

NOVEL TECHNIQUES FOR TIME-COMPRESSING SPEECH: AN EXPLORATORY STUDY

Simon Tucker and Steve Whittaker

Department of Information Studies, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP, UK
Email: {s.tucker,s.whittaker}@sheffield.ac.uk

ABSTRACT

We present a novel technique for time-compressing speech, semantic compression, which uses an ASR transcript to determine which elements of the speech are presented. We carried out an exploratory user study comparing semantic compression to other novel types of time-compression techniques. We found that users feel they have a greater understanding of recordings compressed using semantic techniques than those compressed using acoustic-based techniques. An approach of using high playback speeds to indicate the location of 'insignificant' speech is not favoured by listeners, who prefer to have such segments removed from recordings.

1. INTRODUCTION

Speech has the benefit of being ubiquitous, expressive and easy to produce compared with text, although until recently we lacked effective tools for accessing speech archives. In the last few years, a number of interface techniques have been developed that use visual representations to browse and access speech e.g. [8,10,11]. However this work presupposes that users will have some form of visual display. In contrast in this paper we focus on situations (such as phone or mobile access) which have speech-only or very limited visual output. Specifically we investigate the utility of techniques for time-compressing speech, i.e. decreasing the time taken to listen to a speech recording while retaining significant content.

Previous research has focused on two main time compression techniques: silence removal and speech rate alteration [2]. It is possible to compress a speech recording by approximately 20% using silence removal although this approach is clearly limited since recordings contain a finite amount of silence. Speech rate alteration is a method of artificially altering the playback speed; techniques for altering the speech rate range from simply altering the sample frequency (which alters the playback rate, but also the pitch of each speaker), to more complex frequency domain alterations [9]. Typically, however, speech rate alteration is carried out using the synchronized overlap add method (SOLA) or one of its variants. In SOLA, short frames of the speech recording are overlapped with one another in order to reduce the playback time. The overall compression rate is determined by the degree of overlap and generally a form of correlation is used to ensure that the overlap does not cause distortions. Listening experiments show that users can process speedups of between 2-3 times normal speaking rate depending on the familiarity of the underlying material [2]. A more recent advance in speech rate alteration uses non-linear compression techniques [5]. Here the degree of

compression varies throughout playback so that different portions of the recording are played back at different speeds - for example when the underlying speech is relatively slow, the playback speed will be relatively high and vice versa. Non-linear compression techniques have been shown to be superior to linear methods at relatively high compression levels [4]. However at compression levels which are considered tolerable for extended listening (ranging from 55-80% compression) the non-linear techniques do not "offer a significant advantage" [6].

The techniques described above focus on the acoustic properties of the speech. In this study we compare these acoustic techniques with time compression which makes use of the semantics of the speech. Semantic compression is based on the observation that when browsing speech, listeners do not pay equal attention to all information elements of the speech - instead they focus on important words or salient parts of the speech ignoring less important elements [3]. Semantic compression analyses transcripts generated using automatic speech recognition (ASR) to identify important elements of underlying speech, which are then played to the user. We can use various methods to identify important elements in the ASR transcript. Here we focus on text summarisation and insignificant word removal.

The current study compares several different semantic and acoustic compression methods. We also wanted to compare differences between techniques that excise unimportant material, and those that present it speeded up. We use an exploratory method, because we want to probe the effectiveness of a range of different techniques. Nevertheless we evaluate three specific hypotheses.

H1: Semantic versus Acoustic Techniques

We compare semantic techniques based on text summarisation and word significance with various acoustic techniques using speed up and silence excision. We expect that users will find that clips processed with semantic techniques are easier to understand than those based on acoustic properties, because semantic techniques allow them to focus better on important material.

H2: Speed Up versus Excision

We also compare speed up techniques which preserve all original material with excision techniques which remove unimportant material. We would expect that users will prefer techniques which make use of speed up over those that make use of excision, as omitting material may make it hard to understand the original clip.

H3: Compression Rates

We also investigate overall compression effects, predicting that users should have a greater understanding of clips presented with low compression than those presented with high compression.

2. EXPERIMENTAL PROCEDURE

2.1. Stimuli

Seventeen two minute clips were selected from the AMI public corpus [1]. The length of the clips meant that subjects would have time to judge the level of effort required to listen to each of the compression techniques. The excerpts were chosen so that the first minute contained a single speaker, following which was a minute of multiple speakers; this means that subjects can hear each technique in a range of listening circumstances. Corresponding speech transcripts were also taken from this corpus. Note that each transcript is split into a number of utterances, the start and end time of each utterance also being recorded for each transcript.

Clips were compressed to be either 70% (low compression) or 40% (high compression) of their original length (therefore being 84 and 48 seconds long respectively) using eight different compression algorithms. Each subject heard each algorithm under both levels of compression; therefore each subject heard a total of sixteen clips. Two basic techniques were employed to produce the compressed clips: speech rate alteration (speed up) and segmental excision. To alter the speech rate, we used a standard overlap add algorithm [7].

2.2. Time-Compression Algorithms

Acoustic Techniques. There were 5 different acoustic techniques based on different types of speed up and silence removal.

Constant Speed Up (CS)

We use a constant speech rate alteration to achieve the desired level of compression. Low compression clips were presented at a speed 1.4 times greater than real time and high compression clips at a speed 2.5 times faster than real-time.

Speed Rate Speed Up (SR)

This technique uses a measure of speech density to control the level of speed up. Specifically, we count the number of words present in each utterance. A word-density curve is then produced by normalising the number of words in each utterance by its corresponding length. This curve is then transformed so that the utterance of the highest density is played back at normal speed, with utterances of lower density being played back at greater speeds. The curve is then suitably compressed or expanded so that, when the speed up is applied, the compressed clip length matches that required by the compression rate.

Utterance Speed Up (US)

This technique tests a hypothesis about utterance level redundancy - namely that the start of each utterance provides predictable context for what follows making the ends of utterances informationally more redundant than their beginnings. We therefore increase speech rate from beginning to end of the

utterance. The start of each utterance is played in real time, with speed linearly increasing over the course of the utterance. The slope of the speed up determines the overall length of the clip and is, therefore, computed so that the clip is of the length required by the compression rate.

Silence Excision (SE) and Silence Speed Up (SS)

Each clip is split into 30 ms frames, with a frame overlap of 5 ms, and we compute the spectrum of each frame. A period of silence is then manually identified and the average of the corresponding spectral frames is computed to produce an exemplar of the silence spectrum in the clip. The similarity between each frame and the exemplar is then computed and the frames are ranked according to their similarity to the silence exemplar.

In the excision case frames with low similarity are progressively included into the compressed recording until the length of the excised clip matches that required by the given compression rate. In the speed up case, frames with a low similarity are played back at normal speed, whilst frames with a high similarity are played at 3.5 times real time. Frames are progressively marked as playing at normal speed until the overall length of the clip matches that required by the compression rate.

Semantic Techniques. There were 3 semantic techniques using text summarisation and insignificant word removal.

Summary Excision (ME) and Summary Speed Up (MS)

We construct increasingly lengthy extractive summaries using utterances from the ASR transcript. Each utterance is then ranked according to the number of summaries it appears in. Thus higher ranking utterances appear in very short summaries, whereas lower ranking utterances only appear in lengthy summaries. The motivation behind this is to rank utterances according to the amount of relevant information they contain.

For the excision case, utterances are progressively included (according to their ranking) until the clip is of the length required by the compression rate. In the speed up case, higher ranking utterances are played at normal speed, whereas the lower ranking utterances are presented at a speed 3.5 greater than real time. One hypothesised benefit of summary speed up is that it provides an auditory cue to the user about how much insignificant material is being skipped over. In the summary excision case users are unaware of where and how much material has been omitted.

Insignificant Word Excision (IW)

This technique begins by constructing a dictionary of all the words contained in the transcripts in the corpus, as well as the frequency of occurrence of each of these words. A similar dictionary for the transcript of the clip being processed is also produced. Each word is ranked according to the term frequency (the number of times the word appears in the current transcript) divided by the document frequency (the number of times the word appears in the corpus). In this way words with a high ranking are deemed to carry more 'significance' than those with a low ranking. We produce the clip by progressively including higher ranking words until the clip is of the required length.

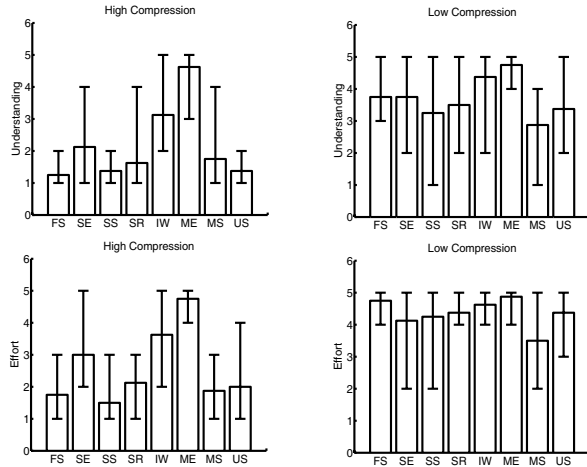


Figure 1. Understanding and Effort ratings organised by compression rate and compression algorithm. The main bar indicates the mean value, with the error bars showing extremities. A higher value means greater agreement with the qualifying statement, "I felt I had an overall understanding of this clip".

2.3. Procedure

Experiments took place in an acoustically isolated booth, with clips being presented to listeners diotically over headphones. Each subject began the experiment by hearing a two minute unprocessed clip in order for them to familiarise themselves with the content type of the clips and the experimental procedure. Subjects were required to rate their overall understanding of the clip and the effort required to listen to the compressed meeting excerpt. Each rating was made on a five point Likert scale, the qualifying statements being "I feel I had an overall understanding of this clip" and "Understanding this clip took little effort". Subjects were instructed that they should focus on their understanding of the speech and not the meeting content; furthermore, subjects were instructed that some speech would be deliberately hard to follow and that they should focus on their understanding of the overall clip. Following the main experiment subjects were presented with a screen which allowed them to replay each clip at high compression. Nine subjects were encouraged to leave specific comments on each of the compression techniques in a small text box.

The order of presentation of clips, compression rates and compression techniques was randomized for each subject with the conditions that no subject heard the same compression technique twice in succession and that each subject heard a different compression technique applied to each clip. Eight subjects were chosen from native English postgraduates at the University of Sheffield. Users were given a small reward for their participation.

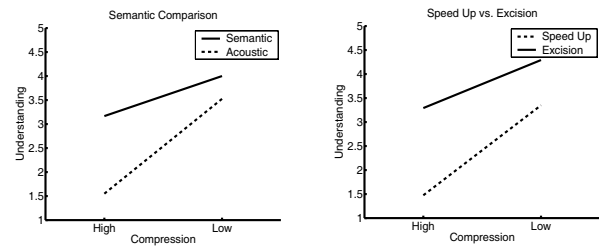


Figure 2. The top panel shows the mean understanding value for semantic and acoustic based techniques, organised by compression rate. The bottom panel shows the mean understanding value for speed up and excision based techniques, organised by compression rate.

3. RESULTS

The overall results are shown in Figure 1. Note that the results indicated a high level of correlation ($r = 0.841$, $p < 0.01$) between subject ratings of understanding and effort (see discussion below). In the text that follows only the results for understanding are discussed.

3.1. Hypothesis Results

H1: Semantic Techniques. Our first hypothesis was confirmed by the results of the experiment ($F(1,96) = 31.045$, $p < 0.001$). As can be seen in Figure 2 it is apparent that subjects felt they had a greater understanding of clips which made use of semantic properties of the transcript than those which were based on acoustic properties of the signal. Amongst the semantic techniques summary excision was most favoured ($\mu_{ME} = 4.688$) although insignificant word excision was statistically indistinct from this ($\mu_{IW} = 3.750$; $t_{ME-IW} = 0.565$, $p < 0.580$). Summary speed up was statistically distinct from both of these techniques at the 5% level ($\mu_{MS} = 2.313$; $t_{IW-MS} = 5.258$, $p < 0.01$).

H2: Speed Up versus Excision. Figure 2 shows the mean understanding grouped by compression type for both levels of compression. As can be seen it is clear that subjects felt they had a greater understanding of clips based on excision than with those based on speed up ($F(1,124) = 54.855$, $p < 0.001$). Performing a specific comparison, subjects felt they had a greater understanding of summary excision than summary speed up ($\mu_{ME} - \mu_{MS} = 2.38$; $t = 8.733$, $p < 0.01$); there was also a preference for silence excision over silence speed up ($\mu_{SE} - \mu_{SS} = 0.62$; $t = 2.44$, $p < 0.028$). We explore these excision results in more detail below.

H3: Compression Rates. As Figure 1 shows it is clear that subjects felt they had an increased understanding of low compressed clips. This effect is confirmed by the ANOVA analysis ($F(1,112) = 90.391$, $p < 0.001$).

3.2. Subject Comments

By analysing the comments taken during the experiment and more general comments made by subjects at the end of the experiment we were able to identify four elements which were commented on by several subjects.

Speed up as a means of indicating missing speech is distracting. We used two techniques, speed up and excision, to deal with parts of speech that were deemed insignificant. The results indicate that subjects preferred that such segments be excised rather than played back at a greatly increased rate; reasons were that the sped up sections were distracting:

"[It's] hard to maintain concentration across period of disruption"

and also they added little useful information:

"Speedup segments were unintelligible. Why not just skip completely?"

No subject indicated that speeding up insignificant segments was a useful cue as to what they were missing.

Meeting Participant Effects. Subjects also commented that their understanding was also highly dependent on the current speaker. Specifically, subjects indicated that people that they knew were easier to understand and, furthermore, that certain accents were more understandable than others:

"I could only understand the Australian and American in this clip"

"I had less trouble understanding the Australian over the other clip"

The comments were not limited to English speakers alone:

"Too fast except for the Indian"

Information load in utterances. We initially thought that the start of an utterance carries more information than the end, this being the motivation behind the utterance speed up condition. This hypothesis was contradicted by both the experimental results and the subjective comments:

"Important info is mashed up...can not understand the gist of the conversation"

Correlation of Effort and Understanding. An examination of the subjective comments indicated that listeners made a correlation between understanding and effort:

"I felt that I could have understood more if I had put in more effort"

This could possibly explain the high levels of correlation in listener judgements of effort and understanding.

4. CONCLUSION

We presented several novel approaches to time-compressing speech recordings that used the speech transcript to identify significant portions of recordings and time-compressed the insignificant segments. Since this is an unintuitive and unexplored space of techniques an exploratory procedure was employed

collecting subjective data and comments from listeners. The results indicated subjects felt they had a greater understanding of semantic techniques over acoustically motivated approaches. Furthermore a technique which used an automatically generated summary to excise utterances of low significance was most favoured by listeners. This technique was superior to both standard silence removal approaches and to a high-level non-linear speed up technique. The novelty of the techniques the experiment described here is necessarily preliminary and future work will evaluate a subset of the techniques described above using a more rigorous experimental procedure.

5. REFERENCES

- [1] *The AMI Public Corpus*, <http://mmm.idiap.ch/>
- [2] B. Arons, "Techniques, perception, and applications of time-compressed speech," *Proceedings 1992 Conference, American Voice I/O Society*, pp. 169-177, 1992.
- [3] B. Arons, "SpeechSkimmer: A system for interactively skimming recorded speech", *ACM Trans. on Computer-Human Interaction*, 4, pp. 3-38, 1997.
- [4] M. Covell, M. Withgott & M. Slaney, "Mach1: nonuniform time-scale modification of speech," *Proc. ICASSP'98*, 1998.
- [5] J. Foote, J. Boreczky, A. Girgensohn & L. Wilcox, "An intelligent media browser using automatic multimodal analysis," *ACM Multimedia 98*, pp. 375-380, 1998.
- [6] L. He & A. Gupta, "User benefits of non-linear time compression," *Microsoft Research Technical Report*, MSR-TR-2000-96, 2000.
- [7] D.J. Hejna Jr., "Real-time time-scale modification of speech via the synchronized overlap-adds algorithm," *Masters Thesis*, MIT, 1990.
- [8] D. Hindus, C. Schmandt, C. Horner, "Capturing, structuring, and representing ubiquitous audio", *ACM Trans. on Information Systems*, 11, pp. 376-400, 1993.
- [9] M.R. Portnoff, "Time-scale modification of speech based on short-time Fourier analysis," *IEEE Trans. on Acoustics, Speech and Signal Processing*, 2, pp. 374-390, 1981.
- [10] S. Vemuri, P. DeCamp, W. Bender, and C. Schmandt, "Improving speech playback using time-compression and speech recognition," *Proc. of CHI 2004*, pp. 295 - 302, 2004.
- [11] S. Whittaker, J. Hitschberg, 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," *Proceedings of CHI 2002*, pp. 275-282, 2002.