Dialogue Act Compression Via Pitch Contour Preservation

Gabriel Murray, Steve Renals

Centre for Speech Technology Research
University of Edinburgh, Edinburgh EH8 9LW
gabriel.murray@ed.ac.uk, s.renals@ed.ac.uk

10.21437/Interspeech.2006-434

Abstract

This paper explores the usefulness of prosody in automatically compressing dialogue acts from meeting speech. Specifically, this work attempts to compress utterances by preserving the pitch contour of the original whole utterance. Two methods of doing this are described in detail and are evaluated *subjectively* using human annotators and *objectively* using edit distance with a human-authored gold-standard. Both metrics show that such a prosodic approach is much better than the random baseline approach and significantly better than a simple text compression method.

Index Terms: speech compression, speech summarization, prosody, pitch contour.

1. Introduction

A common approach to automatic text and speech summarization is *extractive* summarization, in which sentences or utterances are extracted from the original document to form a summary. Recent research in speech summarizarization [1, 2] has indicated that the informativeness of extractive summaries is postively correlated with the length of the extracted units. Consequently, summarizer output tends to consist of fewer, but longer dialogue acts. If the compression rate for summarizing an hour-long meeting is quite low, e.g., 300 words, then few dialogue acts will be extracted. For that reason, it is very desirable to automatically compress these dialogue acts so that more can be extracted without exceeding the overall length limit.

The fragmented and disfluent nature of meeting speech means that implementing text compression techniques is not always feasible. Meeting dialogue acts cannot be reliably parsed, and we are thus limited as to how we can both determine the essential components of a given dialogue act and have a resulting compression that is readable. This paper explores the use of prosody in compressing informative dialogue acts from meeting speech. More specifically, the techniques described below compress the dialogue acts by trying to preserve the original pitch contour as much as possible in the compressed dialogue act. The simple intuition behind this method is that *prosody is meaning* [3] and that preserving this aspect of the prosody may preserve a great deal of the meaning as well.

Two methods of using prosody for speech compression are described below. They are first evaluated subjectively by humans grading on both informativeness and readability criteria, alongside human-authored gold-standards and random baseline compressions. The second evaluation metric is edit distance, objectively measuring the string distance between the automatic approaches and the gold-standards. In addition to the prosodic and random approaches, a simple text compression method was implemented and included for this edit distance evaluation.

2. Previous Work

The majority of previous speech compression work has been done on Broadcast News corpora [4, 5, 6, 7]. Because much of Broadcast News consists of read speech rather than spontaneous dialogue, it is often possible to reliably parse the data and use techniques inherited from textual sentence compression. In that respect, automatic compression of meeting utterances is a much more difficult task.

In work by Hori et. al. [4] and Kikuchi et. al. [5], a sentence compression method is described and results on English and Japanese broadcast news are given. The authors combine word confidence scores, word significance scores, trigram language scores, and word concatentation scores to determine the optimal compression of a given sentence using dynamic programming. The difference between the language score and the word concatentation score is that the former relies solely on trigram language probabilities while the latter is based on the dependency structure of the sentence.

Again on Broadcast News data, Kolluru et. al. [6] presents a multi-stage compaction method using a sequence of multi-layer perceptrons. First, confidence scores are used to remove incorrectly transcribed words. A chunk parser identifies intra-sentential chunks and a subset of the chunks are then chosen based on the presence of Named Entities and *tf.idf* scores.

Zechner [8] describes a speech summarization system in which false starts, repetitions and filled pauses are identified and removed, thereby increasing the coherency of the summaries and compressing the individual sentences. Zechner focuses especially on using part-of-speech tags, *trigger* words with high predictive potential, and turn boundary information in his work, and suggests that prosodic information could lead to further improvement in disfluency detection.

Ohtake et. al. [7] use prosodic features for speech-to-speech newscast compression and do not use ASR for compression at all. They locate accent phrase boundaries by analyzing fall-rise F0 patterns, determine which adjacent accent phrases belong together as single summary units, and then compare two prosodic methods for selecting the most important summary units. For example, summary units can be eliminated if their mean energy level falls below a pre-determined threshold or if a derived F0 summary unit score is above a speaker-dependent threshold. Ohtake et. al. also attempt to use prosodic features to determine whether a given summary unit depends on the preceding summary unit, so that when a summary unit is eliminated, its dependants are also eliminated.

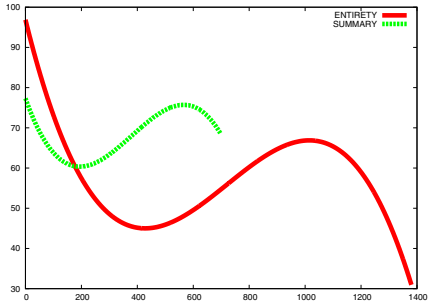

Figure 1: *Sample Dialogue Act and Summary Contours (first prosodic method)*

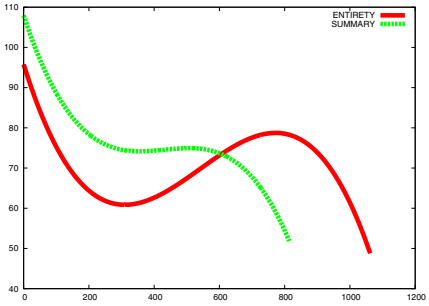

Figure 2: *Sample Dialogue Act and Summary Contours (second prosodic method)*

3. Compression Methods

This section presents the compression methods in detail. First, two prosodic methods are described, both of which attempt to compress the utterance by preserving the pitch contour. A simple textual method is presented, as well as a baseline compression method.

3.1. Prosodic Methods

The compression rate is between 0.65 and 0.70 for all of the automatic compression methods. A first step for each method is to remove simple filled pauses such as *uh* and *erm* as well as immediate repetitions of a word.

3.1.1. First Method

The first prosody method begins by breaking the utterance into prosodic phrases or chunks. The primary cue for phrase boundary is pause length, with pauses of 100 ms or more being considered a boundary. A secondary method is to look for instances of pitch reset which would signal the beginning of a new prosodic phrase. More specifically, we are looking for areas where the pitch falls to a low level for at least 300 ms before rising sharply again, with the fall-rise pattern signalling the pitch declination of one phrase and the beginning of another. We first attempt to locate the boundaries using only pause, as it is considered more reliable, but if we are unable to break the dialogue act into at least 3 chunks, we revert to looking at pitch reset as well.

Once the prosodic phrases are located, the overall pitch slope for each phrase is measured. We then begin an iterative process, wherein for each phrase we measure the pitch slopes of its constituent words and select the word whose slope is closest to that of the phrasal slope. If a phrase has no more than two words, we skip it altogether as it is likely to be a disfluent fragment. We continue the iterative process until the desired number of words has been selected for the compression.

Figure 1 shows cubic regressions for the pitch contours of the following utterance and summary pair:

Original: *So given these um these features or or these these examples um critical examples which they call support f- support vectors then um gien a new example if the new example falls um away from the boundary in one direction then it's classified as being a part of this particular class*

Compression: *So given these features or these examples critical examples which they call support vectors then given a new example if new example falls boundary in one direction then being a part of this particular class*

3.1.2. Second Method

The second method is more crude and does not depend on recognizing phrase boundaries. Instead, the pitch contour for the entire dialogue act is represented as a vector of F0 values. Compression proceeds by deleting words one at a time, based on how large an effect each word's deletion has on the pitch contour. For each iteration of the procedure, each word has its F0 values deleted from the pitch vector and replaced with interpolated values between its former neighbouring words. This new pitch vector is then compared with the original pitch vector by using cosine similarity. The word with the highest cosine similarity is deleted, as the removal of its F0 values had little effect on the overall pitch contour. Again, the procedure continues until the desired length is reached.

Essentially, the two prosodic methods are working from opposite directions, one iteratively selecting words while the other is iteratively eliminating words. There are significant differences, however, as the latter method does not use phrasal information and thus would not ignore short fragments as the former method would. This second method also relies on overall pitch vector similarity, which may not be as reliable as measuring slope at the phrasal and word levels.

Figure 2 shows cubic regressions for the pitch contours of the following utterance and summary pair:

Original: *And the interesting thing is that even though yes it's a digits task and that's a relatively small number of words and there's a bunch of digits that you train on it's just not as good as having a a l- very large amount of data and training up a a a nice good big HMM*

Compression: *And interesting thing is that though yes it's digits task and that's relatively small words and there's bunch digits you train on it's just not good as having a large amount and training up a nice good big HMM*

3.2. Simple Text Method

For the second evaluation scheme described below, we implemented a simple text compression method for comparison. As in the methods described above, we first delete filled pauses and repetitions. We then assign each word in the dialogue act a *tf.idf* score, a metric which gives high ranks to words that are frequent within a document but rare across multiple documents. We select the words with the highest *tf.idf* scores until the desired compres-

sion length is reached. This text compression method is quite simple but nevertheless would give a reasonable expectation of high informativeness.

3.3. Baseline

To assess baseline performance, we randomly select the desired number of words and present them in the original order.

3.4. Gold Standard

The gold standard for compression is human-authored compressions. Manual compressions were made with a compression rate between 60% and 70%. The manual compressions were restricted to using only words from the original dialogue act and had to be presented in the original order, as with the automatic methods. The slightly wider window for the compression rate is because it is not feasible to require human annotators to compress an utterance to a precise percentage of the original.

4. Experiments

Two methods of evaluation were carried out, the first being a subjective analysis using human annotators who rated each compression on two criteria, and the second being a measure of edit-distance to a gold-standard compression. The text compression method was not implemented until after the human evaluation was complete, and so it is only included in the edit-distance evaluation.

4.1. Data

The corpus used was the ICSI meeting corpus, consisting of 75 unrestricted meetings averaging about an hour in length each [9]. These experiments utilize manual dialogue act annotation [10], and thirty dialogue acts from the corpus were chosen which were output from the summarizer described in [2]. These dialogue acts average about 27 words in length. The content of the dialogue acts was quite technical, and though it would have been possible to select less technical and shorter dialogue acts, we are fundamentally concerned with how our compression method performs on actual summarizer output.

4.2. Subjective Evaluation

Five human judges were presented with the output of four compression methods on the test set, for a total of 120 compressions to be evaluated by each annotator. These four methods were random baseline compressions, human-authored gold-standard compressions, and the two prosodic compression methods. The judges were asked to rate each compression for two criteria, informativeness and readability. The ratings were made on a 1-5 Likert scale with 1 being 'Very Poor' and 5 being 'Very Good.'

4.2.1. Informativeness

When rating a given compression in terms of its informativeness, judges were asked to keep in mind whether the compression retained the most important parts of the original utterance and refrained from including irrelevant or unnecessary parts of the original. They were instructed that this is a distinct and separate rating from readability, so that a compression may score high on informativeness and still do very poorly on readability.

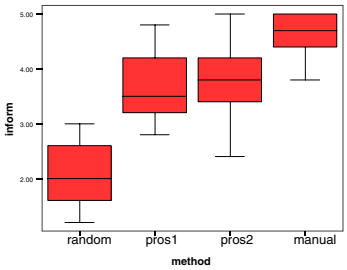

Figure 3: *Informativeness Scores for Four Compression Methods*

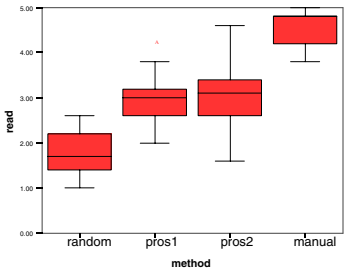

Figure 4: *Readability Scores for Four Compression Methods*

4.2.2. Readability

When rating a given compression in terms of its readability, judges were asked to consider whether the compression seemed grammatical and fluent relative to the original and whether the compression was generally readable. The term *relative* was included in the instructions because a compression which is an ungrammatical fragment should not be scored very low if the original utterance was also an ungrammatical fragment, for example.

4.3. Edit Distance

The second method of evaluation is edit distance, which utilizes our human-authored compressions as a gold-standard for an objective comparison. The edit distance between two strings is defined as $1 - (I + D + S)/R$, where R is the number of words in the reference string and I, D and S are insertions, deletions and substitutions, respectively. This metric thus objectively measures how close an automatically compressed string comes to the ideally compressed string. For this evaluation, four compression approaches were measured against the reference string, with the four approaches being random, text-based, and two prosodic approaches.

5. Results

5.1. Subjective

Figure 3 shows the averaged informativeness scores for the four compression methods. The inter-annotator agreement was very good, with the correlation of macroaveraged scores above 0.9 for each annotator pair. The manual compressions were rated signif-

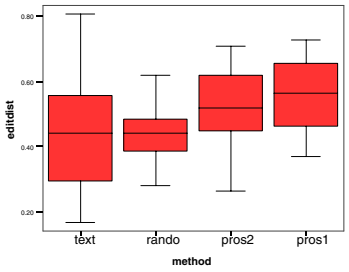

Figure 5: *Edit Distance for Four Compression Methods*

icantly higher than the others ($p<0.05$), with an average informativeness score of 4.65. Both of the prosodic approaches were significantly better than random ($p<0.05$) but were not significantly different from one another. The first prosodic approach had an average informativeness score of 3.69 and the second prosodic approach had an average of 3.82. The random compressions averaged 2.08 in terms of informativeness.

Figure 4 shows the averaged readability scores for the four compression methods. The inter-annotator agreement was again very good, with correlations above 0.9 for each annotator pair. The significant effects are the same as those of the informativeness scores, with the manual compressions rating significantly higher than the other approaches ($p<0.05$) and the prosodic approaches being signficantly better than random ($p<0.05$) but not significantly different from one another. The manual compressions had an average readability score of 4.6, the first prosodic approach averaged 2.93, the second prosodic approach averaged 3.15, and the random compressions averaged 1.77 in terms of readability. Interestingly, while the random and prosodic approaches had readability scores significantly lower than their informativeness scores, the manual compressions scored comparably on both readability and informativeness.

5.2. Edit Distance

Figure 5 shows the results of the edit distance metric, in which the manual gold-standard compressions were compared with the random and prosodic approaches, as well as a simple *tf.idf* approach. The most striking aspect of these results is that the *tf.idf* method performed only at the level of the random method. The prosodic approaches were significantly better ($p<0.05$), with an average edit distance of 0.56 and 0.53, respectively. The *tf.idf* and random approaches each had an average edit distance of 0.44.

6. Conclusion

This paper has presented a novel method of compressing utterances by preserving the pitch contour of the original within the compressed version. This compression method was meant to be robust to the disfluencies and ungrammaticalities of meeting speech, which prevent reliable parsing or dependency extraction, and the results are very encouraging. We report the findings of a pilot study evaluating two implementations of this approach. Based on both subjective and objective evaluation metrics, the prosodic approaches are far better than random compression. Objective evaluation using edit-distance also shows the prosodic methods outperforming a keyword-based compression approach. Relative to human-authored gold-standards, the readability of the prosodic compressions suffers but there are quite high levels of informativeness.

Though the second prosodic method was thought to be cruder than the first, it performed slightly but not significantly better in terms of both readability and informativeness. Future work may combine the two methods in order to optimize the compression results.

7. Acknowledgements

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-174).

8. References

[1] S. Maskey and J. Hirschberg, "Comparing lexial, acoustic/prosodic, discourse and structural features for speech summarization," in *Proc. of Interspeech 2005, Lisbon, Portugal*, September 2005.

[2] G. Murray, S. Renals, J. Moore, and J. Carletta, "Incorporating speaker and discourse features into speech summarization," in *Proc. of HLT-NAACL 2006, New York City, USA*, June 2006, p. to appear.

[3] M. Steedman, *The syntactic process*. Cambridge, MA, USA: MIT Press, 2000.

[4] C. Hori, S. Furui, R. Malkin, H. Yu, and A. Waibel, "Automatic speech summarization applied to english broadcast news speech," in *Proc. of the International Conference on Acoustics, Speech and Signal Processing, Orlando, USA*, 2002, pp. 9–12.

[5] T. Kikuchi, S. Furui, and C. Hori, "Two-stage automatic speech summarization by sentence extraction and compaction," in *Proc. of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, Japan*, 2003, pp. 207–210.

[6] B. Kolluru, Y. Gotoh, and H. Christensen, "Multi-stage compaction approach to broadcast news summarisation," in *Proc. of Interspeech 2005, Lisbon, Portugal*, September 2005.

[7] K. Ohtake, K. Yamamoto, Y. Toma, S. Sado, S. Masuyama, and S. Nakagawa, "Newscast speech summarization via sentence shortening based on prosodic features," in *Proc. of the ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, Japan*, April 2003,.

[8] K. Zechner, "Automatic summarization of spoken dialogues in unrestricted domains," Ph.D. dissertation, Carnegie Mellon University, 2001.

[9] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proc. of IEEE ICASSP 2003, Hong Kong, China*, April 2003.

[10] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, , and H. Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proc. of the 5th SIGdial Workshop on Discourse and Dialogue, Cambridge, MA, USA*, April-May 2004, pp. 97–100.