

# Evaluating Methods of Temporally Compressing Speech

Simon Tucker and Steve Whittaker

**Abstract**—Efficient browsing of speech recordings is problematic. The linear nature of speech, coupled with the lack of abstraction that the medium affords, means that listeners have to listen to long segments of a recording to locate points of interest. We explore *temporal compression* algorithms that attempt to reduce the amount of time users require to listen to speech recordings, whilst retaining the important content. This paper implements two main approaches to temporal compression: *artificial speech rate alteration (speed-up)* and *unimportant segment removal (excision)*. We evaluate the effectiveness of these approaches by having listeners rate comprehension and listening effort for different types of temporal compression. For different compression levels, we compare performance of various implementations of speed-up and excision as well as techniques based on semantic features and acoustic features. Our results indicate that listeners prefer low compression levels, excision over speed-up, and algorithms based on semantic rather than acoustic features. Finally listeners were negative about hybrid algorithms that used speed-up to indicate missing regions in an excised recording.

**Index Terms**—Speech Processing, Text Processing, Information Retrieval, User Interfaces

## I. INTRODUCTION

SEVERAL large-scale projects (e.g. [1]–[3]) have built complex visual tools to help users review multimedia meeting records. These systems record audio and video media, along with other relevant meeting events such as logs of personal notes, whiteboard markings and presentation slides. They supplement this raw data to construct a rich meeting record, by adding derived data (such as automatic speech recognition (ASR) transcripts) and annotations (e.g. emotion types [4], or ‘hotspots’ [5]). The meeting record can then be browsed using a variety of indices (e.g. speaker, topic, ‘hotspots’, slide changes, personal notes or ‘interesting’ visual events - see [6], for a review). These indices allow users to locate specific facts or regions within the recording.

One limitation of these systems is that they require complex visual displays, along with storage and indexing of multiple data types. In contrast, our focus here is on efficient access to speech using small displays such as those available on PDAs and mobile phones. We explore different *temporal compression* methods that aim to reduce the amount of time it takes to listen to a speech recording whilst retaining all of its important information. We investigate two different compression techniques: *excision* where unimportant portions of the recording are removed and *speed-up* where the playback rate

is altered while keeping speaker pitch constant [7]. Throughout this paper the percentage of the original duration is used as a measure of the level of compression applied to a recording. We first discuss prior research in Temporal Compression in order to motivate the new techniques we explore here.

Excision reduces the recording length by removing automatically selected portions of audio data, effectively compressing it. One simple excision technique removes the between-word silences in the recording [8]. However the amount of compression that can be achieved using silence removal is restricted by the amount of silence in the original recording. For the corpus used in these experiments, on average only 25% of the meeting recordings were silence. We therefore examine an acoustically motivated excision technique which attempts to remove parts of the recording which are acoustically similar to silence regardless of whether these parts contain speech. We also explored a new temporal compression technique using semantic information at a coarse grained word level, as opposed to the sentence level compression techniques described in the literature. An alternative excision method makes use of text-based summarization techniques applied to (usually human-generated) meeting transcripts. It uses meeting-specific discourse features to produce an extractive summary to determine which parts of the recording should be played to listeners ([9], [10]). While this approach is promising for broadcast news [11], it is unclear how well it will generalize to meetings data which is of higher acoustic complexity and is much less structured.

Speed-up constructs a new recording in which the speaking rate has been artificially altered. Techniques for altering speech rate range from simple sample-frequency alteration (which changes both the playback rate and speaker pitch), to more complex frequency domain procedures [12]. Alteration is largely carried out in the time-domain; a technique, popular for its efficiency and quality, is the synchronized overlap add method (SOLA) which successively overlaps small segments of the recording [13]. The amount of compression applied to the recording is determined by the amount of overlap between these segments. User studies have shown linear speed-up to be effective. Users can comprehend information played at twice its normal rate and after exposure to sped up speech, they prefer it to the normal speech rate ([14], [15]). More recent work explores non-linear speed-up (e.g. [16]). While linear techniques apply constant compression throughout the recording, non-linear approaches vary compression continuously according to external factors. [17] describe a nonlinear compression technique which implements a variable playback rate mimicking the way human speakers naturally increase their talking speed. Once the playback rate has been determined, a

Manuscript received January 20, 2002; revised November 18, 2002. This work was supported by the IEEE.

The authors are with the Department of Information Studies, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP (e-mail: {s.tucker,s.whittaker}@shef.ac.uk).

modified SOLA technique is used to perform the compression. Non-linear approaches are superior at higher compression levels; but at compression levels considered tolerable for extended listening (ranging from 55-80%), nonlinear techniques do not offer a significant user advantage [18]. The non-linear approaches described in the literature derive the playback rate from acoustic properties - in this experiment we additionally examine non-linear speed-up techniques which use semantic information to derive the playback rate. For example, we explore the effects of speed-up *within* an utterance - motivated by the finding that listeners can often infer what will be said towards the end of an utterance [19]. Finally we devised novel *hybrid* techniques that combine speed-up and excision. Having identified unimportant utterances or silences we can speed through these instead of excising them altogether.

While some prior work has explored the effect of different temporal compression techniques on the user ([8], [9], [14], [15], [18]), these evaluations have not been *comparative*; For example, they have not compared the utility of different speed-up techniques with excision techniques. We compare user judgments of the comprehensibility and effort required to listen to different compressed audio recordings, for three different types of temporal compression techniques. Firstly, we compare excision and speed-up techniques. Secondly, we compare compression techniques based on the *acoustic* waveform (e.g. silence removal, or speed-up) with *semantic* techniques based on lexical information extracted from the transcript. Although these algorithms may differ from natural speaker strategies we nevertheless wanted to test their effects. Finally we evaluate hybrid techniques that combine speed-up and excision. Altogether we compare eight different algorithms. We also compare the effects of overall compression level and how this interacts with different techniques, e.g. is speed-up effective at low compression levels but not at high levels? We explore this for two levels of compression: 70% and 40%. In summary, the study compares multiple acoustic and semantically based compression methods which make use of both excision and speed-up as a means of reducing the length of the recording. We use subjective evaluation, as our focus is on the user's listening experience.

## II. COMPRESSION TECHNIQUES

The temporal compression techniques we developed and evaluated can be classified into six groups (see Table I), according to Compression Type (whether they make use of Excision, Speed-up or a Hybrid combination of both for compression), or Compression Basis (whether they use Semantic or Acoustic features for compression). Each cell of this table is described in detail below.

### A. Acoustic Excision

The first cell of Table I, acoustic excision, covers one acoustically motivated technique (*silence excision*) that removes silences, on the grounds that these do not convey important information. Standard silence excision techniques [8] are constrained by the amount of silence in the original recording. To overcome this limitation, we therefore developed

a technique that measures the *silence similarity* of audio segments. We then set a threshold, with segments above the threshold labeled as silence-like, which are then excised.

Each clip is split into 30 ms frames with an overlap of 5 ms. A Hanning window is applied to each frame and the spectrum is computed using the Fourier transform. A short period of silence is then manually identified. Whilst several silence identification algorithms exist (e.g. [20]), a manual approach is used to prevent automatic identification errors from affecting the resulting temporal compression. The spectrum of the silence portion of the recording is then computed and this is used to produce a spectral exemplar of silence for the entire recording.

The similarity between each spectral frame of the recording and the silence exemplar is then computed using the Euclidean distance. Each frame is then ranked according to this measure. Frames are progressively included in the compressed recording according to this ranking, starting with an empty file. We begin with the lowest silence-ranked segments - and continue adding frames until the length of the excised clip matches that required by the given compression rate. Frames are presented in the order they occurred in the original recording.

### B. Semantic Excision

The second category of techniques, semantic excision, use a meeting transcript and information retrieval techniques to identify less important speech segments which are then excised from the recording. We examine two different semantic excision methods - *insignificant utterance* and the novel techniques of *insignificant word* excision. Insignificant utterance excisions are derived from the hand-generated transcript. We first compute an importance measure for each utterance in the meeting. Important utterances are retained, while unimportant utterances are excised - allowing users to focus on important parts of the meeting. Insignificant word excision computes the importance of each word in the transcript, and then excises less important words. The motivation behind this approach is that listeners should be able to fill in the missing lexical gaps themselves, based on the gist of what they have heard.

We compute an importance score for each word in the transcript using a simple term frequency, inverse document frequency score (TF\*IDF) [21]. In this measure, the number of times a term appears in a single document is multiplied by the inverse frequency of the term appearing in the set of documents, thus the importance (*imp*) of a term, *t*, appearing a particular transcript, *d*, can be computed as:

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

Where  $count_{td}$  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 and  $N_t$  is the number of transcripts which contain the term *t*. The set of transcripts from the AMI corpus is used as the document set.

From this word level measurement, we compute the importance of each utterance to be the mean importance of each word in the utterance. Utterances are then ranked according

to their importance. Compression then continues in a similar way to that described above; high-importance utterances are progressively included into an empty file until the desired level of compression is reached. Utterances are presented in the order in which they occurred in the original recording.

Insignificant word excision works in a similar way. We again compute the importance of each word of the transcript using  $TF*IDF$ . The words themselves are then ranked according to this importance measure. The compressed clip is again constructed by progressively including words of highest importance until the desired level of compression is reached. Words are presented in the order in which they occurred in the original recording. A consequence of producing the files in this way is that between-word silences are removed during compression. Thus the effect on the listener is a stream of the important words contained in the recording.

### C. Acoustic Speed-up

The third group of techniques examine acoustic speed-up where we explore constant speed-up and speech-rate speed-up. With constant speed-up, clips compressed at the low compression rate are played back at a *constant speed-up* - 1.4 times real time, whilst clips compressed at the high rate are played back at 2.5 times real time - corresponding to 70% and 40% compression respectively.

We also implement *speech-rate speed-up*. Other approaches compute speech rate at the part-of-speech level. For example, Mach1 [17] adjusts the playback rate at phoneme level. In contrast, we compute speech rate for the overall utterance, using speech density to control the playback rate. We calculate density by counting the number of words in each utterance and then construct a word-density curve by dividing each utterance word count by the duration of the corresponding utterance. This curve is then normalized, so that the utterance which has the highest word density measure is played back with no speed-up and utterances of lower densities being played back at higher speeds. The curve is then expanded or compressed in order to achieve the overall compression rate. The motivation of this technique is to approximate a normalization of the underlying speech speed. When a speaker is talking fast, their speech density is high, causing that portion of the recording to be played back relatively at a slower rate than a portion when they are talking slowly.

### D. Semantic Speed-up

The fourth set of techniques investigates the novel notion of information distribution within the utterance. Psycholinguistic studies indicate that the more of an utterance listeners have heard, the better they are able to predict what will next be said - suggesting that end-of-utterance information is partially redundant [19]. *Utterance position speed-up* exploits this by progressively speeding up each utterance - so that the beginning of each utterance is played back in real time, but as it continues the playback speed increases. To achieve this affect, the gradient of the speed increase is constant across each utterance, but is altered until the desired compression rate is reached.

Finally we implemented two sets of novel hybrid techniques that combine speed-up and excision, one using acoustic information and the other using semantic cues to identify unimportant information. Instead of completely excising unimportant information, hybrid techniques present it sped up. This is to address a potential problem with excision; listeners may be unaware both that unimportant material has been excised, and also how much material is missing. Presenting unimportant material sped up should indicate the location and approximate duration of that missing material.

### E. Hybrid Acoustic Speed-up

*Silence speed-up* identifies silence-like segments as described above, but instead of excising these, presents them sped up. Silence speed-up works in the same way as silence excision, except that all frames are classified as having either high or low silence similarity. The high silence similarity frames are then played back at 3.5 times real time whilst the low similarity frames are played back in real time. The balance between low and high frames is chosen so that the required compression level is met. Thus the effect on the listener is that the recording sounds normal, but with silence-like portions played back at an increased rate. The relatively high value of 3.5 speed-up is chosen so that users are aware of when they are missing portions of the recording, but they are only able to determine 'glimpses' of the context that they are missing [22].

### F. Hybrid Semantic Speed-up

A similar technique, *insignificant utterance speed-up*, identifies unimportant utterances and then presents these sped up. Again this is similar to insignificant utterance excision, except that the utterances are divided into two groups; important utterances and unimportant utterances, according to the measure of importance we used. The unimportant utterances are then played back at 3.5 times real time, with the important utterances played back in real time. Again, the balance between important and unimportant utterances is chosen so that the required level of compression is achieved.

## III. EXPERIMENTAL PROCEDURE

### A. Stimuli

The stimuli for the experiment were excerpts taken from the AMI public corpus [23], which is a set of semi-scripted meetings recorded in a designated meeting room [24]. Seventeen two-minute clips were manually selected from this corpus; two minutes was chosen as it was felt that this gave listeners sufficient time to judge the effort of listening as well as their overall understanding of the excerpt. Clips were chosen so they featured one minute of monologue followed by one minute of discussion between two or more meeting participants, allowing listeners to judge the compression techniques with a varying number of speakers. Each meeting had been hand transcribed and segmented into a number of utterances.

Clips were then compressed at 70% (which is subsequently referred to as low compression) and at 40% (which is referred

to as high compression) of their original length, thus the clips were 84 and 48 seconds long respectively. We chose these levels as other research finds that 70% is acceptable for listeners [8] and 40% represents a compression level at which non-redundant information begins to be lost [25]. We used the eight different compression techniques described above. Each subject heard each algorithm at both levels of compression, thus sixteen clips were presented to each subject.

## B. Procedure

Experiments took place in an acoustically isolated booth, with clips being presented to listeners diotically over Sennheiser HD250 headphones. A Matlab script was used to both present the excerpts and collect the results. Each subject began the experiment by listening to a two minute unprocessed training clip in order to familiarize themselves with the content of the meeting excerpts and the experimental procedure. This was taken from the main corpus and was the same for each subject.

After listening to each clip the subjects were asked to rate their understanding and effort on two separate 5-point Likert scales. The qualifying statement for rating understanding was "I feel I had an overall understanding of this clip" and the statement for effort was "Understanding this clip took little effort". To ensure that subjects were not affected by the discussion topics, they were instructed to concentrate on their understanding of the speech, and not the specific meeting content. We also informed them that some of the conditions were deliberately made to be difficult to understand, so they should focus on their overall understanding rather than specific details. Once subjects had completed the main experiment, they were presented with a webpage that allowed them to replay each condition at the high level of compression. We then asked them to leave free text comments comparing and contrasting the different techniques. The order of presentation of clips, compression rates and techniques were randomized for each subject with the exception that no subject heard the same compression technique twice in succession and that no clip was heard twice by a single subject. The 8 subjects were native English postgraduates at the University of Sheffield.

## C. Hypotheses

We evaluated three main hypotheses:

1) *H1: Compression:* We predict that users should have a greater perceived understanding of low compression clips than those presented with high compression, as high compression induces additional cognitive processing. For the same reason they should also judge they expend less effort in understanding low compression clips.

2) *H2: Semantic versus Acoustic Techniques:* We predict that users will find that clips processed with semantic excision techniques easier to understand than those based on acoustic properties, because semantic techniques allow them to focus better on important material.

3) *H3: Speed-up versus Excision:* Here we have two hypotheses.

**H3a:** We expect that users will prefer speed-up over excision, as excision omits material - resulting in loss of context, and making speech harder to understand.

**H3b:** For the same reasons, we expect hybrid techniques that speed-up insignificant segments will be preferred to excision techniques that simply remove these segments. This is again because excision may result in loss of context for significant materials, making them hard to understand.

## IV. RESULTS

The overall results for the experiment are shown in Fig. 1. It is apparent from the graphs that there is little difference between subjective judgments of effort and understanding, possibly because subjects considered that something that was difficult to understand also required substantial effort to process. This close relation between understanding and effort is supported by the Pearson product moment correlation coefficient and associated t-test ( $r_{(128)} = 0.841, p < 0.01$ ).

### A. Observations

Figure 1 shows judgments for the various techniques. At low compression levels, users judge themselves to be able to understand the clips, suggesting compression has little effect on perceived understanding. But there are differences between techniques: semantic excision techniques (IWE, IUE) were favoured by listeners, even at low levels of compression. At high levels of compression the picture is rather different; only excision techniques are highly ranked, with subjective effort and comprehension being much lower for other techniques.

### B. Hypotheses

To investigate hypotheses H1-3 two ANOVAs were carried out with Compression Level (High vs. Low), Type (Speed-up vs. Excision vs. Hybrid) and Basis (Semantic vs. Acoustic) as independent variables and ratings of comprehension and effort as dependent variables.

1) *H1: Compression:* As Fig. 1 shows, confirming H1, there is a clear preference for low over high compression for both comprehension ( $F_{(1,128)} = 79.6, p < 0.001$ ) and effort ( $F_{(1,128)} = 136.1, p < 0.001$ ). There was also an ANOVA interaction between Level and Basis ( $F_{(2,128)} = 3.1, p < 0.081$ ) and ( $F_{(2,128)} = 4.2, p < 0.05$ ) for comprehension and effort respectively, which as Fig. 2 shows was due to low ratings of acoustic techniques under high compression. There was also an interaction between Level and Type ( $F_{(2,128)} = 7.1, p < 0.001$ ) and ( $F_{(2,128)} = 5.0, p < 0.01$ ) for comprehension and effort respectively, which post hoc tests indicated was due to low evaluations of speed-up and hybrid techniques at high compression.

2) *H2: Semantic vs. Acoustic Techniques:* As Fig. 2 shows, semantic techniques were judged better than acoustic ones. This is supported by the ANOVA for comprehension with a main effect of Basis ( $F_{(1,128)} = 4.5, p < 0.05$ ) but not for effort, ( $F_{(1,128)} = 1.9, p < 0.169$ ). Our results therefore

partially confirm H2. The lack of an effect for effort could be explained by the more uniform ratings of effort under low compression compared with understanding (See Fig. 2).

3) *H3a: Speed-up vs. Excision*: Fig. 2 and the ANOVA disconfirm our predictions for H3a and H3b. As expected there was a main effect of Compression Type ( $F_{(2,128)} = 21.2, p < 0.001$ ) and ( $F_{(2,128)} = 20.7, p < 0.001$ ) for Understanding and Effort respectively, but contrary to our expectations, planned comparisons showed that people preferred excision to speed-up for both measures ( $F_{(1,125)} = 24.9, p < 0.001$ ) and ( $F_{(1,125)} = 21.1, p < 0.001$ ) for effort and understanding respectively.

4) *H3b: Hybrid vs. Excision*: Again our predictions were disconfirmed. Planned comparisons of Compression Type showed that people preferred excision over hybrid techniques despite the fact that hybrids provided information about missing materials; ( $F_{(1,94)} = 24.3, p < 0.001$ ) and ( $F_{(1,94)} = 12.8, p < 0.001$ ) for effort and understanding respectively.

### C. Comments

Subjects made comments contrasting compression techniques during the experiment, at its end, and after the experiment. An analysis of the comments indicated four themes common across a number of subjects. Speed-up is distracting as a means of indicating excised speech in the hybrid techniques. Subjects preferred unimportant segments to be excised rather than played back at a greatly increased rate; they felt that the sped up sections were distracting: *"[It's] hard to maintain concentration across period of disruption."* Sped-up segments also seemed to add little useful information: *"Speed-up 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. There were also observations about the information load in utterances. Subjects did not feel utterance position speed-up was effective: *"Important info is mashed up...cannot understand the gist of the conversation."*

In common with other studies [26], subjects also commented that their understanding was highly speaker dependent. They indicated that certain accents were more understandable than others: *"I could only understand the Australian and American in this clip"* and *"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"* Subjects felt was a strong relation between understanding and effort *"I felt I could have understood more if I had put in more effort"*. This could explain why we found such high correlation between these responses in the experiment.

## V. SUMMARY AND DISCUSSION

We carried out an exploratory study developing new temporal compression algorithms and comparing their effects on the perceived understanding of compressed speech. In particular the silence, excision word level excisions and the hybrid techniques described are unique to this paper, as is the subjective comparison between speed-up and excisions manipulations. Subjects found most techniques acceptable at low compression levels, with differences only emerging strongly

at higher levels. Overall, as we had anticipated, participants preferred semantic over acoustic techniques, as these allowed them to focus on important information. Contrary to our expectations however, subjects preferred excision over speed-up, which subjective comments suggest might result from the greater cognitive load imposed by speed-up. More specifically, subjects did not favour the condition in which the speed-up was constant, nor when speech density was used to control the playback rate. We also explored hybrid combinations of excision and speed-up where high levels of speed-up were intended to cue listeners about the size and nature of excised segments. Although our specific implementations did not exhaust the full range of possibilities, hybrid technique was not favoured by listeners, who would rather these unimportant portions were removed entirely. Although there may be specific reasons why certain techniques were not liked, overall these preferences for semantic over acoustic and excision over speed-up were strong and reliable.

Our results have direct implications for current temporal compression research. This either: (a) uses sophisticated summarization techniques to determine what should be played to the listener ([9], [10]); or (b) combines speed-up with visualizations of the underlying speech (e.g. derived from an ASR transcript) [26]. Our results suggest that speed-up may not be the optimal technique for performing temporal compression - as participants prefer excision. We also had good reactions to insignificant word removal, indicating that standard insignificant utterance excision is not the only excision technique that results in an understandable compression.

The focus of this experiment was on the general level of understanding of a short meeting excerpt, i.e. identifying its gist. But research suggests there are other tasks that users need to carry out with speech data, such as fact-finding or summarizing ([27], [28]). Future research should address how well our new techniques support these different types of task, as well as for longer speech excerpts where we might expect greater benefits for compression.

Our experiment presented users with a single clip played from beginning to end. But a single compression technique, rigidly applied, need not be the only effective way to access speech. Different access tasks may demand a combination (though, as noted above, not a simultaneous combination) of compression techniques. In addition, users may want to control the amount and type of compression to suit their processing needs. For example, they may want to determine the level of playback granularity according to the time they have available to listen to the recording and type of information they wish to collect. Again we need to examine how effective different techniques are when users have more control over what they hear [29].

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. Several projects (e.g. [26], [27], [30]) have shown that users can exploit errorful transcripts to access information. In this study we did not use such errorful transcripts, because we wanted to test the upper bound of excision techniques, but

future work needs to explore how these techniques fare with errorful transcripts.

In conclusion, we have developed various techniques for temporally compressing speech. We have also shown, in user studies, that excision techniques and those that rely on semantics are more acceptable than speed-up and acoustic techniques that have previously been deployed. In future work we plan to develop and evaluate these techniques further in particular to test, whether using our techniques, users can understand speech at even higher compression levels.

## REFERENCES

- [1] "AMI project," 2005. [Online]. Available: <http://www.ami.org>
- [2] "IM2 project," 2005. [Online]. Available: <http://www.im2.ch>
- [3] "M4 project," 2005. [Online]. Available: <http://www.m4project.org>
- [4] B. Schuller, R. Müller, M. Lang, and G. Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles," in *INTERSPEECH 2005, Special Session: Emotion Speech Analysis and Synthesis: Towards a Multimodal Approach*, 2005, pp. 805–809.
- [5] B. Wrede and E. Shriberg, "Spotting "hot spots" in meetings: Human judgements and prosodic cues," in *European Conference On Speech and Communication Technology*, 2003, pp. 2805–2808.
- [6] S. Tucker and S. Whittaker, "Accessing multimodal meeting data: Systems, problems and possibilities," in *Lecture Notes In Computer Science*, B. S. and H. Bourlard, Eds., 2005, vol. 3361, pp. 1–11.
- [7] —, "Novel techniques for time-compressing speech: An exploratory study," in *International Conference on Acoustics, Speech and Signal Processing 2005*, 2005.
- [8] B. Arons, "Techniques, perception, and applications of time-compressed speech," in *1992 Conference, American Voice I/O Society*, September 1992, pp. 169–177.
- [9] C. Hori and S. Furui, "A new approach to automatic speech summarization," *IEEE Multimedia*, vol. 5, no. 3, pp. 368–378, 2003.
- [10] K. McKeown, J. Hirschberg, M. Galley, and S. Maskey, "From text to speech summarisation," in *ICASSP*, 2005.
- [11] S. Maskey and J. Hirschberg, "Automatic summarisation of broadcast news using structural features," in *Eurospeech*, 2003.
- [12] M. Portnoff, "Time-scale modification of speech based on short-time fourier analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 29, no. 3, pp. 374–390, June 1981.
- [13] D. Hejna, "Real-time time-scale modification of speech via the synchronized overlap-add algorithm," Master's thesis, M.I.T., 1990.
- [14] D. Beasley and J. Maki, "Time and frequency altered speech," in *Contemporary Issues In Experimental Phonetics*, N. Lass, Ed. Academic Press, 1976, pp. 419–458.
- [15] T. Sticht, "Comprehension of repeated time-compression recordings," *Journal of Experimental Education*, vol. 37, no. 4, 1969.
- [16] J. Foote, G. Boreczky, and L. Wilcox, "An intelligent media browser using automatic multimodal analysis," in *ACM Multimedia*, September 1998, pp. 375–380.
- [17] M. Covell, M. Withgott, and M. Slaney, "Mach1 for nonuniform time-scale modification of speech," in *ICASSP*, 1998.
- [18] L. He and A. Gupta, "User benefits of non-linear time compression," Microsoft Research, Tech. Rep. MSR-TR-2000-96, September 2000.
- [19] H. Clark and D. Wilkes-Gibbs, "Referring as a collaborative process," *Cognition*, vol. 22, pp. 1–39, 1986.
- [20] L. Rabiner and M. Sambur, "Application of an LPC distance measure to the voiced-unvoiced-silence detection problem," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 25, no. 4, pp. 338–343, 8 1977.
- [21] K. Sparck Jones, "A statistical interpretation of term specificity and it's application in retrieval," *Journal of Documentation*, vol. 28, pp. 11–21, 1972.
- [22] M. Cooke, "A glimpsing model of speech perception in noise," *Journal of the Acoustical Society of America*, vol. 119, pp. 1562–1573, 2006.
- [23] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, M. F., D. Moore, P. Wellner, and H. Bourlard, "Modelling human interaction in meetings," in *ICASSP*, April 2003.
- [24] D. Moore, "The IDIAP smart meeting room," IDIAP, Tech. Rep. IDIAP-COM02-07, 2002.
- [25] G. Heiman, R. Leio, H. Leighbody, and B. K., "Word intelligibility decrements and the comprehension of time-compressed speech," *Perception and Psychophysics*, vol. 40, no. 6, pp. 407–411, 1986.

- [26] S. Vemuri, P. DeCamp, W. Bender, and C. Schmandt, "Improving speech playback using time-compression and speech recognition," in *Proceedings of CHI 2004*, April 2004, pp. 295–302.
- [27] 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 CHI 2002*, April 2002.
- [28] S. Whittaker, J. Hirschberg, and C. Nakatani, "Play it again: A study of the factors underlying speech browsing behaviours," in *Proceedings of CHI 1998*, 2002, pp. 275–282.
- [29] S. Tucker and S. Whittaker, "Time is of the essence: An evaluation of temporal compression algorithms," in *Proceedings of CHI 2006*, 2006.
- [30] S. Whittaker and B. Amento, "Semantic speech editing," in *Proceedings of CHI 2004*, 2004, pp. 527–534.

PLACE  
PHOTO  
HERE

**Simon Tucker** received the M.Eng. degree in software engineering in 2000 and the Ph.D. degree in computer science in 2003, both from the University of Sheffield, Sheffield, U.K. Since 2004 he has worked in the Department of Information Studies, University of Sheffield, as a Research Associate on the EU Augmented Multi-Party Interaction (AMI) project. He has research interests in automatic recognition and browsing of audio, machine learning and temporal compression of speech.

PLACE  
PHOTO  
HERE

**Steve Whittaker** has been Chair of Information Retrieval in the Information Studies Department at Sheffield University, since 2003. His research interests are in the theory, design and evaluation of communication and collaborative systems, as well as human computer interaction, multimedia access and retrieval. He has authored or co-authored over 90 refereed journal or conference papers (excluding workshop papers) and these have been cited over 2000 times. In the past he has researched, designed and built 10 novel systems supporting aspects of communication, personal information management and multimedia interfaces. He is holder of 10 US and UK patents, and is currently on the Editorial Board of the *Human Computer Interaction* and *Discourse Processes* journals and was Chair of ACM Computer Supported Cooperative Work 2000. He has worked for Hewlett Packard Labs (US and UK), Lotus, IBM Research, Bell Labs and AT&T Labs-Research. He has also been visiting faculty at Stanford and Edinburgh Universities.

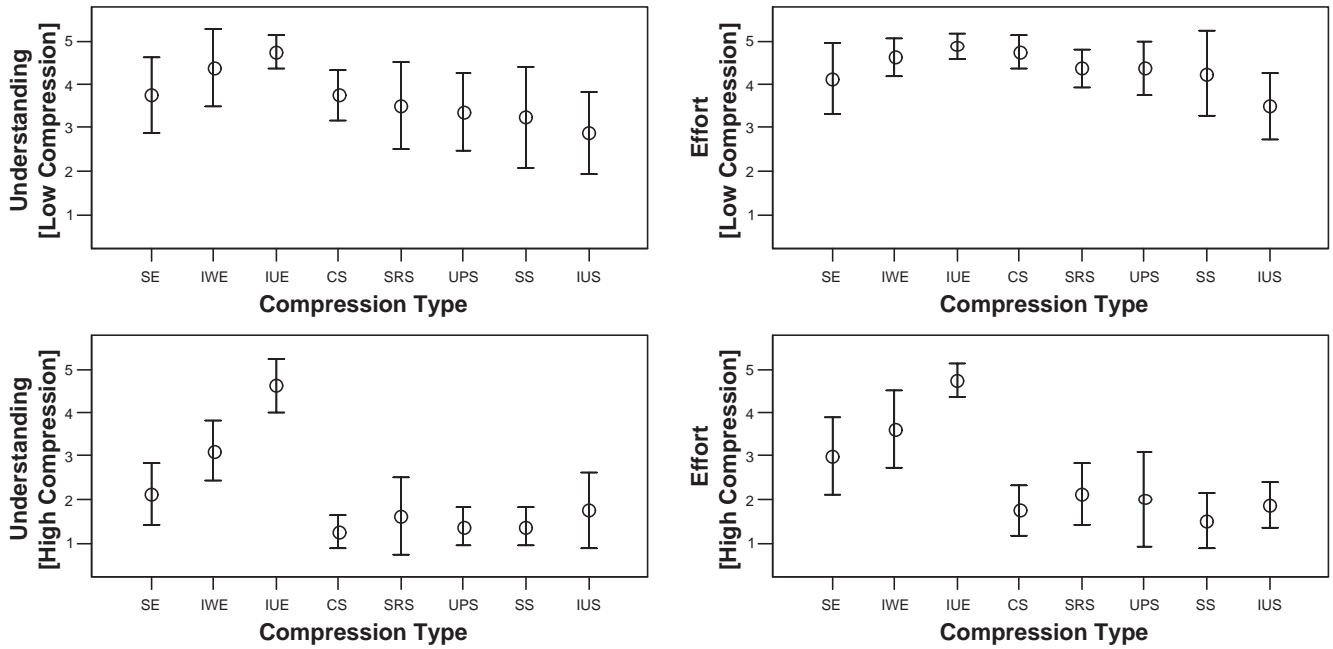


Fig. 1. Mean understanding and effort ratings (left and right columns respectively) organized by compression technique for low and high compression rates (top and bottom rows respectively - 95with the target statements. Key to abscissa labels, SE: Silence Excision, IWE: Insignificant Word Excision, IUE: Insignificant Utterance Excision, CS: Constant Speed-up, SRS: Speech Rate Speed-up, UPS: Utterance Position Speed-up, SS: Silence Speed-up, IUS: Insignificant Utterance Speed-up.

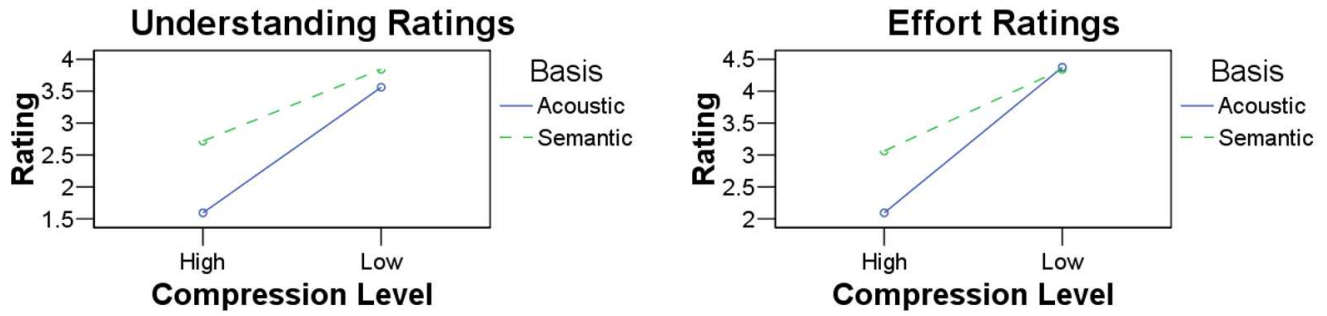


Fig. 2. Rating means for understanding and effort for comparisons between Compression Level and the Basis of the Technique

TABLE I  
CLASSIFICATION OF TEMPORAL COMPRESSION TECHNIQUES

		Compression Process		
		Excision	Speed-up	Hybrid
Basis of Technique	Acoustic	Silence Excision	Constant Speed-up Speech Rate Speed-up	Silence Speed-up
	Semantic	Insignificant Word Excision Insignificant Utterance Excision	Utterance Level Speed-up	Insignificant Utterance Speed-up