Incorporating Speaker and Discourse Features into Speech Summarization

Gabriel Murray, Steve Renals,
Jean Carletta, Johanna Moore

University of Edinburgh, School of Informatics
Edinburgh EH8 9LW, Scotland
gabriel.murray@ed.ac.uk, s.renals@ed.ac.uk,
jeanc@inf.ed.ac.uk, j.moore@ed.ac.uk

Abstract

We have explored the usefulness of incorporating speech and discourse features in an automatic speech summarization system applied to meeting recordings from the ICSI Meetings corpus. By analyzing speaker activity, turn-taking and discourse cues, we hypothesize that such a system can outperform solely text-based methods inherited from the field of text summarization. The summarization methods are described, two evaluation methods are applied and compared, and the results clearly show that utilizing such features is advantageous and efficient. Even simple methods relying on discourse cues and speaker activity can outperform text summarization approaches.

1. Introduction

The task of summarizing spontaneous spoken dialogue from meetings presents many challenges: information is sparse; speech is disfluent and fragmented; automatic speech recognition is imperfect. However, there are numerous speech-specific characteristics to be explored and taken advantage of. Previous research on summarizing speech has concentrated on utilizing prosodic features [1, 2]. We have examined the usefulness of additional speech-specific characteristics such as discourse cues, speaker activity, and listener feedback. This speech features approach is contrasted with a second summarization approach using only textual features—a centroid method [3] using a latent semantic representation of utterances. These individual approaches are compared to a combined approach as well as random baseline summaries.

This paper also introduces a new evaluation scheme for automatic summaries of meeting recordings, using a weighted precision score based on multiple human annotations of each meeting transcript. This evaluation scheme is described in detail below and is motivated by previous findings [4] suggesting that n-gram based metrics like ROUGE [5] do not correlate well in this domain.

2. Previous Work

In the field of speech summarization in general, research investigating speech-specific characteristics has focused largely on prosodic features such as F0 mean and standard deviation, pause information, syllable duration and energy. Koumpis and Renals [1] investigated prosodic features for summarizing voicemail messages in order to send voicemail summaries to mobile devices. Hori et al. [6] have developed an integrated speech summarization approach, based on finite state transducers, in which the recognition and summarization components are composed into a single finite state transducer, reporting results on a lecture summarization task. In the Broadcast News domain, Maskey and Hirschberg [7] found that the best summarization results utilized prosodic, lexical, and structural features, while Ohtake et al. [8] explored using *only* prosodic features for summarization. Maskey and Hirschberg similarly found that prosodic features alone resulted in good quality summaries of

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 367–374,
New York, June 2006.

Broadcast News.

In the meetings domain (using the ICSI corpus), Murray et al. [2] compared text summarization approaches with feature-based approaches using prosodic features, with human judges favoring the feature-based approaches. Zechner [9] investigated summarizing several genres of speech, including spontaneous meeting speech. Though relevance detection in his work relied largely on *tf.idf* scores, Zechner also explored cross-speaker information linking and question/answer detection, so that utterances could be extracted not only according to high *tf.idf* scores, but also if they were linked to other informative utterances.

Similarly, this work aims to detect important utterances that may not be detectable according to lexical features or prosodic prominence, but are nonetheless linked to high speaker activity, decision-making, or meeting structure.

3. Summarization Approaches

The following subsections give detailed descriptions of our two summarization systems, one of which focuses on speech and discourse features while the other utilizes text summarization techniques and latent semantic analysis.

3.1. Speech and Discourse Features

In previous summarization work on the ICSI corpus [2, 4], Murray et al. explored multiple ways of applying latent semantic analysis (LSA) to a term/document matrix of weighted term frequencies from a given meeting, a development of the method in [10]. A central insight to the present work is that additional features beyond simple term frequencies can be included in the matrix before singular value decomposition (SVD) is carried out. We can use SVD to project this matrix of features to a lower dimensionality space, subsequently applying the same methods as used in [2] for extracting sentences.

The features used in these experiments included features of speaker activity, discourse cues, listener feedback, simple keyword spotting, meeting location and dialogue act length (in words).

For each dialogue act, there are features indicating which speaker spoke the dialogue act and whether the same speaker spoke the preceding and succeeding dialogue acts. Another set of features indicates how many speakers are active on either side of a given dialogue act: specifically, how many speakers were active in the preceding and succeeding five dialogue acts. To further gauge speaker activity, we located areas of high speaker interaction and indicated whether or not a given dialogue act immediately preceded this region of activity, with the motivation being that informative utterances are often provocative in eliciting responses and interaction. Additionally, we included a feature indicating which speakers most often uttered dialogue acts that preceded high levels of speaker interaction, as one way of gauging speaker status in the meeting. Another feature relating to speaker activity gives each dialogue act a score according to how active the speaker is in the meeting as a whole, based on the intuition that the most active speakers will tend to utter the most important dialogue acts.

The features for discourse cues, listener feedback, and keyword spotting were deliberately superficial, all based simply on detecting informative words. The feature for discourse cues indicates the presence or absence of words such as *decide, discuss, conclude, agree*, and fragments such as *we should* indicating a planned course of action. Listener feedback was based on the presence or absence of positive feedback cues following a given dialogue act; these include responses such as *right, exactly* and *yeah*. Keyword spotting was based on frequent words minus stopwords, indicating the presence or absence of any of the top twenty non-stopword frequent words. The discourse cues of interest were derived from a manual corpus analysis rather than being automatically detected.

A structural feature scored dialogue acts according to their position in the meeting, with dialogue acts from the middle to later portion of the meeting scoring higher and dialogue acts at the beginning and very end scoring lower. This is a feature that is well-matched to the relatively unstructured ICSI meetings, as many meetings would be expected to have informative proposals and agendas at the beginning and perhaps summary statements and conclusions at the end.

Finally, we include a dialogue act length feature motivated by the fact that informative utterances will tend to be longer than others.

The extraction method follows [11] by ranking sentences using an LSA sentence score. The

matrix of features is decomposed as follows:

$$A = USV^T$$

where U is an $m \times n$ matrix of left-singular vectors, S is an $n \times n$ diagonal matrix of singular values, and V is the $n \times n$ matrix of right-singular vectors. Using sub-matrices S and V^T, the LSA sentence scores are obtained using:

$$Sc_i^{LSA} = \sqrt{\sum_{k=1}^{n} v(i,k)^2 * \sigma(k)^2},$$

where $v(i,k)$ is the kth element of the ith sentence vector and $\sigma(k)$ is the corresponding singular value.

Experiments on a development set of 55 ICSI meetings showed that reduction to between 5–15 dimension was optimal. These development experiments also showed that weighting some features slightly higher than others resulted in much improved results; specifically, the discourse cues and listener feedback cues were weighted slightly higher.

3.2. LSA Centroid

The second summarization method is a textual approach incorporating LSA into a centroid-based system [3]. The centroid is a pseudo-document representing the important aspects of the document as a whole; in the work of [3], this pseudo-document consists of keywords and their modified *tf.idf* scores. In the present research, we take a different approach to constructing the centroid and to representing sentences in the document. First, *tf.idf* scores are calculated for all words in the meeting. Using these scores, we find the top twenty keywords and choose these as the basis for our centroid. We then perform LSA on a very large corpus of Broadcast News and ICSI data, using the Infomap tool[1]. Infomap provides a query language with which we can retrieve word vectors for our twenty keywords, and the centroid is thus represented as the average of its constituent keyword vectors [12] [13].

Dialogue acts from the meetings are represented in much the same fashion. For each dialogue act, the vectors of its constituent words are retrieved, and the dialogue act as a whole is the average of its word vectors. Extraction then proceeds by finding the dialogue act with the highest cosine similarity with the centroid, adding this to the summary, then continuing until the desired summary length is reached.

[1]http://infomap.stanford.edu

3.3. Combined

The third summarization method is simply a combination of the first two. Each system produces a ranking and a master ranking is derived from these two rankings. The hypothesis is that the strength of one system will differ from the other and that the two will complement each other and produce a good overall ranking. The first system would be expected to locate areas of high activity, decision-making, and planning, while the second would locate information-rich utterances. This exemplifies one of the challenges of summarizing meeting recordings: namely, that utterances can be important in much different ways. A comprehensive system that relies on more than one idea of importance is ideal.

4. Experimental Setup

All summaries were 350 words in length, much shorter than the compression rate used in [2] (10% of dialogue acts). The ICSI meetings themselves average around 10,000 words in length. The reasons for choosing a shorter length for summaries are that shorter summaries are more likely to be useful to a user wanting to quickly overview and browse a meeting, they present a greater summarization challenge in that the summarizer must be more exact in pinpointing the important aspects of the meeting, and shorter summaries make it more feasible to enlist human evaluators to judge the numerous summaries on various criteria in the future.

Summaries were created on both manual transcripts and speech recognizer output. The unit of extraction for these summaries was the dialogue act, and these experiments used human segmented and labeled dialogue acts rather than try to detect them automatically. In future work, we intend to incorporate dialogue act detection and labeling as part of one complete automatic summarization system.

4.1. Corpus Description

The ICSI Meetings corpus consists of 75 meetings, lasting approximately one hour each. Our test set consists of six meetings, each with multiple human annotations. Annotators were given access to a graphical user interface (GUI) for browsing an individual meeting that included earlier human annotations: an orthographic transcription time-synchronized with the audio, and a topic segmentation based on a shallow hierarchical decomposition with keyword-based text labels describing each topic segment. The annotators were told to construct a textual summary of the meeting aimed at someone who is interested in the research being carried out, such as a researcher who does similar work elsewhere, using four headings:

- general abstract: "why are they meeting and what do they talk about?";
- decisions made by the group;
- progress and achievements;
- problems described

The annotators were given a 200 word limit for each heading, and told that there must be text for the general abstract, but that the other headings may have null annotations for some meetings. Annotators who were new to the data were encouraged to listen to a meeting straight through before beginning to author the summary.

Immediately after authoring a textual summary, annotators were asked to create an extractive summary, using a different GUI. This GUI showed both their textual summary and the orthographic transcription, without topic segmentation but with one line per dialogue act based on the pre-existing MRDA coding [14]. Annotators were told to extract dialogue acts that together would convey the information in the textual summary, and could be used to support the correctness of that summary. They were given no specific instructions about the number or percentage of acts to extract or about redundant dialogue acts. For each dialogue act extracted, they were then required in a second pass to choose the sentences from the textual summary supported by the dialogue act, creating a many-to-many mapping between the recording and the textual summary. Although the expectation was that each extracted dialogue act and each summary sentence would be linked to something in the opposing resource, we told the annotators that under some circumstances dialogue acts and summary sentences could stand alone.

We created summaries using both manual transcripts as well as automatic speech recognition (ASR) output. The AMI-ASR system [15] is described in more detail in [4] and the average word error rate (WER) for the corpus is 29.5%.

4.2. Evaluation Frameworks

The many-to-many mapping of dialogue acts to summary sentences described in the previous section allows us to evaluate our extractive summaries according to how often each annotator linked a given extracted dialogue act to a summary sentence. This is somewhat analogous to Pyramid weighting [16], but with dialogue acts as the SCUs. In fact, we can calculate weighted precision, recall and f-score using these annotations, but because the summaries created are so short, we focus on weighted precision as our central metric. For each dialogue act that the summarizer extracts, we count the number of times that each annotator links that dialogue act to a summary sentence. For a given dialogue act, it may be that one annotator links it 0 times, one annotator links it 1 time, and the third annotator links it two times, resulting in an average score of 1 for that dialogue act. The scores for all of the summary dialogue acts can be calculated and averaged to create an overall summary score.

ROUGE scores, based on n-gram overlap between human abstracts and automatic extracts, were also calculated for comparison [5]. ROUGE-2, based on bigram overlap, is considered the most stable as far as correlating with human judgments, and this was therefore our ROUGE metric of interest. ROUGE-SU4, which evaluates bigrams with intervening material between the two elements of the bigram, has recently been shown in the context of the Document Understanding Conference (DUC)[2] to bring no significant additional information as compared with ROUGE-2. Results from [4] and from DUC 2005 also show that ROUGE does not always correlate well with human judgments. It is therefore included in this research in the hope of further determining how reliable the

[2] http://duc.nist.gov

ROUGE metric is for our domain of meeting summarization.

5. Results

The experimental results are shown in figure 1 (weighted precision) and figure 2 (ROUGE-2) and are discussed below.

5.1. Weighted Precision Results

For weighted precision, the speech features approach was easily the best and scored significantly better than the centroid and random approaches (ANOVA,$p<0.05$), attaining an averaged weighted precision of 0.52. The combined approach did not improve upon the speech features approach but was not significantly worse either. The randomly created summaries scored much lower than all three systems.

The superior performance of the speech features approach compared to the LSA centroid method closely mirrors results on the ICSI development set, where the centroid method scored 0.23 and the speech features approach scored 0.42. For the speech features approach on the test set, the best feature by far was dialogue act length. Removing this feature resulted in the precision score being nearly halved. This mirrors results from Maskey and Hirschberg [7], who found that the length of a sentence in seconds and its length in words were the two best features for predicting summary sentences. Both the simple keyword spotting and the discourse cue detection features caused a lesser decline in precision when removed, while other features of speaker activity had a negligible impact on the test results.

Interestingly, the weighted precision scores on ASR were not significantly worse for any of the summarization approaches. In fact, the centroid approach scored very slightly higher on ASR output than on manual transcripts. In [17] and [2] it was similarly found that summarizing with ASR output did not cause great deterioration in the quality of the summaries. It is not especially surprising that the speech features approach performed similarly on both manual and ASR transcripts, as many of its features based on speaker exchanges and speaker activity would be unaffected by ASR errors. The speech features approach is still significantly better than the random and centroid summaries, and is not significantly better than the combined approach on ASR.

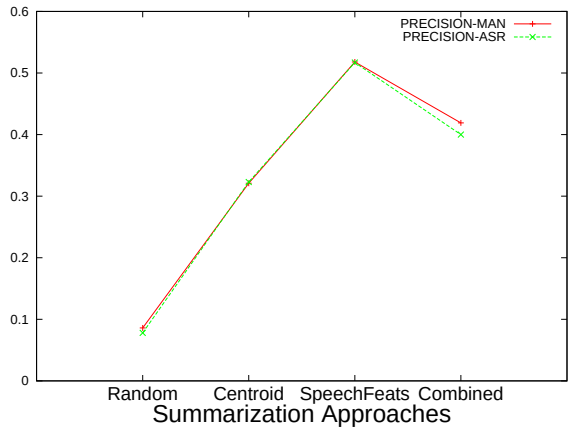

Figure 1: *Weighted Precision Results on Test Set*

5.2. ROUGE Results

The ROUGE results greatly differed from the weighted precision results in several ways. First, the centroid method was considered to be the best, with a ROUGE-2 score of 0.047 compared with 0.041 for the speech features approach. Second, there were not as great of differences between the four systems according to ROUGE as there were according to weighted precision. In fact, the random summaries of manual transcripts are not significantly worse than the other approaches, according to ROUGE-2. Neither the combined approach nor the speech features approach is significantly worse than the centroid system, with the combined approach generally scoring on par with the centroid scores.

The third difference relates to summarization on ASR output. ROUGE-2 has the random system and the combined system showing sharp declines when applied to ASR transcripts. The speech features and centroid approaches do not show declines. Random summaries are significantly worse than both the centroid summaries ($p<0.1$) and speech features summaries ($p<0.05$). Though the combined approach declines on ASR output, it is not significantly worse than the other systems.

To get an idea of a ROUGE-2 upper bound, for each meeting in the test set we left one human abstract out and compared it with the remaining abstracts. The result was an average ROUGE-2 score of .086.

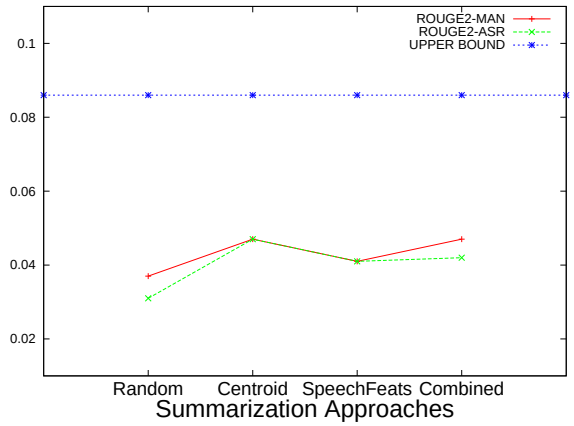

Figure 2: *ROUGE-2 Results on Test Set*

ROUGE-1 and ROUGE-SU4 show no significant differences between the centroid and speech features approaches.

5.3. Correlations

There is no significant correlation between macroaveraged ROUGE and weighted precision scores across the meeting set, on both ASR and manual transcripts. The Pearson correlation is 0.562 with a significance of $p < 0.147$. The Spearman correlation is 0.282 with a significance of $p < 0.498$. The correlation of scores across each test meeting is worse yet, with a Pearson correlation of 0.185 ($p<0.208$) and a Spearman correlation of 0.181 ($p<0.271$).

5.4. Sample Summary

The following is the text of a summary of meeting Bed004 using the speech features approach:

-so its possible that we could do something like a summary node of some sort that

-and then the question would be if if those are the things that you care about uh can you make a relatively compact way of getting from the various inputs to the things you care about

-this is sort of th the second version and i i i look at this maybe just as a you know a a whatever uml diagram or you know as just a uh screen shot not really as a bayes net as john johno said

-and um this is about as much as we can do if we dont w if we want to avoid uh uh a huge combinatorial explosion where we specify ok if its this and this but that is not the case and so forth it just gets really really messy

-also it strikes me that we we m may want to approach the point where we can sort of try to find a uh a specification for some interface here that um takes the normal m three l looks at it

-so what youre trying to get out of this deep co cognitive linguistics is the fact that w if you know about source source paths and goals and nnn all this sort of stuff that a lot of this is the same for different tasks

-what youd really like of course is the same thing youd always like which is that you have um a kind of intermediate representation which looks the same o over a bunch of inputs and a bunch of outputs

-and pushing it one step further when you get to construction grammar and stuff what youd like to be able to do is say you have this parser which is much fancier than the parser that comes with uh smartkom

-in independent of whether it about what is this or where is it or something that you could tell from the construction you could pull out deep semantic information which youre gonna use in a general way

6. Discussion

Though the speech features approach was considered the best system, it is unclear why the combined approach did not yield improvement. One possibility relates to the extreme brevity of the summaries: because the summaries are only 350 words in length, it is possible to have two summaries of the same meeting which are equally good but completely non-overlapping in content. In other words, they both extract informative dialogue acts, but not the same ones. Combining the rankings of two such systems might create a third system which is comparable but not any better than either of the first two systems alone. However, it is still possible that the combined system will be better in terms of balancing the two types of importance discussed above: utterances that contain a lot of informative content and keywords and utterances that relate to decision-making and meeting structure.

ROUGE did not correlate well with the weighted precision scores, a result that adds to the previous evidence that this metric may not be reliable in the domain of meeting summarization.

It is very encouraging that the summarization approaches in general seem immune to the WER of the ASR output. This confirms previous findings such as [17] and [2], and the speech and structural features used herein are particularly unaffected by a moderately high WER. The reason for the random summarizaton system not suffering

a sharp decline when applied to ASR may be due to the fact that its scores were already so low that it couldn't deteriorate any further.

7. Future Work

The above results show that even a relatively small set of speech, discourse, and structural features can outperform a text summarization approach on this data, and there are many additional features to be explored. Of particular interest to us are features relating to speaker status, i.e. features that help us determine who is leading the meeting and who it is that others are deferring to. We would also like to more closely investigate the relationship between areas of high speaker activity and informative utterances.

In the immediate future, we will incorporate these features into a machine-learning framework, building support vector models trained on the extracted and non-extracted classes of the training set.

Finally, we will apply these methods to the AMI corpus [18] and create summaries of comparable length for that meeting set. There are likely to be differences regarding usefulness of certain features due to the ICSI meetings being relatively unstructured and informal and the AMI hub meetings being more structured with a higher information density.

8. Conclusion

The results presented above show that using features related to speaker activity, listener feedback, discourse cues and dialogue act length can outperform the lexical methods of text summarization approaches. More specifically, the fact that there are multiple types of important utterances requires that we use multiple methods of detecting importance. Lexical methods and prosodic features are not necessarily going to detect utterances that are relevant to agreement, decision-making or speaker activity. This research also provides further evidence that ROUGE does not correlate well with human judgments in this domain. Finally, it has been demonstrated that high WER for ASR output does not significantly decrease summarization quality.

9. Acknowledgements

Thanks to Thomas Hain and the AMI-ASR group for speech recognition output. This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multiparty Interaction, FP6-506811, publication AMI-150).

10. References

[1] K. Koumpis and S. Renals, "Automatic summarization of voicemail messages using lexical and prosodic features," *ACM Transactions on Speech and Language Processing*, vol. 2, pp. 1–24, 2005.

[2] G. Murray, S. Renals, and J. Carletta, "Extractive summarization of meeting recordings," in *Proceedings of the 9th European Conference on Speech Communication and Technology, Lisbon, Portugal*, September 2005.

[3] D. Radev, S. Blair-Goldensohn, and Z. Zhang, "Experiments in single and multidocument summarization using mead," in *The Proceedings of the First Document Understanding Conference, New Orleans, LA*, September 2001.

[4] G. Murray, S. Renals, J. Carletta, and J. Moore, "Evaluating automatic summaries of meeting recordings," in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Workshop on Machine Translation and Summarization Evaluation (MTSE), Ann Arbor, MI, USA*, June 2005.

[5] C.-Y. Lin and E. H. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *Proceedings of HLT-NAACL 2003, Edmonton, Calgary, Canada*, May 2003.

[6] T. Hori, C. Hori, and Y. Minami, "Speech summarization using weighted finite-state transducers," in *Proceedings of the 8th European Conference on Speech Communication and Technology, Geneva, Switzerland*, September 2003.

[7] S. Maskey and J. Hirschberg, "Comparing lexial, acoustic/prosodic, discourse and structural features for speech summarization," in *Proceedings of the 9th European Conference on Speech Communication and Technology, Lisbon, Portugal*, September 2005.

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

[9] K. Zechner, "Automatic summarization of open-domain multiparty dialogues in diverse genres," *Computational Linguistics*, vol. 28, no. 4, pp. 447–485, 2002.

[10] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, Louisiana, USA*, September 2001, pp. 19–25.

[11] J. Steinberger and K. Ježek, "Using latent semantic analysis in text summarization and summary evaluation," in *Proceedings of ISIM 2004, Roznov pod Radhostem, Czech Republic*, April 2004, pp. 93–100.

[12] P. Foltz, W. Kintsch, and T. Landauer, "The measurement of textual coherence with latent semantic analysis," *Discourse Processes*, vol. 25, 1998.

[13] B. Hachey, G. Murray, and D. Reitter, "The embra system at duc 2005: Query-oriented multi-document summarization with a very large latent semantic space," in *Proceedings of the Document Understanding Conference (DUC) 2005, Vancouver, BC, Canada*, October 2005.

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

[15] T. Hain, J. Dines, G. Garau, M. Karafiat, D. Moore, V. Wan, R. Ordelman, I.Mc.Cowan, J.Vepa, and S.Renals, "An investigation into transcription of conference room meetings," *Proceedings of the 9th European Conference on Speech Communication and Technology, Lisbon, Portugal*, September 2005.

[16] A. Nenkova and B. Passonneau, "Evaluating content selection in summarization: The pyramid method," in *Proceedings of HLT-NAACL 2004, Boston, MA, USA*, May 2004.

[17] R. Valenza, T. Robinson, M. Hickey, and R. Tucker, "Summarization of spoken audio through information extraction," in *Proceedings of the ESCA Workshop on Accessing Information in Spoken Audio, Cambridge UK*, April 1999, pp. 111–116.

[18] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Proceedings of MLMI 2005, Edinburgh, UK*, June 2005.