.
2. Using subjective ratings meant that we were unable to objectively compare the effectiveness of different compression techniques.
3. We used a continuous presentation procedure, where users listened to compressed excerpts without being able to interact with them. Users complained they could not pause, remove compression or replay parts of the recording they couldn't quite understand.

We now discuss how we will address these limitations.

Using Longer Speech Excerpts

Our initial study and the prior literature (e.g. [10]) make use of reasonably short speech excerpts (typically 5 minutes or less) But as the duration of the excerpt increases, the amount of time saved by compression should also increase. We therefore investigated algorithm performance when compressing excerpts that were 30 minutes long.

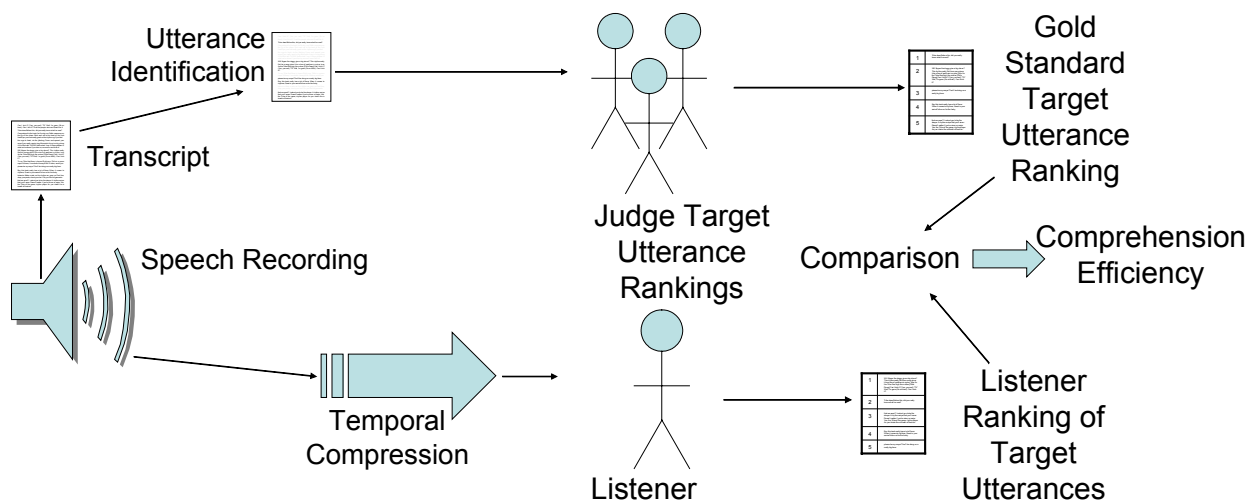


Figure 1. Overview of the assessment procedure. Judges examine the transcript ranking selected target utterances to produce a gold standard ranking. Listeners then hear corresponding audio excerpts which have been temporally compressed and rank the same set of target utterances. This ranking is then compared to the gold standard to produce the comprehension efficiency score.

Objectively Measuring the Effectiveness of Temporal Compression

We also wanted to objectively measure the difference between compression techniques. As stated above, our goal in temporal compression is to allow users to rapidly extract the gist of a recording. Our focus on gist means we cannot employ existing evaluation metrics that measure users' ability to extract specific factual information [25].

Instead we looked into evaluation metrics used in summarization. A common approach for assessing summarization algorithms is to have humans generate a summary (referred to as the 'gold-standard') and then measure the similarity of any automatic summary to this gold-standard.

While generating the gold-standard is relatively costly, once constructed it is a reusable resource that allows the automatic evaluation of different summarization algorithms [13,17]. In our case however, whilst we can produce a similar human generated gold-standard summary to evaluate each temporal compression technique, we would still need listeners to manually produce a summarization of what they had heard. This would make any evaluation prohibitively time-consuming.

Because fact-finding is inappropriate and the summarization approach is too time-consuming, we devised a hybrid approach to capitalise on the advantages of both techniques. This process is illustrated in Figure 1. We produce a gold-standard *ranking* of utterances from the excerpt by selecting representative target utterances (see below) and asking human judges to rank these target utterances in order of their importance.

To evaluate a temporal compression algorithm we compress the same excerpt and present it to human listeners. We then

ask those listeners to rank the importance of the same subset of target utterances previously evaluated by the judges. If the temporal compression algorithm supports effective extraction of gist, we should find that the judges' and users' importance rankings are highly correlated.

User Control and Supporting the Active Browsing Process

Finally, we were interested in how people use compression techniques in a browsing context. Our initial experiment used a continuous presentation procedure, in which users were unable to uncompress, pause or replay excerpts. They found this unacceptable, however, indicating that they wanted to turn the compression off, pause or re-listen to a recent portion of the recording when processing became difficult or they felt they had missed something important.

To investigate active browsing we constructed a simple interface (see Figure 2) - refined after piloting with representative users. This allowed listeners to pause, toggle temporal compression on or off (using the 'comp' button) and also to rewind the recording (using the 'back' button) by a fixed amount of five seconds. The interface also showed users their current position in the recording and allowed them to alter the volume.

The interface was designed for the purpose of evaluating the different compression techniques. It was deliberately restrictive, in order to generate revealing usage behaviour in response to compression, rather than support all possible browsing activities. Thus we allow listeners to reverse a fixed amount, but not skip forward in the recording. This rewind function allows listeners to clarify material they have previously heard, but not override compression by jumping forwards. Furthermore, whilst we felt that users would most likely want to rewind the recording and turn the

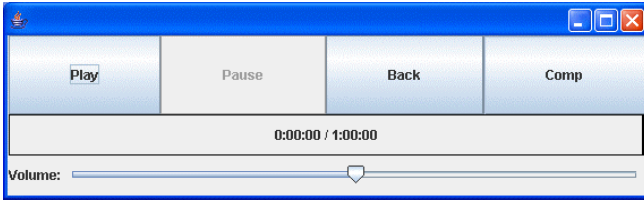


Figure 2. Interface for accessing compressed speech. Users can pause playback, go back five seconds or toggle compression off and on. The ‘Comp’ button toggles to indicate whether compression is on or off.

compression off in one action, we kept these functions separate in the interface to investigate how each function is used by listeners.

By supporting active browsing, it follows that we are not able, as experimenters, to control the exact amount of compression listeners hear, because different listeners will use the browser to listen to the same material in different ways. This makes strictly controlled comparisons of the algorithms problematic in the browsing condition. We therefore had two phases to the experiment. In the first phase, listeners heard relatively short continuous speech excerpts but were not able to control browsing. In the second phase, we used longer clips but gave users the browsing interface described above. This experimental design also has the benefit that, in the second browsing phase, we do not have to explain a particular compression technique to users as they have already experienced it directly.

ASSESSMENT PROCEDURE

The goal of the experiment was to objectively compare the promising excision techniques developed in the initial study (insignificant utterance and insignificant word removal) with speedup – the standard temporal compression technique. We also wanted to know whether these techniques were better than uncompressed playback in allowing users to identify the gist of the recording, and how people used compression techniques when they were allowed active control over their browsing.

Our algorithms derived insignificant utterances and words from human generated transcripts. But in many practical situations these would not be available, and our algorithms would have to rely on errorful ASR-generated transcripts, where exact error rates depend on many factors including recording conditions, ASR algorithm, language and acoustic models. In this study we did not use such errorful transcripts, because we wanted to test the upper bound of excision techniques.

This section provides more details about the procedure we used.

Preliminary Preparations

We chose 36 four-minute and 8 thirty-minute meeting excerpts from the ICSI meeting corpus [16], using manual transcripts supplied with the corpus. From these transcripts we measured the importance of non-stop words (stop words are common, non-informational words like ‘the’, ‘it’, ‘them’ etc.) using simple information retrieval measures of term frequency * inverse document frequency (*tfidf*) [3]:

$$imp_{td} = \frac{\log(count_{td} + 1)}{\log(length_d)} \times \log\left(\frac{N}{N_t}\right)$$

where imp_{td} is the importance of a term t in a document d , $count_{td}$ is the frequency with which term t appears in transcript d , $length_d$ is the number of unique terms in transcript d , N is the number of transcripts in the corpus and N_t is the number of transcripts in the corpus which contain the term t .

We then derived the importance of each utterance in the transcript by computing the mean importance of non-stop words contained in each utterance. For the short excerpts, five target utterances were manually selected and sampled from a range of different importance levels. Thus we chose target utterances that were highly important, unimportant and a number with intermediate levels of importance, ensuring that each was meaningful and non-repetitive. For the longer excerpts, twenty target utterances were chosen using the same criteria. Target utterances were less than a minute long, representing speech from a single speaker, they were a mean of 16 words in length.

Constructing the Gold-Standard

We then built a small web-based application to collect the judges' gold-standard target utterance rankings. Each judge was assigned either twelve short excerpts or four long excerpts to judge. Other research reports that judges experience problems in ranking collections of spoken utterances, but make more accurate judgments when presented with transcripts of the speech [23]. We therefore allowed judges to read transcripts of the recordings and rank the importance of target utterances on-line. They were given one month to complete their rankings. Fifteen judges were used, meaning that we collected three independent rankings for each meeting excerpt.

To ensure that judges were in agreement we measured Kendall's coefficient of concordance [9] for the rankings provided for each meeting. The concordance coefficient in each case was greater than 0.6, with mean concordance 0.75, indicating agreement ($p < 0.05$). We then constructed the gold-standard rankings by computing the mean ranking for each target utterance across judges, assuming that ranking are linear. Note that this means that target utterances can be assigned non-integral ratings

The number of gold-standard statements we asked users to rank depended on the length of the speech excerpt - 5 utterances for the short, and 20 for the longer excerpts. The

judges had an unlimited amount of time to rank utterances; however, the experimental users' time was naturally limited, making ranking problematic for longer excerpts. Thus for longer recordings we asked listeners to rate the importance of each target utterance in isolation rather than trying to compare and rank each target utterance. Specific details of how the procedure was implemented are described below.

Compression Techniques

We used three different algorithms to produce the compressed excerpts: insignificant utterance excision, insignificant word excision and a heuristically motivated speedup algorithm. These were compared with a non-compressed baseline.

Insignificant Utterance Excision

There are many possible ways to implement utterance excision [2,12,14]. We used a simple method that did not require complex natural language or acoustic processing. We first computed utterance importance scores using the *fidf* measure described above. We then used these scores to rank utterances according to their importance. To ensure that the compressed recording contained all the target utterances selected for this excerpt, we began with an audio file that included each of our target utterances. This meant that users were not biased by being asked to rank utterances that they did not hear. We then progressively included high-ranking utterances until the file was of the duration determined by the compression amount. Utterances were presented in the order in which they occurred in the original recording.

Insignificant Word Excision

This method is identical to utterance excision, except that importance scores are calculated for each word and high ranking words added to the file containing target utterances until the desired level of compression is achieved. We removed stop words from target utterances prior to producing the compressed excerpt, so that target utterances would not appear unusual, when users heard them. The following example shows the effects of excision:

Initial: *'yeah, i have to tell you for the uh - for the admin meeting that we have, lilah does that um every time before an admin meeting'*.

Compressed: *'tell admin meeting lilah time admin meeting'*.

Non-Linear Speech Rate Alteration

We used the mach1 speedup algorithm [6] which aims to replicate the phonetic speed variations which occur when humans naturally modify their speech rate. We first compute a measure of the relative speech rate for each part of the recording. We then linearly transform this relative speech rate contour to reach the desired level of compression. This transformed contour is then used to dynamically alter the speech rate using a standard SOLA algorithm [11].

Control: Non-Compressed Excerpts

Uncompressed control excerpts allow us to directly compare performance between compressed and uncompressed excerpts.

Compression Levels

In addition to modifying the *type* of compression we also modified the *level* of compression to see whether a technique's effectiveness depends on the amount of compression. This allows us to tell, for example, whether speedup works at low compression levels but is ineffective at higher ones. The levels of compression were chosen to reflect levels beginning at those cited as being comfortable for listeners [10]. For the short excerpts we applied three levels of compression (66, 50 and 40% of the original duration, corresponding to 1.5, 2 and 2.5 times normal speed) and for the longer excerpts two levels of compression were applied (66 and 50% of the original length).

Users

The users were university staff and students. They were aged between 20 and 40. None reported any hearing difficulties and each was paid for their participation.

Experimental Procedure

All experiments took place in a noise-reduced acoustic booth, excerpts were presented diotically (monaural to both ears) over Sennheiser HD250 Linear II headphones. We used a Java program to present and collect the results.

Users attended experiments over four days, and were presented with a single compression technique each day. Each experiment contained two phases, the first phase involved short continuous excerpts and the second phase longer excerpts where people used the browser to control what they heard.

In the first phase, users heard 9 different compressed excerpts (3 repetitions of each of the 3 compression rates), and 3 uncompressed excerpts. After hearing each complete excerpt, they were presented with the set of target utterances from the excerpt, and asked to rank the importance of each target utterance. Users performed their ranking by choosing labels ('important' to 'unimportant') from drop-down menus next to the target utterances. The target utterances were presented in a random ordering to each user.

Before carrying out the second phase, we gave users a short web-based tutorial that explained the functioning of the browsing interface. (see Figure 2). The tutorial explained each interface feature (e.g. pause, backup, uncompress, replay), encouraging listeners to practice using that feature. They were allowed as much time as they liked to familiarize themselves with the interface before proceeding to the next part of the task. Interface practice was provided

each day as different compression techniques may demand different usage strategies.

In the second phase users had thirty minutes to listen to each excerpt using the browser. We imposed a time limit as we were interested in *efficient* processing of speech. We chose 30 minutes as the enforced time limit as this was the uncompressed duration of the excerpt. The experiment finished either after thirty minutes elapsed, or the user decided they had fully processed the excerpt, whichever happened sooner (typically users finished before the thirty minute deadline, except in the uncompressed condition). To keep users aware of their progress, the interface included a timer showing how much time remained for them to complete the task. It was made clear that users needed to leave themselves enough time to listen to the entire recording. They were then presented with the twenty target utterances and the same rankings as before, although as discussed above users were just asked to judge the importance in this case. Users had unlimited time to perform their rankings or judgements in both phases.

Performance Measures and Data Collected

We used Kendall's tau to measure the agreement between the gold standard rankings and the user rankings or judgements. The performance score was thus computed using the following equation:

$$\tau = 1 - \left(\frac{2i}{p}\right)$$

where i is the number of inversions between ranking pairs and p is the total number of ranking pairs. Thus we compute the proportion of utterances pairs which users have ranked in a different order from the gold-standard. By computing Kendall's tau in this way we overcome any problems associated with non-integral rankings (since we are only interested in the *direction* of pairwise orderings). In the case where we are comparing importance *judgements* to gold-standard rankings, if two sentences are judged to be of equal importance we only count this as a misordering if the corresponding gold-standard ranking differs by four or more (since there were five judgement levels for twenty target utterances). Thus the same scoring technique can be used for both lengths of meeting excerpt.

This simple agreement measure does not, however, capture a key aim of temporal compression - which is to reduce the *time* taken to effectively process a recording. We therefore normalized the success score to allow for the length of time users took in listening to the recording. In this way we control for the fact that uncompressed speech is easier to process, but takes longer to listen to. For short excerpts users have no control over what they hear, so this listening time is the excerpt length. In the case of the longer excerpts we measure the amount of time each listener took to process the recording. We call this normalized measure *comprehension efficiency* (C_e),

$$C_e = \frac{\tau}{t}$$

where t is the total amount of time the user spends listening to speech.

We also collected subjective data at the end of each phase of the experiment, examining both users' perception of the techniques and, in the second phase, their perceived use of the interface. The phrases used for Likert responses were:

- "I found it easy to get the gist of the discussion"
- "I felt I was missing important information"
- "I felt the speech was too fast"

In the second phase we also used the phrases:

- "I repeatedly had to go back in the speech"
- "If I missed something I could easily find out what it was"

In the second phase we also logged all user interface actions, their timing and duration. This enabled us to determine how frequently users paused, rewound and replayed each excerpt, as well as whether they were listening with or without compression.

HYPOTHESES

For both short (S) and long (L) excerpts, we made predictions about compression level, type and subjective reactions to different compression types, and about how people would actively use the browser to control different types of compression.

Efficiency Hypotheses

S1/L1: Compression Level

Greater compression should allow users to identify gist more efficiently. We therefore expected greater comprehension efficiency at higher compression levels for all compression techniques.

S2/L2: Compression Type

Our previous study led us to expect differences between compression techniques. Specifically we predicted that excision techniques would be more efficient than both speedup and uncompressed speech.

Subjective Hypotheses

S3/L3: Subjective Responses

We expected users' subjective responses to replicate the findings of our initial study, showing a preference for excision over speed up.

L4: Interface Responses

Our previous work indicated that speedup can present information too fast. We therefore expected users to report increased use of the rewind functions in the speedup case.

Interface Interaction Hypotheses

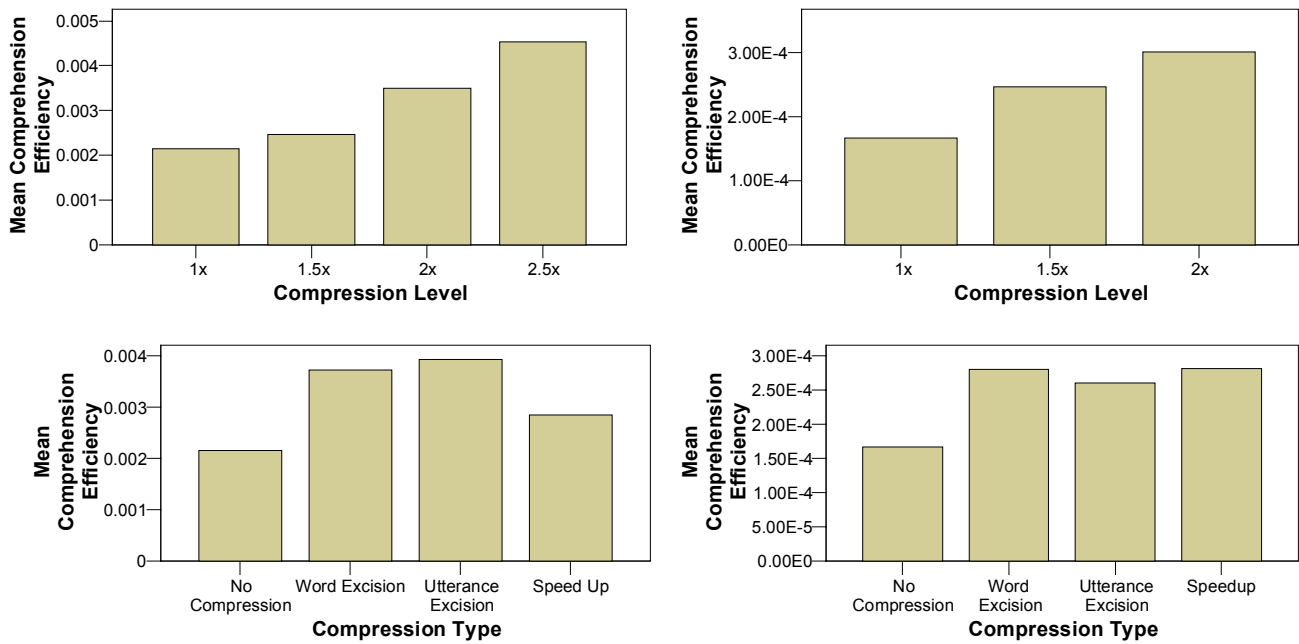


Figure 3. Bar graphs showing comprehension efficiency against compression type for the short (left panel) and long (right panel) excerpts.

We also predicted how people would make use of the browser in the different compression conditions.

P1: Reactions in the Speedup Condition

A compression technique such as speed up induces cognitive overload. We therefore expected that speedup would lead users to pause or switch off compression to reduce this.

P2: Reactions in the Utterance Excision Condition

In contrast, utterance excision omits material, which introduces discontinuities in what users hear. They may understand the utterance they have just heard but have no context for it. We therefore expected users to be more likely with utterance excision to rewind and replay an utterance, to establish its prior context.

P3: Reactions in the Word Excision Condition

We expected word excision to occupy the middle ground between the two other compression techniques. Removing a large number of words may both reduce comprehensibility of the current utterance as well as leading to a loss of context for subsequent utterances. Because of this, we

predicted that users would be more likely to both turn the compression off and rewind the recording in this condition.

RESULTS

Our predictions and results are summarized in Table 1. We now discuss specific hypotheses.

Efficiency Hypotheses

To assess the efficiency hypotheses we conducted a 4 (compression level) X 4 (compression type) ANOVA for the short excerpts and a 3 (compression level) X 4 (compression type) ANOVA for the long excerpts. In both cases the dependent variable was comprehension efficiency.

S1/L1: Compression Level

As expected comprehension efficiency is increased at higher compression levels. Figure 3 (top left and right) shows the relationship between comprehension efficiency and compression level for both long and short excerpts. ANOVAs confirmed the effect of compression level on comprehension efficiency ($F=4.43, p<0.01$; $F=16.00, p<0.05$, for the short and long excerpts respectively). There is a significant correlation between comprehension efficiency and compression level ($r=0.21, p<0.001$ (short); $r=0.47, p<0.001$ (long)). This shows that users are able to extract gist even at the highest levels of compression.

S2/L2: Compression Type

The results confirm the superiority of excision techniques over speedup for short excerpts. There were no differences for long excerpts. The mean comprehension efficiency organized by compression type is shown in Figure 3 (bottom left and right). The ANOVAs indicate that an effect of compression type is only present for short excerpts

Hypothesis	Confirmed	Hypothesis	Confirmed
S1	yes	L1	yes
S2	yes	L2	no
S3	yes	L3	yes
P1	yes	P2	yes
P3	no		

Table 1: Results of Hypotheses.

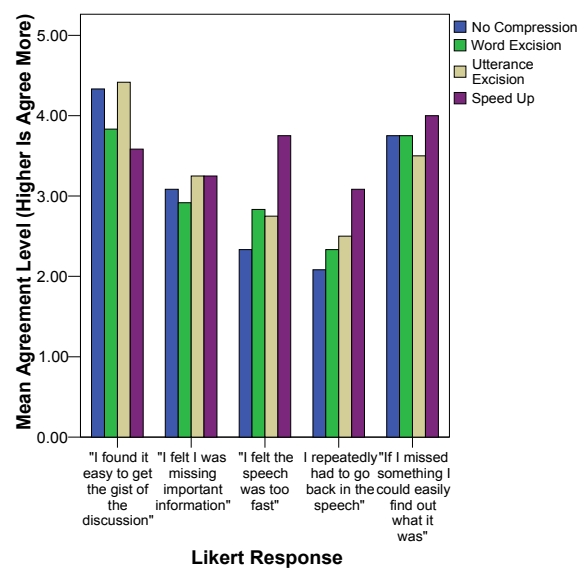
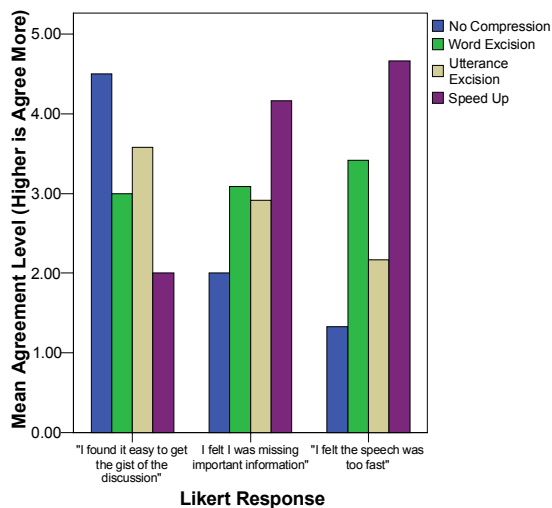


Figure 4. Graph of subjective results (in each case higher is agree more with the target statements) organised by compression type for the short (left panel) and long (right panel) excerpts.

($F=3.84$, $p<0.05$, for short, $F=0.54$, $p > 0.6$, for long). Tukey planned comparisons here indicate, as we predicted, that the excision techniques outperform speedup for short excerpts ($p<0.05$).

There were also no interactions between level and type of compression indicating that the relative success of a technique did not depend on the amount of compression.

Subjective Hypotheses

Excision was perceived as generally better than speedup for short excerpts, with speedup perceived as ‘too fast’ for long excerpts. We conducted a second set of ANOVAs, with the appropriate subjective ratings as dependent variables (see Figure 4).

S3/L3: Subjective Responses

For short excerpts (Figure 4, left hand graph), the ANOVA indicated an effect of compression type on each of the 3 subjective responses (all $p<0.05$). Users found it hardest to ‘extract gist’ with speedup, although this difference was only a trend when comparing word excision and speedup. A similar set of responses was seen for the ‘missing information’ question. For “*I felt the speech was too fast*”, users stated that the sped up speech was too fast, compared with both excision conditions.

For the long excerpts, (Figure 4, right hand graph), all techniques were judged the same for all questions except for the “*I felt the speech was too fast*”. Here again sped up speech was judged as too fast compared with excision conditions and uncompressed speech (all $p<0.05$).

L4: Interface Responses

For the long excerpts, (as shown in Figure 4, right hand graph, two far right columns), we also examined users’ impressions of how they used the browser to address

processing problems. While users felt they could recover missing information equally well in each compression condition, the ANOVA confirmed an effect of compression type on the response for “*I repeatedly had to go back in the speech*”. Planned comparisons indicated that backing up was felt to occur most in the speedup condition ($p<0.05$).

Interface Interaction Hypotheses

To analyse how the browser was used in the different compression conditions, we clustered logfile entries into three categories according to problems users had described in our initial study. These categories were uncompress (in response to information generally being presented too fast), rewind (in response to missing some information), as well as combined rewind-uncompress actions (in response to being unable to understand what one has just heard). The mean number of these actions organized by the compression algorithm is shown in Figure 5.

We found different usage patterns for the different compression techniques: with uncompress being most used with speedup and rewind with utterance excision.

As we predicted (P1), use of the uncompress button depends on compression type ($F=4.76$, $p<0.01$), with planned comparisons showing this is more frequent in the speedup case than other conditions (all $p<0.05$).

Consistent with prediction P2, use of the rewind (or back) button differs for different compression types ($F=3.46$, $p<0.05$). Planned comparisons showed that the back button is most used in the utterance excision condition compared with other conditions (all $p<0.05$).

Contrary to prediction P3, combinations of back-uncompress actions occur independently the compression techniques ($F=0.04$, $p > 0.9$).

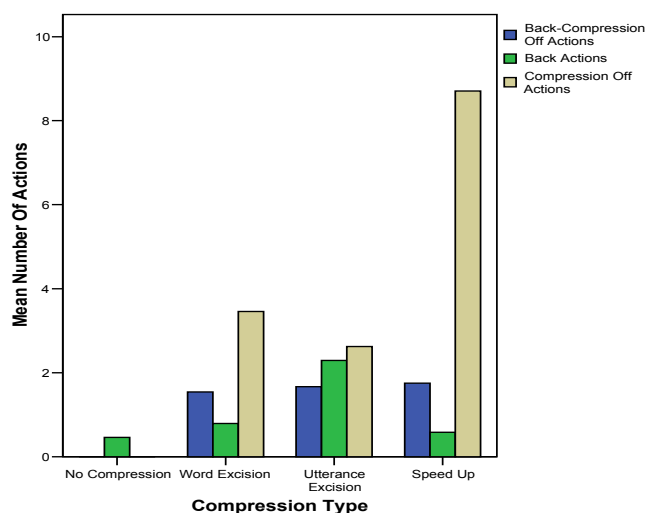


Figure 5. Bar graph of interface actions organised by compression technique. In the no compression case listeners only have access to the ‘back’ function.

Informal Comments

Users’ posthoc informal comments mirrored the experimental data. People were usually positive about utterance excision, (“*This was all very clear*”), although they noted that the discontinuities it introduced could mean that they lacked context for what they were just hearing. (“*[I] found the flow was disrupted when I was listening to a sentence...and it then skips onto another one*”). This comment also indicates that users were sensitive to exactly how the different techniques were presenting information. They were slightly less positive about word excision, feeling it presented information as a ‘stream of words’ and that this could make it hard to understand, even increasing the perceived speed at which information was presented. (“*The stream of words came so quickly at times that understanding became difficult*”). People were much less positive about speedup, however, saying: “*The speech was too fast*” and “*Important information...was simply too garbled*”.

Several users commented that inconsistencies in the recordings (e.g. volume changes, sudden sounds etc.) were amplified in the compressed recordings. They also reported difficulty with overlapping speakers. As with [21,22] users reported their understanding was also dependent on speaker nationality.

SUMMARY AND DISCUSSION

Building on an exploratory study, we developed a method for comparing various temporal compression techniques intended to help users extract the main points of a recording, without having to rely on feature rich visual displays.

Our main study applied this method to compare existing speedup techniques with novel excision techniques that remove insignificant information. We found that excision

techniques were generally better than speedup, and they also outperformed an uncompressed baseline. Users preferred excision techniques to speedup, and were less likely with excision to turn off compression. We did not find predicted differences between excision and speedup for the long excerpts, but potential differences may be masked by active user browsing, e.g. turning off speedup for long excerpts. Finally although utterance excision was slightly preferred, lexical excision was surprisingly successful given the simplicity of the technique.

We also analysed users’ behaviours to infer how the different techniques affect comprehension. Although excision techniques were generally successful, excision and speed up affect comprehension in different ways. Speedup can induce general overload, leading users to uncompress incoming material without replaying what they have heard, whereas excision can lead to context-loss, with the need to replay recently heard materials.

We also found successful comprehension at higher compression levels than prior studies. For example, prior studies of speedup show that users are effective with compression rates of up to 2 times, whereas our users were successful at 2.5 times compression. One possible explanation for this discrepancy is that excision presents users with fewer processing problems than speedup. This is supported by our usage data. Another possibility is that we measured users’ ability to extract gist, whereas other studies measured fact finding or sustainable speed [10] and these are inherently harder tasks.

Despite these demonstrable advantages for excision, our implementation currently exploits a human generated transcript, whereas in practice errorful ASR transcripts will often have to be used. Our results therefore show the upper bound to excision techniques. But ASR errors may not worsen performance significantly. Other research has shown that speech search can be highly effective even with errorful transcripts ([22],[26],[27]), partly because of the nature and type of ASR errors [8]. Furthermore, there may be ways to avoid using transcripts by exploiting other ways to identify important information, e.g. prosody [2]. We intend to conduct more research to examine the effects of ASR errors on gisting ability, and to explore the viability of these other techniques.

Our results also suggest further research questions. Firstly, is excision useful for tasks other than extracting the gist of a recording? For example, could excision techniques, implemented in a working browser, also help users to answer detailed factual questions about a recording? Could people use excision to navigate to relevant regions of a recording and explore these in more depth using standard play operations to answer factual questions? Furthermore, here we used a simple implementation of utterance excision, but how might more complex summarization techniques exploiting text processing [12,14] or acoustic analysis [2] fare in comparison?

Our results have specific implications for speech browser design. They suggest different browser designs for different compression techniques – as users are more reliant on backing up with excision and uncompressing with speedup. They also suggest utterance excision needs a combined one-button function that rewinds and replays prior material uncompressed, whereas speedup demands a button for toggling off compression.

There may be other applications of temporal compression outside the domain of meetings, e.g. news, voicemail or recorded presentations. Another possible application is to teleconferencing, where compression might allow latecomers to rapidly catch up on what has already been discussed. These domains might also allow us to develop variants of the excision algorithm that exploit information from other media sources to identify points of importance. For example, vision processing might identify important events in a video and this information could be used in feature-rich multimedia browsers.

In conclusion, our results identify a new set of general methods and techniques that help to browse and access speech, without requiring sophisticated interfaces. By applying these techniques we should be able to support more effective user access to the rapidly growing number of speech archives.

REFERENCES

1. AMI Project. <http://www.amiproject.org/>.
2. Arons, B. SpeechSkimmer: A system for interactively skimming recorded speech. *ACM Trans. Computer-Human Interaction* 4, 1 (1997), 3-38.
3. Baeza-Yates, R. and Ribeiro-Neto, B. *Modern Information Retrieval*. Addison Wesley, 1999.
4. Beasley, D.S. and Maki, J.E. Time and frequency altered speech. In *Contemporary Issues in Experimental Phonetics*, Academic Press, (1976), 419-458.
5. Chalfonte, B.L., Fish, R.S. and Kraut, R. Expressive richness: A comparison of speech and text as Media for Revision. *Proc. CHI 1991*, (1991), 21-26.
6. Covell, M., Withgott, M. and Slaney, M. Mach1: Nonuniform time-scale modification of speech. *Proc. IEEE ICASSP 1998*, (1998), 493-496.
7. Cutler, R., Rui, Y., Gupta, A. Cadiz, J.J. Tashev, I., He, L., Colburn, A., Zhang, Z., Liu, Z. and Silverberg, S. Distributed meetings: A meeting capture and broadcasting system. *Proc. 10th ACM International Conf on Multimedia*, (2002), 503-512.
8. Garofolo, J., Auzanne, C.G.P. and Voorhees, E.M. The TREC-9 spoken document retrieval track: A success story. *Proc. RIAO-2000*, (2000).
9. Hays, W.L. *Statistics for the Social Sciences*. Holt, Rinehart and Winston, 1973.
10. He, L. and Gupta, A. User benefits of non-linear time compression. *Microsoft Research Technical Report MSR-TR-2000-96*, Microsoft, (2000).
11. Hejna, D. *Real-time time-scale modification of speech via the synchronized overlap-add algorithm*. MSc Dissertation, M.I.T., (1990).
12. Hori, C. and Furui, S. A new approach to automatic speech summarization. *IEEE Trans. Multimedia* 5, 3 (2003), 368-378.
13. Lin, C-W. ROUGE: A package for automatic evaluation of summaries. *Proceedings of ACL 2004*, (2004), 56-60.
14. McKeown, K., Hirschberg, J., Galley, M. and Maskey, S.. From text to speech summarization. In *Proc. of ICASSP 2005*, (2005).
15. MLMI 2005. <http://groups.inf.ed.ac.uk/mlmi05/techprog.html>.
16. Morgan, N., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Janin, A., Pfau, T., Shriberg, E. and Stolcke, A. The meeting project at ICSI. *Proc. HLT Conference*, (2001), 246-252.
17. Nenkova, A. and Passonneau, R. Evaluating content selection in summarization: the pyramid model. In *Proc HLT-NAACL 2004*, (2004), 145-152.
18. Sticht, T.G. Comprehension of repeated time-compression recordings. *Journal of Experimental Education* 37, 4 (1969).
19. Stifelman, L. Augmenting real-world objects: A paper-based audio notebook. In *Proc. CHI 1996*, (1996), 199-200.
20. Tucker, S. and Whittaker, S. Accessing multimodal meeting data: systems, problems and possibilities. In *Lecture Notes in Computer Science 3361*, (2005), 1-11.
21. Tucker, S. and Whittaker, S. Novel techniques for time-compressing speech: An exploratory study. In *Proc of ICASSP 2005*, (2005).
22. Vemuri, S., DeCamp, P., Bender, W. and Schmandt, C. Improving speech playback using time-compression and speech recognition. In *Proc. CHI 2004*, (2004), 295-302.
23. Voorhees, E.M. and Buckland, L.P. *The Thirteenth Text REtrieval Conference Proceedings*. NIST Special Publication, (2004).
24. Walker, M., Prasad, R. and Stent, A. A trainable generator for recommendations in multimodal dialog. In *EUROSPEECH: European Conference on Speech Processing*, (2003), 1697-1701.
25. Wellner, P., Flynn, M., Tucker, S. and Whittaker, S. A meeting browser evaluation test. In *Proc. CHI 2005*, (2005).
26. 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 *Proc. CHI 2002*, (2002), 275-282.
27. Whittaker, S., and Amento, B. Semantic speech editing. In *Proc. CHI 2004*, (2004), 527-534.