# Meta Comments for Summarizing Meeting Speech

Gabriel Murray and Steve Renals. Meta comments for summarizing meeting speech. In Machine Learning for Multimodal Interaction (Proc. MLMI '08), number 5237 in Lecture Notes in Computer Science, pages 236-247. Springer, 2008.

# Meta Comments for Summarizing Meeting Speech

Gabriel Murray[1] and Steve Renals[2]

[1] University of British Columbia, Vancouver, Canada
gabrielm@cs.ubc.ca
[2] University of Edinburgh, Edinburgh, Scotland
s.renals@ed.ac.uk

**Abstract.** This paper is about the extractive summarization of meeting speech, using the ICSI and AMI corpora. In the first set of experiments we use prosodic, lexical, structural and speaker-related features to select the most informative dialogue acts from each meeting, with the hypothesis being that such a rich mixture of features will yield the best results. In the second part, we present an approach in which the identification of "meta-comments" is used to create more informative summaries that provide an increased level of abstraction. We find that the inclusion of these meta comments improves summarization performance according to several evaluation metrics.

## 1 Introduction

Speech summarization has attracted increasing interest in the past few years. There has been a variety of work concerned with the summarization of broadcast news [19, 14, 8, 3], voicemail messages [11], lectures [9, 21] and spontaneous conversations [18, 22]. In this paper we are concerned with the summarization of multiparty meetings. Small group meetings provide a compelling setting for spoken language processing, since they feature considerable interaction (up to 30% of utterances are overlapped), and informal conversational speech. Previous work in the summarization of meeting speech [20, 16, 6] has been largely based on the extraction of informative sentences or dialogue acts (DAs) from the source transcript. The extracted portions are then concatenated to form a summary of the meeting, with informativeness gauged by various lexical and prosodic criteria, among others.

In this work we first present a set of experiments that aim to identify the most useful features for the detection of informative DAs in multiparty meetings. We have applied this extractive summarization framework to the ICSI and AMI meeting corpora, described below. Extractive summaries of multiparty meetings often lack coherence, and may not be judged to be particularly informative by a user. In the second part of the paper, we aim to produce summaries with a greater degree of abstraction through the automatic extraction of "meta" DAs: DAs in which the speaker refers to the meeting itself. Through the inclusion of such DAs in our summaries, we hypothesize that the summaries will be more coherent and more obviously informative to an end user. Much as human abstracts tend to be created in a high-level fashion from a third-party perspective, we aim to automatically create extracts with similar attributes, harnessing the self-referential quality of meeting speech. Using an expanded feature set, we report results on the AMI corpus and compare with our previously generated extractive summaries.

## 2 Experimental Setup

We have used the the AMI and ICSI meeting corpora. The AMI corpus [1] consists of about 100 hours of recorded and annotated meetings, divided into *scenario* and *non-scenario* meetings. In the scenario portion, groups of four participants take part in a series of four meetings and play roles within a fictitious company. While the scenario given to them is artificial, the speech and the actions are completely spontaneous and natural. There are 138 meetings of this type in total. The length of an individual meeting ranges from 15 to 45 minutes, depending on which meeting in the series it is and how quickly the group is working. For these experiments, we use only the scenario meetings from the AMI corpus.

The second corpus used herein is the ICSI meeting corpus [10], a corpus of 75 naturally occurring meetings of research groups, approximately one hour each in length. Unlike the AMI scenario meetings and similar to the AMI non-scenario meetings, there are varying numbers of participants across meetings in the ICSI corpus, ranging from three to ten, with an average of six participants per meeting.

Both corpora feature a mixture of native and non-native English speakers and have been transcribed both manually and using automatic speech recognition(ASR) [7]. The resultant word error rates were 29.5% for the ICSI corpus, and 38.9% for the AMI corpus.

### 2.1 Summary Annotation

For both the AMI and ICSI corpora, annotators were asked to write abstractive summaries of each meeting and to extract the DAs in the meeting that best conveyed or supported the information in the abstractive summary. A many-to-many mapping between transcript DAs and sentences from the human abstract was obtained for each annotator. It is also possible for a DA to be extractive but unlinked. The human-authored abstracts each contain a general abstract summary and three subsections for "decisions," "actions" and "problems" from the meeting.

Kappa values were used to measure inter-annotator agreement. The ICSI test set has a lower kappa value (0.35) compared with the AMI test set (0.48), reflecting the difficulty in summarizing the much less structured (and more technical) ICSI meetings.

### 2.2 Summary Evaluation

To evaluate automatically produced extractive summaries we have extended the weighted precision measure [17] to weighted precision, recall and F-measure. This evaluation scheme relies on the multiple human annotated summary links described in the previous section. Both weighted precision and recall share the same numerator

$$num = \sum_{i=1}^{M} \sum_{j=1}^{N} L(s_i, a_j)$$

where $L(s_i, a_j)$ is the number of links for a DA $s_i$ in the machine extractive summary according to annotator $a_i$, $M$ is the number of DAs in the machine summary, and $N$ is

the number of annotators. Weighted precision is defined as:

$$precision = \frac{num}{N \cdot M}$$

and weighted recall is given by

$$recall = \frac{num}{\sum_{i=1}^{O} \sum_{j=1}^{N} L(s_i, a_j)}$$

where $O$ is the total number of DAs in the meeting, $N$ is the number of annotators, and the denominator represents the total number of links made between DAs and abstract sentences by all annotators. The weighted F-measure is calculated as the harmonic mean of weighted precision and recall.

We have also used the ROUGE evaluation framework [13] for the second set of experiments, in particular ROUGE-2 and ROUGE-SU4. We believe that ROUGE is particularly relevant for evaluation in that case, as we are trying to create extracts that are more abstract-like, and ROUGE compares machine summaries to gold-standard human abstracts.

## 3 Features for Meeting Summarization

In this section we outline the features and classifiers used for extractive summarization of meetings, presenting results using the AMI and ICSI corpora.

Table 1 lists and briefly describes the set of the features used. The prosodic features consist of energy, F0, pause, duration and a rate-of-speech measure. We calculate both the duration of the complete DA, as well as of the uninterrupted portion. The structural features include the DA's position in the meeting and position within the speaker's turn (which may contain multiple DAs). There are two measures of speaker dominance: the dominance of the speaker in terms of meeting DAs and in terms of total speaking time. There are two term-weighting metrics, *tf.idf* and *su.idf*, the former favoring words that are frequent in the given document but rare across all documents, and the latter favoring words that are used with varying frequency by the different speakers [15]. The prosodic and term-weight features are calculated at the word level and averaged over the DA. In these experiments we employed a manual DA segmentation, although automatic approaches are available [5].

For each corpus, a logistic regression classifier is trained on the seen data as follows, using the *liblinear* toolkit[3]. Feature subset selection is carried out using a method based on the $f$ statistic:

$$F(i) = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{D^{(+)} + D^{(-)}}$$

$$D^{(\pm)} = \frac{1}{n_\pm - 1} \sum_{k=1}^{n_\pm} (x_{k,i}^{(\pm)} - \bar{x}_i^{(\pm)})^2$$

[3] `http://www.csie.ntu.edu.tw/~cjlin/liblinear/`

| Feature ID | Description |
|---|---|
| **Prosodic Features** | |
| **ENMN** | mean energy |
| **F0MN** | mean F0 |
| **ENMX** | max energy |
| **F0MX** | max F0 |
| **F0SD** | F0 stdev. |
| **PPAU** | precedent pause |
| **SPAU** | subsequent pause |
| **ROS** | rate of speech |
| **Structural Features** | |
| **MPOS** | meeting position |
| **TPOS** | turn position |
| **Speaker Features** | |
| **DOMD** | speaker dominance (DAs) |
| **DOMT** | speaker dominance (seconds) |
| **Length Features** | |
| **DDUR** | DA duration |
| **UINT** | uninterrupted length |
| **WCNT** | number of words |
| **Lexical Features** | |
| **SUI** | su.idf sum |
| **TFI** | tf.idf sum |
| **ACUE** (experiment 2) | abstractive cuewords |
| **FPAU** (experiment 2) | filled pauses |

**Table 1.** Features Key

where $n_+$ and $n_-$ are the number of positive instances and negative instances, respectively, $\bar{x}_i$, $\bar{x}_i^{(+)}$, and $\bar{x}_i^{(-)}$ are the means of the $i$th feature for the whole, positive and negative data instances, respectively, and $x_{k,i}^{(+)}$ and $x_{k,i}^{(-)}$ are the $i$th features of the $k$th positive and negative instances [2]. The $f$ statistic for each feature was first calculated, and then feature subsets of size $n = 3, 5, 7, 9, 11, 13, 15, 17$ were tried, with the $n$ best features included at each step based on the $f$ statistic. The feature subset size with the highest balanced accuracy during cross-validation was selected as the feature set for training the logistic regression model.

The classifier was then run on the unseen test data, and the class probabilities were used to rank the candidate DAs for each meeting and create extracts of 700 words. This length was chosen so that the summaries would be short enough to be read by a time-constrained user, much as a short human abstract might be quickly consulted, but long enough to index the most important points of the meeting. This short summary length also necessitates a high level of precision since we extract relatively few DAs.

### 3.1 AMI Results

For the AMI data the best feature subset according to the feature selection method includes all 17 features, for both manual and ASR transcriptions. For both transcription types, the best five features (in order) were DA word count, *su.idf* score, DA duration, uninterrupted length of the DA, and *tf.idf* score. Figure 1 shows the histograms of the feature $f$ statistics using both the manual and ASR transcriptions.

We calculated the ROC curves and areas under the curve (AUROC) for the classifiers that identified the extractive DAs, using both manual and ASR transcriptions. For

the manual transcripts AUROC = 0.855, for the ASR transcripts AUROC= 0.850, with chance level classification at 0.5.

Figure 3 illustrates the weighted F-measures for the 700-word summaries on manual and ASR transcripts using the feature-based approach. There is no significant difference between the manual and ASR F-measures according to paired t-test, and the ASR scores are on average slightly higher.

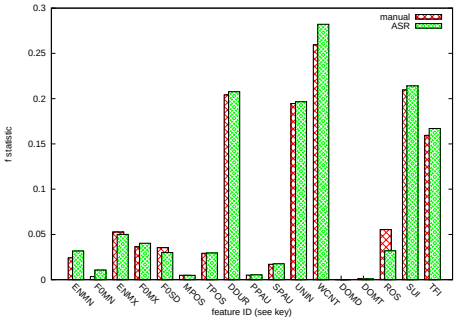



**Fig. 1.** $f$ statistics for AMI database features

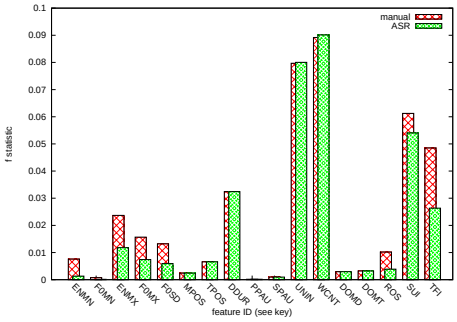



**Fig. 2.** $f$ statistics for ICSI database features

### 3.2 ICSI Results

For the ICSI corpus using manual transcripts, the optimal feature subset consisted of 15 features according to balanced accuracy, excluding mean F0 and precedent pause. The best 5 features according to the $f$ statistic were DA word count, uninterrupted length, *su.idf* score, *tf.idf* score and DA duration. The optimal subset for ASR transcripts consisted of the same 15 features. Figure 2 shows the histograms for the feature $f$ statistics using both the manual and ASR databases.

We calculated the ROC and AUROC for each classifier applied to the 6 test set meetings. For manual transcripts AUROC = 0.818, and for ASR transcripts AUROC = 0.824.

Figure 3 shows the weighted F-measures for the 700-word summaries for both manual and ASR transcripts. As with the AMI corpus, there is no significant difference between manual and ASR results and the ASR average is again slightly higher.

### 3.3 Discussion

In this first experiment we have shown that a rich mixture of features yields good results, based on feature subset selection with the $f$ statistic. We have also compared the AMI and ICSI corpora in terms of feature selection. For both corpora, summarization is slightly better on ASR than on manual transcripts, in terms of weighted F-measure. It is worth pointing out, however, that the weighted F-measure only evaluates whether the correct DAs have been extracted and does not penalize misrecognized words within an

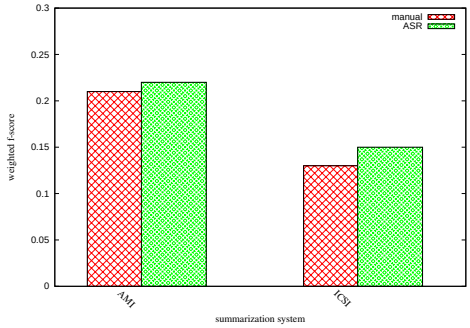



**Fig. 3.** Weighted F-Measures for AMI and ICSI Corpora, Manual and ASR Transcripts

extracted DA. Such ASR errors create a problem for textual summaries, but are less important for multimodal summaries (e.g. those produced by concatenating audio and/or video segments).

In the next section we provide a more detailed analysis of the effectiveness of various feature subsets for an altered summarization task.

## 4 Meta Comments in Meeting Speech

In the second experiment we aim to improve our results through the identification of meta DAs to be included in machine summaries. These are DAs in which the speaker refers to the meeting itself. We first describe scheme we used to annotate meta DAs, then present an expanded feature set, and compare summarization results with the first experiment.

The AMI corpus contains *reflexivity* annotations: a DA is considered to be reflexive if it refers to the meeting or discussion itself. Reflexive DAs are related to the idea of meta comments, but the reflexivity annotation alone is not sufficient. Many of the DAs deemed to be reflexive consist of statements like “Next slide, please.” and “Can I ask a question?” in addition to many short feedback statements such as “Yeah” and “Okay.” Although such DAs do indeed refer to the flow of discussion at a high level, they are not particularly informative. We are not interested in identifying DAs that are *purely* about the flow of discussion, but rather we would like to detect those DAs that refer to low-level issues in a high-level way. For example, we would find the DA “We decided on a red remote control” more interesting than the DA “Let’s move on”.

In light of these considerations, we created an annotation scheme for meta DAs, that combined several existing annotations in order to form a new binary meta/non-meta annotation for the corpus. The ideal condition would be to consider DAs as meta only if they are labelled as both extractive and reflexive. However, there are relatively few such DAs in each meeting. For that reason, we also consider DAs to be meta if they are linked to the “decisions,” “actions” or “problems” subsections of the abstract. The intuition

behind using the DA links to those three abstract subsections is that areas of a discussion that relate to these categories will tend to indicate where the discussion moves from a lower level to a higher level. For example, the group might discuss technical issues in some detail and then make a decision regarding those issues, or set out a course of action for the next meetings.

For this second experiment, we trained the classifier to extract only these newly-labelled meta DAs rather than all generally extract-worthy DAs as in the first experiment. We analyze which individual features and feature subsets are most effective for this novel extraction task. We then evaluate our brief summaries using weighted F-measure and ROUGE and make an explicit comparison with the previously generated summaries. This work focuses solely on the AMI data, for two reasons: the ICSI data does not contain the reflexivity annotation, and the ICSI abstracts have slightly different subsections than the AMI abstracts.

### 4.1 Filled Pause and Cueword Features

In these experiments we have two additional lexical features to the feature set used in the previous section, which we hypothesise to be relevant to the meta DA identification task. The first new feature is the number of filled pauses in each DA. This is included because the fluency of speech might change at areas of conversational transition, perhaps including more filled pauses than on average. These filled pauses consist of terms such as "uh", "um", "erm", "mm," and "hmm."

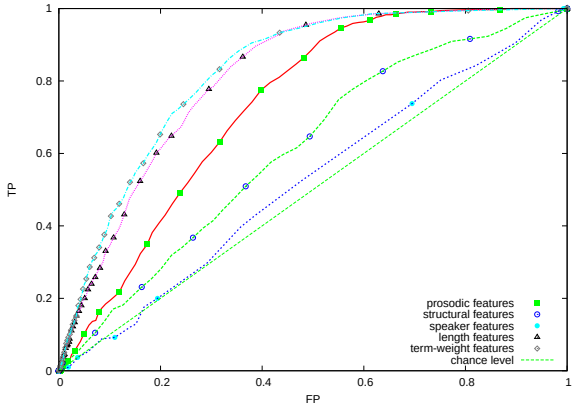
The second new feature is the presence of abstractive or meta cuewords, as automatically derived from the training data. Since we are trying to create summaries that are somehow more abstract-like, we examine terms that occur often in the abstracts of meetings but less often in the *extracts* of meetings. We score each word according to the ratio of these two frequencies,

$$TF(t,j)/TF(t,k)$$

where $TF(t,j)$ is the frequency of term $t$ in the set of abstracts $j$ from the training set meetings and $TF(t,k)$ is the frequency of term $t$ in the set of extracts $k$ from the training set meetings. These scores are used to rank the words from most abstractive to least abstractive, and we keep the top 50 words as our list of meta cuewords. The top 5 abstractive cuewords are "team", "group", "specialist", "member", and "manager." For both the manual and ASR feature databases, each DA then has a feature indicating how many of these high-level terms it contains.

### 4.2 Evaluation of Meta DA Extraction

We evaluated the resulting 700-word summaries using three metrics: weighted F-measures using the new extractive labels, weighted F-measures using the old extractive labels, and ROUGE. For the second of those evaluations, it is not expected that the summaries derived from meta DAs will fare as well as using the original extractive summaries, since the vast majority of previously extractive DAs are now considered members of the negative class and the evaluation metric is based on the previous extractive/non-extractive labels; the results are included out of interest nonetheless.




| Fea. Subset | AUROC |
|---|---|
| **Prosodic** | 0.734 |
| **Structural** | 0.611 |
| **Speaker** | 0.524 |
| **Length** | 0.811 |
| **Term-Weight** | 0.826 |

**Fig. 4.** AUROC Values, Manual Transcripts

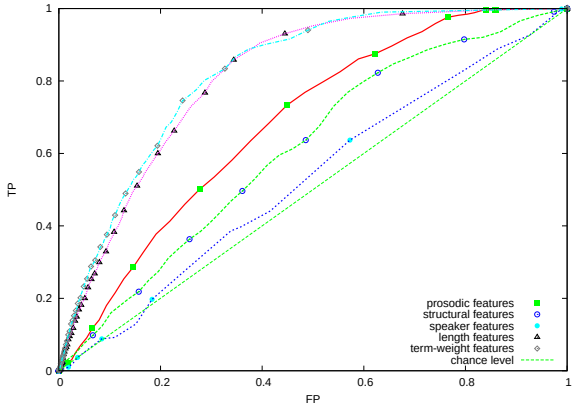
We experimented using the AMI corpus. With manual transcripts, the feature subset that was selected consisted of 13 features, which excluded mean F0, position in the speaker's turn, precedent pause, both dominance features, and filled pauses. The best five features in order were *su.idf*, DA word-count, *tf.idf*, DA duration, and uninterrupted duration. In the case of ASR transcription, all 19 features were selected and the best five features were the same as for the manual transcripts.

We calculated the ROC and AUROC for the meta DA classifiers applied to the 20 test set meetings using both manual and ASR transcription. For manual, AUROC = 0.843 and for ASR, AUROC = 0.842. This result is very encouraging, as it shows that it is possible to discriminate the meta DAs from other DAs (including some marked as extractive). Given that we created a new positive class based on a DA satisfying one of four criteria, and that we consider everything else as negative, this result shows that DAs that meet at least one of these extraction criteria do have characteristics in common with one another and can be discerned as a separate group from the remainder.

### 4.3 Feature Analysis

The previous sections have reported a brief features analysis according to each feature's $f$ statistic for the extractive/non-extractive classes. This section expands upon that by examining how useful different subsets of features are for classification on their own. While we found that the optimal subset according to automatic feature subset selection is 13 and 19 features for manual and ASR, respectively, it is still interesting to examine performance using only certain classes of features on this data. We therefore divide the features into five categories of **prosodic** features, **length** features, **speaker** features, **structural** features and **lexical** features. Note that we do not consider DA duration to be a prosodic feature.

Figure 4 shows the ROC curves and AUROC values for each feature subset for the manual transcriptions. We find that no individual subset matches the classification performance found by using the entire feature set, but that several classes exhibit credible




| Fea. Subset | AUROC |
|---|---|
| **Prosodic** | 0.684 |
| **Structural** | 0.612 |
| **Speaker** | 0.527 |
| **Length** | 0.812 |
| **Term-Weight** | 0.822 |

**Fig. 5.** AUROC Values, ASR Transcripts

individual performance. The length and term-weight features are clearly the best, but we find that prosodic features alone perform better than structural or speaker features.

Figure 5 shows the ROC curves and AUROC values for each feature subset for the ASR transcriptions. The trend is largely the same as above: no individual feature type is better than the combination of feature types. The principal difference is that prosodic features alone are worse on ASR, likely due to extracting prosodic features aligned to erroneous word boundaries, while term-weight features are about the same as on manual.

### 4.4 Summary Evaluation

Figure 6 presents the weighted F-measures using the novel extractive labelling, for the new meta summaries as well as for the summaries created and evaluated in the first experiment. For manual transcripts, the new summaries outperform the old summaries with an average F-measure of 0.17 versus 0.12. The reason for the scores overall being lower than the F-measures reported in the previous chapter using the original formulation of weighted precision/recall/F-measure is that there are now far fewer positive instances in each meeting since we are restricting the positive class to the "meta" subset of informative DAs. The meta summaries are significantly better than the previous summaries on this evaluation according to paired t-test ($p<0.05$).

For ASR, we find both the new meta summaries and older non-meta summaries performing slightly better than on manual transcripts according to this evaluation. The meta summaries again are rated higher than the non-meta summaries, with an average F-measure of 0.19 versus 0.14 and are significantly better according to paired t-test ($p<0.05$).

We would expect the new meta extractive summaries to perform better in terms of weighted F-measure with respect to the new extractive labelling, since the classifiers were trained in a consistent manner. However, when using the old extractive labelling

the weighted F-measures for these new summaries are also slightly higher than the F-measures reported in the previous section. The F-measure for manual transcripts is 0.23 compared with 0.21 previously, and 0.24 for ASR compared with 0.22 earlier. This is a surprising and encouraging result, that our new annotation and subsequent "meta" DA extraction experiments have led not only to finding areas of high-level meta comments in the meetings but also to improved general summary informativeness. Kappa statistics also suggest that it is easier for annotators to agree on DAs that meet these specific meta criteria ($\kappa$=0.45) than DAs that simply support the general abstract portion of the human summary ($\kappa$=0.40).

We also evaluate the meta summaries using the ROUGE-2 and ROUGE-SU4 metrics [13], which have previously been found to correlate well with human judgements in the DUC summarization tasks [12, 4]. We calculate precision, recall and F-measures for each, and ROUGE is run using the parameters utilized in the DUC conferences, plus removal of stopwords.

Again the meta summaries outperform the summaries created in the first experiments. For ROUGE-2, using manual transcripts, the meta summaries average a score of 0.039, compared with 0.033 for the previous non-meta summaries.On the ASR transcripts, the meta summaries scored slightly higher with an average of 0.041 compared with 0.032 for the non-meta summaries, which is significant at $p<0.05$. According to ROUGE-SU4, on manual transcripts the meta summaries outperform the low-level summaries with an average of 0.066 compared with 0.061, respectively. On ASR transcripts, the meta summaries average 0.069 compared with 0.064 for the low-level summaries. Both differences are significant at $p<0.05$. Figure 7 shows the ROUGE-SU4 scores for meta and non-meta summaries compared with human extracts of the same length.

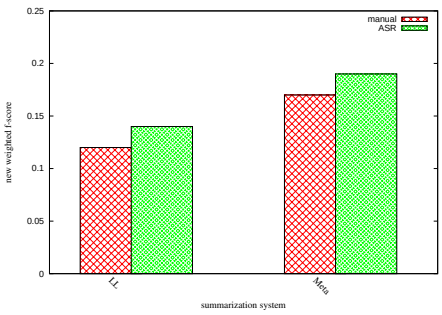



**Fig. 6.** New Weighted F-measures

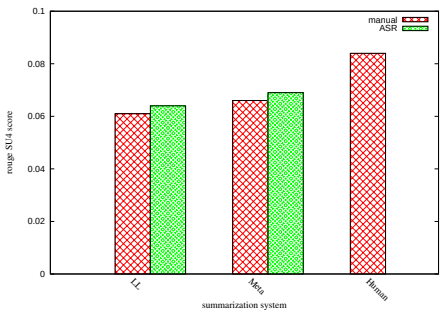



**Fig. 7.** ROUGE-SU4 Scores

**LL**=low-level summaries from first experiment, **Meta**=novel meta summaries

The following two DAs from meeting TS3003c are examples of DAs that are extracted for the meta summary but not for the previously generated non-meta summary of the same meeting.

– **speaker A**: so the industrial designer and user interface designer are going to work together on this one

– **speaker D**: i heard our industrial designer talk about flat, single- and double-curved.

### 4.5 Discussion

According to multiple intrinsic evaluations, our novel meta summaries are superior to the previously generated summaries. We believe that the criteria for *informativeness* are more meaningful, that the output is more flexible, and that these high-level summaries would be more coherent from the perspective of a third-party end user.

Of the two novel feature types in the expanded features database, abstractive cue-words are found to be very good indicators of meta DAs, while the presence of filled pauses is much less useful. It may be the case that the presence of filled pauses would be a helpful feature for a general extraction task but is simply not indicative of meta DAs.

There are interesting possibilities for new directions with this research. For example, by training on individual classes one could create a complex extractive summary that first lists DAs relating to decisions, followed by DAs that identify action items for the following meeting. A hierarchical summary could also be created, with high-level DAs at the top, linked to related lower-level DAs that might provide more detail. It is also possible that these meta summary DAs would lend themselves to further interpretation and generation of automatic abstracts.

## 5 Conclusion

The aim of this work has been two-fold: to help move the state-of-the-art in speech summarization further along the extractive-abstractive continuum, and to determine the most effective feature subsets for the summarization task. We have shown that informative meta DAs can be reliably identified, and have described the effectiveness of various feature sets in performing this task. While the work has been firmly in the extractive paradigm, it has moved beyond previously used simplistic notions of "informative" versus "uninformative" in order to create more informative and high-level summary output.

*Acknowledgements* This work is supported by the European IST Programme Project AMIDA (FP6-0033812). Thanks to the AMI-ASR team for providing the ASR.

## References

1. 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 *Proc. of MLMI 2005, Edinburgh, UK*, pages 28–39, 2005.
2. Y.-W. Chen and C.-J. Lin. Combining SVMs with various feature selection strategies. In I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, editors, *Feature extraction, foundations and applications*. Springer, 2006.

3. H. Christensen, Y. Gotoh, and S. Renals. A cascaded broadcast news highlighter. *IEEE Transactions on Audio, Speech and Language Processing*, 16:151–161, 2008.
4. H. Dang. Overview of duc 2005. In *Proc. of the Document Understanding Conference (DUC) 2005, Vancouver, BC, Canada*, 2005.
5. A. Dielmann and S. Renals. DBN based joint dialogue act recognition of multiparty meetings. In *Proc. of ICASSP 2007, Honolulu, USA*, pages 133–136, 2007.
6. M. Galley. A skip-chain conditional random field for ranking meeting utterances by importance. In *Proc. of EMNLP 2006, Sydney, Australia*, pages 364–372, 2006.
7. T. Hain, L. Burget, J. Dines, G. Garau, V. Wan, M. Karafiat, J. Vepa, and M. Lincoln. The AMI system for transcription of speech in meetings. In *Proc. of ICASSP 2007,*, pages 357–360, 2007.
8. C. Hori and S. Furui. Speech summarization: An approach through word extraction and a method for evaluation. *IEICE Transactions on Information and Systems*, E87-D(1):15–25, 2004.
9. T. Hori, C. Hori, and Y. Minami. Speech summarization using weighted finite-state transducers. In *Proc. of Interspeech 2003, Geneva, Switzerland*, pages 2817–2820, 2003.
10. 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*, pages 364–367, 2003.
11. K. Koumpis and S. Renals. Automatic summarization of voicemail messages using lexical and prosodic features. *ACM Transactions on Speech and Language Processing*, 2:1–24, 2005.
12. C.-Y. Lin. Looking for a few good metrics: Automatic summarization evaluation - how many samples are enough. In *Proc. of NTCIR 2004, Tokyo, Japan*, pages 1765–1776, 2004.
13. C.-Y. Lin and E. H. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proc. of HLT-NAACL 2003, Edmonton, Calgary, Canada*, pages 71–78, 2003.
14. S. Maskey and J. Hirschberg. Comparing lexial, acoustic/prosodic, discourse and structural features for speech summarization. In *Proc. of Interspeech 2005, Lisbon, Portugal*, pages 621–624, 2005.
15. G. Murray and S. Renals. Term-weighting for summarization of multi-party spoken dialogues. In *Proc. of MLMI 2007, Brno, Czech Republic*, pages 155–166, 2007.
16. G. Murray, S. Renals, and J. Carletta. Extractive summarization of meeting recordings. In *Proc. of Interspeech 2005, Lisbon, Portugal*, pages 593–596, 2005.
17. G. Murray, S. Renals, J. Moore, and J. Carletta. Incorporating speaker and discourse features into speech summarization. In *Proc. of the HLT-NAACL 2006, New York City, USA*, pages 367–374, 2006.
18. N. Reithinger, M. Kipp, R. Engel, and J. Alexandersson. Summarizing multilingual spoken negotiation dialogues. In *Proc. of ACL 2000, Hong Kong*, pages 310–317, Morristown, NJ, USA, 2000. Association for Computational Linguistics.
19. R. Valenza, T. Robinson, M. Hickey, and R. Tucker. Summarization of spoken audio through information extraction. In *Proc. of the ESCA Workshop on Accessing Information in Spoken Audio, Cambridge UK*, pages 111–116, 1999.
20. K. Zechner. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485, 2002.
21. J. Zhang, H. Chan, P. Fung, and L. Cao. Comparative study on speech summarization of broadcast news and lecture speech. In *Proc. of Interspeech 2007, Antwerp, Belgium*, pages 2781–2784, 2007.
22. X. Zhu and G. Penn. Summarization of spontaneous conversations. In *Proc. of Interspeech 2006, Pittsburgh, USA*, pages 1531–1534, 2006.