

Time-Compressing Speech: ASR Transcripts are an Effective Way to Support Gist Extraction

Simon Tucker, Nicos Kyprianou and Steve Whittaker

Department of Information Studies, University of Sheffield, Sheffield, UK
{s.tucker,s.whittaker}@shef.ac.uk

Abstract. A major problem for users exploiting speech archives is the laborious nature of speech access. Prior work has developed methods that allow users to efficiently identify and access the gist of an archive using textual transcripts of the conversational recording. Text processing techniques are applied to these transcripts to identify unimportant parts of the recording and to excise these, reducing the time taken to identify the main points of the recording. However our prior work has relied on human-generated as opposed to automatically generated transcripts. Our study compares excision methods applied to human-generated and automatically generated transcripts with state of the art word error rates (38%). We show that both excision techniques provide equivalent support for gist extraction. Furthermore, both techniques perform better than the standard speedup techniques used in current applications. This suggests that excision is a viable technique for gist extraction in many practical situations.

1 Introduction

Large archives of speech recordings are becoming more common, as the cost of storage continues to decrease. Such speech archives include: meeting records [9], news [3] and voicemail [18]. Recent tools for accessing these archives either assist users in locating recordings of interest [3] or on supporting the listening process by providing complex visual browsers for access [16]. However people are increasingly using simple, mobile devices (phones or PDAs), where rich displays are not available. In previous work we have developed and evaluated a number of *temporal compression* techniques which do not rely on complex visual displays. They aim to reduce the time taken to listen to a speech recording while still allowing users to extract the most important information. Such techniques are designed to support *gist extraction*, i.e. general understanding of a recording rather than providing access to specific facts.

There are two main ways to reduce the amount of time required to listen to a recording. We can either excise unimportant portions to reduce its length, or alter the playback rate, i.e. speed it up. *Speedup* is used in many commercial applications, e.g. voicemail. With speedup, playback rate is altered so as to not

affect speaker pitch, and the speedup can be non-linear, reflecting the way in which humans naturally increase their speech rate [2, 6]. Whilst speed up ensures that listeners hear the complete recording, *excision* compresses it by removing parts - using semantic or acoustic cues to identify unimportant information for subsequent removal.

In previous experiments [15, 14] we evaluated various novel temporal compression techniques, including speedup, excision and hybrid methods that combined speedup and excision techniques. In [14] we used a measure of gist extraction to compare excision and non-linear speed up against an uncompressed control. Our findings show that excision leads to objectively better listener performance than speedup, and excision is also preferred to speedup. Both excision and speedup compression methods lead to more efficient gist extraction than uncompressed speech. Whilst there is evidence that using ASR as opposed to a manual transcript has only a small effect on textual summarization [12] an outstanding question is what effect transcript accuracy has on the ability of listeners to *extract useful information* from temporally compressed audio. This paper examines this question and determine whether listeners are able to extract gist from temporally compressed recordings when the underlying transcripts contain ASR errors.

We therefore compare excision performance for human generated transcripts, with ASR transcripts containing a word error rate (WER) of 38%. This error rate represents state of the art recognition quality for meetings corpora [4]. We compare these with standard speedup techniques for different levels of temporal compression. We also investigate the effects of compression under two sets of conditions: passive exposure to speech excerpts where users listen to speech clips without being able to stop or replay what they hear; and more active exploration of clips using a simple browsing interface. We first describe the evaluation procedure, following which we outline and discuss the results of the study.

2 Assessment Procedure

The assessment procedure aims to objectively and efficiently measure users' ability extract gist from spoken materials. Typical measures of understanding used in this domain focus on recall of specific facts [17]. However our techniques are intended to support *gist extraction* rather than factual knowledge making these techniques inappropriate. An alternative way to measure gist extraction asks users to summarize what they have heard, which is then scored against a gold standard summary [1]. However summarization and evaluation are time-consuming for subjects. More importantly, there is no consensus about metrics or methods for the effective evaluation of summaries [13]. We assume that effective gist extraction requires users to distinguish the importance of different utterances, being able to say which utterances are central to the meeting and which are peripheral. We therefore devised a hybrid evaluation method which employs judges to produce an initial gold standard importance ranking of representative utterances from the recording. In the evaluation stage, we ask experimental sub-

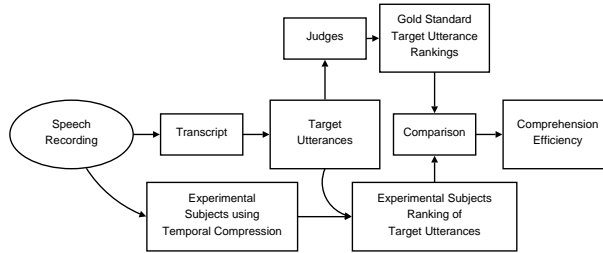


Fig. 1. Overview of assessment procedure. Judges examine the transcript ranking selected target utterances

jects to rank the same utterances after they have listened to different types of temporally compressed recordings. The objective measure of gist extraction is then the correlation between gold standard rankings and those obtained under temporal compression. Our measure of the quality of temporal compression is the extent to which subjects listening using temporal compression are able to replicate the judges gold standard rankings. A high correlation between subject rankings and the gold standard indicates the temporal compression technique provides good support for gist extraction. The evaluation process is described in detail below.

2.1 Algorithms

We used 33 transcripts altogether; 27 four-minute and 6 thirty-minute meeting excerpts from the ICSI meeting corpus [10], using the human generated transcripts supplied with the corpus. The automatically produced (ASR) transcripts were generated using a six pass architecture using 16 component Gaussian HMMs as acoustic models[4]; the WER for the ASR transcripts was 38%. This error rate is state of the art for meetings data.

From both of these transcript sets we first identified stop words (such as ‘the’, ‘and’, ‘them’, etc.). We then separately determined the importance of non-stop words using measures of term frequency * inverse document frequency (TF*IDF)[7]:

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

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 .

2.2 Gold Standard Measurements

We then computed the importance of each *utterance* in the transcripts as the mean importance of the non-stop words which appear in the utterance. These

importance scores were then used to select a subset of utterances which we presented to the judges for gold standard ranking. We did this for long and short extracts. The short excerpts five *target utterances* were manually chosen from the full range of importance levels. Thus both highly important and unimportant utterances were chosen, as well as utterances distributed across the intermediate levels of importance. Utterances were manually selected to ensure that the selected utterances were meaningful and non-repetitive. For the longer excerpts twenty utterances were chosen using the same criteria. Target utterances were chosen to be less than a minute long and represented speech from a single speaker and were a mean of 16 words in length.

We built a small web-based application to collect judges' target utterance rankings. Each judge was assigned a selection of either short or long excerpts and at least three independent rankings were collected for each meeting excerpt.

The judges ranked the five selected utterances for the short excerpts and twenty utterances for the long excerpts. They were given an unlimited amount of time to perform their rankings. To determine reliability we measured Kendall's coefficient of concordance for these rankings. The coefficient in all cases was greater than 0.6 with a mean concordance of 0.75, indicating a good level of agreement ($p < 0.05$). We then constructed the gold-standard rankings by computing the mean ranking for each target utterance across judges, with the assumption that the rankings can be evenly spread on a linear scale. Note that this means that target utterances can be assigned non-integral rankings.

2.3 Compression Techniques

We evaluated two different temporal compression algorithms, one that used excision and the other that used a standard speedup technique [2]. The excision technique removed unimportant utterances, and was applied to both ASR and human generated transcripts.

Insignificant Utterance Excision Our excision approach relies purely on the words in the transcript and does not require complex natural language or acoustic processing. We first compute utterance importance scores using TF*IDF ([7]) using the method described above, to rank the utterances contained within the transcript in order of their importance. The compressed clip is constructed by adding utterances to an empty file in order of their importance until the file reaches the length required by the specified compression level. Utterances are presented in the order in which they occurred in the original recording. We apply the approach to both ASR and human generated transcripts to generate two insignificant utterance excision files - one generated from ASR transcripts and the other from the human generated transcripts.

Non-Linear Speech Rate Alteration We used the Mach1 speedup algorithm [2] which aims to replicate the phonetic variations which occur when humans naturally modify their speech rate. We first compute a measure of the relative

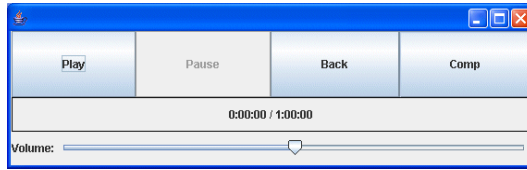


Fig. 2. The interface used in the active condition.

speed rate for each part of the recording. We then linearly transform this relative speed contour so that the entire excerpt duration meets the desired level of compression. This transformed contour is then used to dynamically alter the speed rate using a standard SOLA algorithm [6].

2.4 Compression Levels

In addition to modifying the *type* of compression we also alter the compression level. We also investigate the effects of compression under two sets of conditions: *passive exposure* to short speech excerpts where users cannot stop or replay what they are hearing; and more *active exploration* of longer clips using a simple browsing interface. This simple interface shown in Fig.2 allows users to turn the compression on/off, pause or re-listen to a recent portion of the recording when processing became difficult or they feel they have missed something important.

For the short excerpts we applied three levels of compression (66, 50 and 40% of the original duration, which corresponds 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). These levels of compression are consistent with accepted comfortable listening levels[5].

2.5 Subjects

Eleven subjects were selected from university staff and students. They were aged between 20 and 40. None reported any hearing difficulties and each received a confectionery reward for taking part.

2.6 Experimental Procedure

All experiments took place in a noise-reduced acoustic booth; excerpts were presented diotically over Sennheiser HD250 Linear II headphones. A Java program was used to present the excerpts and to collect the results.

Subjects attended experiments over three days and were presented with a single compression technique for each day. Each day consisted of two phases, a passive phase and an active phase.

In the first (passive) phase, users heard nine different compressed excerpts (three repetitions of each of the three compression levels). After hearing each

complete excerpt they were presented with the set of target utterances for that excerpt and asked to rank the importance of each of the utterances in the context of the whole excerpt. Subjects performed their rankings by choosing labels from a five point Likert scale ('important' to 'unimportant') from drop down menus next to the target utterances. The ordering of the target utterances was randomized for each user.

Before carrying out the active exposure using the simple browser, subjects were given a short tutorial that explained the various functions of the browser interface. They then briefly experimented with the browser on a short speech excerpt until they felt they were comfortable using it. In the active phase subjects had thirty minutes to explore each excerpt using the browser. The experiment was time limited and either ended after thirty minutes, or when the subjects had listened to the excerpt in its entirety. Typically, subjects finished before the thirty minute deadline. The interface indicated how much time was remaining and it was made clear to subjects that they should attempt to use the interface controls for replaying or decompressing sparingly to ensure that they have enough time to listen to the full recording. When they had finished listening they were presented with the twenty target utterances and the same five ranking levels were used to judge (rather than rank) the importance level of each of the target utterances. In both phases subjects were given an unlimited amount of time to perform their rankings or judgments.

2.7 Performance Measures and Data Collected

We used Kendall's tau to measure the level of agreement between the gold standard rankings and the subject ranking and judgments. The performance score was computed using the following equation:

$$\tau = 1 - 2i/p, \tag{2}$$

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 target utterance pairs which users have ranked in a different order from the ordering present in the gold standard. By computing Kendall's tau in this way we overcome any problems associated with the non-integral rankings (since we focus on the direction of the pairwise orderings). The same scoring technique is used for both the long and short meeting excerpts. We additionally normalize the scores by the mean τ across subjects and conditions for clarity.

This simple measure of agreement, however, does not capture a key aim of temporal compression - which is to reduce the time taken to effectively process a recording. We therefore normalize the success scores to take account of listening time as a function of the original recording length. We call this normalized measure Comprehension Efficiency (C_e):

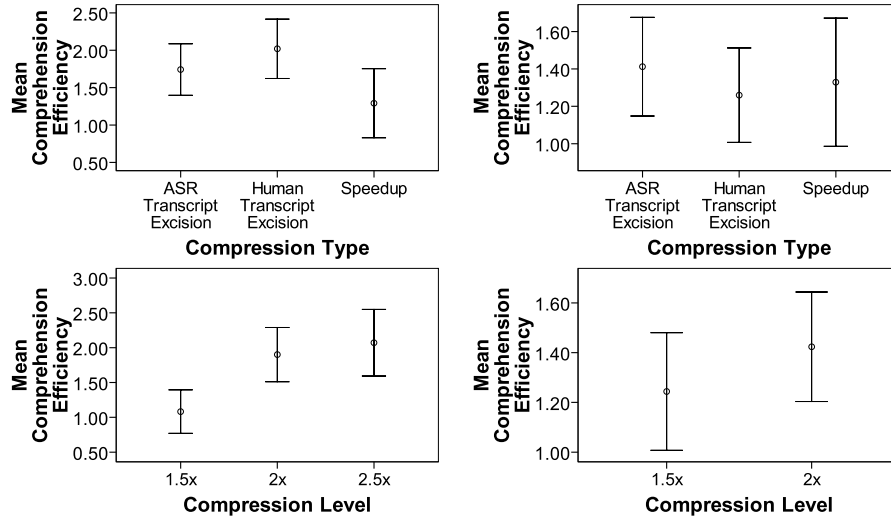


Fig. 3. Error bar graphs showing comprehension efficiency against compression type (bottom) and compression level (top) for the short (left) and long (right) conditions

$$C_e = \frac{\tau/\hat{\tau}}{t_{listening}/t_{original}}, \quad (3)$$

where, $\hat{\tau}$ is the mean tau across all subjects and conditions, $t_{listening}$ is the total listening time and $t_{original}$ is the uncompressed length of the recording.

In addition to Comprehension Efficiency measure, we collected users' subjective reactions to the various compression techniques, asking them to compare and contrast their listening experiences. We also logged their use of the browser, specifically different browser actions such as stopping playback, replaying speech or removing compression to determine how this was used in the different compression conditions.

2.8 Hypotheses

We made separate predictions for the passive and active listening conditions.

Passive Condition We did not expect ASR to affect comprehension efficiency. We expected subjects to be able to extract gist equally efficiently with both the ASR and human transcript based excision methods. As in [14], we expect excision to be superior to speedup, regardless of the type of transcript. We also expect that increased compression levels will lead to greater levels of comprehension efficiency since subjects will be able to extract gist more rapidly, given that our comprehension measure is normalized according to clip length.

Active Condition In the active condition, because the user is able to recover from listening errors using the browser, we do not expect the compression technique to affect comprehension efficiency. Our previous experiment found that comprehension efficiency increased with the compression level and we expect this to be the same here. Again, we expect that users will make more active use of the browsing interface when processing sped up speech since this condition requires the listener to clarify key points uncompressed.

3 Results

3.1 Passive Condition

To assess the objective results for the passive condition we conducted a 3 (compression level) X 3 (compression type) ANOVA with comprehension efficiency as the dependent variable. The results are shown in the two left hand graphs in Fig.3.

As predicted, we found an overall main effect of compression type on comprehension efficiency ($F_{(2,288)} = 3.386, p < 0.05$). Planned comparisons confirmed that there was no significant difference between comprehension efficiency in the two excision conditions regardless of whether the transcript was human or ASR generated ($p > 0.3$). However, as predicted, comprehension efficiency in the speed up condition was worse than in the two excision conditions combined ($p < 0.02$).

The ANOVA confirmed the effect of compression level on comprehension efficiency ($F_{(2,288)} = 7.005, p < 0.05$), with planned comparisons confirming that there was greater comprehension efficiency at the two higher compression levels compared with 1.5 times compression ($p < 0.05$).

3.2 Active Condition

To assess the objective results for the long condition we conducted a 2 (compression level) X 3 (compression type) MANOVA with comprehension efficiency and interface actions as the dependent variables.

Consistent with our predictions, there was no effect of compression technique on comprehension efficiency ($F_{(2,60)} = 0.302, p > 0.7$). As in our previous study it seems that the use of the interface allows subjects to extract gist efficiently, independently of the interface condition by stopping, uncompressing the speech and by replaying elements they failed to understand.

We found no effect of compression level on the comprehension efficiency in the long condition ($F_{(1,60)} = 1.254, p > 0.25$). We think that this could be an effect of the interface controls causing any potential processing gains afforded by the shorter playing time to be dampened.

As in our previous study we found that there was an effect of compression technique on the number of interface actions performed ($F_{(2,60)} = 9.593, p < 0.01$). Planned comparisons indicated that more actions were used in the speed up condition ($p < 0.01$) but there was no significant difference between the actions used in the excision conditions ($p > 0.7$). There was no indication therefore,

that subjects had to make more adjustments with ASR than human-transcript excision.

3.3 Qualitative Results

An analysis of the questionnaire results for the passive condition shows a main effect of condition on the subject answers ($F_{(2,30)} = 5.117, p < 0.01$). Tukey planned comparisons indicate that this was a result of the differences in the answers for the excision and speed up (each $p < 0.05$); there was no difference between the answers given for the excision conditions (each $p > 0.08$). In the active condition listeners felt that the “speech was too fast” in the speed up case compared with excision ($p < 0.03$) and that they “repeatedly had to go back in the speech” ($p < 0.05$) in speedup compared with excision conditions. We found no subjective difference between the excision conditions ($p > 0.08$).

4 Discussion

This paper examines the effects of using ASR transcripts to construct compressed meeting recordings to support gist extraction. We found no differences in gist extraction performance between human generated and ASR transcripts for state of the art ASR error levels (38% WER). ASR also had no effect on the interaction strategies employed by listeners when processing temporally compressed speech nor on the subjective assessment of the techniques. However ASR methods were superior to state of the art techniques currently used for speedup [2]. Other findings are consistent with our previous work[14].

Our results extend our previous work on temporal compression and add to the body of literature which shows that errorful transcripts can be highly useful in a variety of other tasks, e.g. speech retrieval or speech browsing ([3, 18]). Given that these experiments were carried out using informal multi-participant conversational speech the approach taken here could also be promising when applied to other speech domains such as news domains or recorded presentations.

Whilst our compression algorithms work well in this domain there are several ways they might be improved. Firstly, we rely exclusively on lexical techniques to compute importance; more sophisticated measures of utterance importance (using syntactic or prosodic information [8] or even information about speaker role[11]) could lead to improvements in comprehension efficiency. Secondly, we could also exploit other sources or metadata to refine our importance scores - for example using visual cues to participant attention [19], or slide usage to indicate regions of high interest to meeting participants.

Future work will examine the use of these techniques for accessing other types of information - for example we have shown that they work well for extracting gist, but it is an open question as to how effective they are when used to answer more specific, factual, styles of questions.

5 Acknowledgments

This work is supported by the European IST Programme Project FP6-0033812.

References

1. R. Baeza-Yates and J.E. Maki. *Modern Information Retrieval*. Addison Wesley, 1999.
2. M. Covell, M. Withgott, and M. Slaney. Mach1 for nonuniform time-scale modification of speech. In *Proceedings of ICASSP 98*, 1998.
3. J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees. The TREC spoken document retrieval track: A success story. In *RIAO 2000: Content-Based Multimedia Information Access*, volume 1, pages 1–20, 2000.
4. T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, I. McCowan, D. Moore, V. Wan, R. Orderland, and S. Renals. The 2005 AMI system for the transcription of speech in meetings. In *Proceedings of the Measuring Behavior 2005 symposium on Annotating and Measuring Meeting Behavior*, 2005.
5. L. He and A. Gupta. User benefits of non-linear time compression. Technical Report MSR-TR-2000-96, Microsoft Research, September 2000.
6. D.J. Hejna. Real-time time-scale modification of speech via the synchronized overlap-add algorithm. Master’s thesis, M.I.T., 1990.
7. K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.
8. K. Koumpis and S. Renals. Automatic summarization of voicemail messages using lexical and prosodic features. *ACM Transactions on Speech and Language Processing*, 2, 2005.
9. I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner. The AMI meeting corpus. In *Proceedings of the Measuring Behavior 2005 symposium on Annotating and Measuring Meeting Behavior*, 2005.
10. N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbert, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke. The meeting project at ICSI. In *Human Language Technologies Conference*, March 2001.
11. G. Murray, P. Hsueh, S. Tucker, J. Kilgour, J. Carletta, J.D. Moore, and S. Renals. Automatic segmentation and summarization of meeting speech. In *Proceedings of NAACL-HLT 2007*, April 2007.
12. G. Murray and S. Renals. Term-weighting for summarization of multi-party spoken dialogues. In A. Popescu-Belis, S. Renals, and H. Bourlard, editors, *Machine Learning for Multimodal Interaction IV*, volume 4892 of *Lecture Notes in Computer Science*, pages 155–166. Springer, 2007.
13. A. Nenkova, R. Passonneau, and K. McKeown. The pyramid method: incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing*, 4(2), 2007.
14. S. Tucker and S. Whittaker. Time is of the essence: An evaluation of temporal compression algorithms. In *Proceedings of CHI '06*, pages 71–80, April 2006.
15. S. Tucker and S. Whittaker. Temporal compression of speech: An evaluation. *IEEE Transactions on Audio, Speech and Language Processing*, 2008. In Press.

16. P. Wellner, M. Flynn, and M. Guillemot. Browsing recording of multi-party interactions in ambient intelligent environments. In *Proceedings of Conference on Human Factors in Computing Systems (CHI)*, April 2004.
17. P. Wellner, M. Flynn, S. Tucker, and S. Whittaker. A meeting browser evaluation test. In *Proceedings of Conference on Human Factors in Computing Systems (CHI)*, April 2005.
18. S. Whittaker, J. Hirschberg, B. Amento, L. Stark, M. Bacchiani, P. Isenhour, L. Stead, G. Zamchick, and A. Rosenberg. SCANMail: A voicemail interface that makes speech browsable, readable and searchable. In *Proceedings of Conference on Human Factors in Computing Systems (CHI)*, April 2002.
19. J. You, G. Liu, L. Sun, and H. Li. A multiple visual models based perceptive analysis framework for multilevel video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(3):273–285, March 2007.