# Exploring Speaker Characteristics for Meeting Summarization

*Fei Liu, Yang Liu*

Computer Science Department
The University of Texas at Dallas, Richardson, TX, USA
feiliu,yangl@hlt.utdallas.edu

## Abstract

In this paper, we investigate using meeting-specific characteristics to improve extractive meeting summarization, in particular, speaker-related attributes (such as verboseness, gender, native language, role in the meeting). A rich set of speaker-sensitive features are developed in the supervised learning framework. We perform experiments on the ICSI meeting corpus. Results are evaluated using multiple criteria, including ROUGE, a sentence-level F-measure, and an approximated Pyramid approach. We show that incorporating speaker characteristics can consistently improve summarization performance on various testing conditions.

## 1. Introduction

Automatic meeting summarization provides an efficient way of indexing meeting archives and has received much attention recently. Meetings are different from traditional written text in many ways, making summarization a more challenging task for this domain. Meeting transcripts lack structural information, such as title, sentence boundaries, or paragraph. Sentences in meetings are often poorly structured with interruptions and disfluencies. Multiple participants in meetings introduce an interleaved discourse structure. High recognition error rate degrades many lexical, syntactic, and semantic analysis techniques.

Researchers have investigated some speech-specific characteristics for speech summarization. For example, [1] proposed to use re-occurring acoustic patterns in speech to estimate utterance similarity, hence identify salient utterances without using transcribed text; [2, 3, 4, 5, 6, 7] combined lexical, structural, and prosodic information (such as pitch, duration, energy, and pause) in the supervised framework for different speech domains, including meetings, broadcast news, and lectures; [4, 8] incorporated discourse cue words, listener feedback, and speaker activity related features in their meeting summarization system; [9] investigated the hierarchical structure in lecture speech and developed a rhetorical state HMM for summarization; they also showed that speaker normalized acoustic features are highly effective for lecture summarization. [10, 11] made use of different representations of speech recognition (ASR) output and confidence scores to improve summarization performance for ASR condition.

In this study, we explore speaker dependent characteristics (such as verboseness, gender, native language, role in the meeting) to improve extractive meeting summarization performance. Meetings are typically multi-party conversations. Speakers differ in their speaking styles and lexical usage. In addition, a speaker' role in a topic discussion has an impact on his/her speaking style. Different from most text domains where a document is generally written by one person, each participant in the meeting can begin a new topic when starting his/her turn. Hence we expect that leveraging the speaker characteristics would be beneficial for summarization. We first perform an analysis of various normalization methods that are motivated to capture speaker characteristics, and then integrate speaker information in a supervised summarization framework. Experiments are performed on the ICSI meeting corpus using both human transcripts and ASR output. Summarization performance is evaluated using ROUGE, Pyramid, and sentence-level F-measures. We observe consistent improvements using our proposed approaches, on both human transcripts and ASR output, and using different evaluation metrics.

## 2. Meeting Corpus

We use the ICSI meeting corpus, which consists of 75 naturally occurring meetings, each about an hour long [12]. All the meetings have been orthographically transcribed and annotated with dialogue acts (DAs), corresponding speakers, topic boundaries, and extractive summaries [3, 13]. The ASR output we used is obtained from a state-of-the-art SRI system [14], with a word error rate (WER) of about 38.2% in the entire corpus. The DA boundaries for ASR output are obtained by aligning the human annotated DA boundaries to the ASR words based on time information. In total, there are about 110K DA units annotated for the corpus. Six meetings (same as in [3, 4]) from this corpus are used for testing to make our work comparable to the state-of-the-art results. 20 meetings are randomly selected as development set, and the rest 48 meetings[1] are used for training our supervised system. Each test meeting has three human reference summaries, while there is only one reference summary for each of the development and training meetings. We use the TreeTagger[2] to lemmatize both human transcripts and ASR output, and the TnT part-of-speech tagger [15] trained from Switchboard data for tagging.

## 3. Summarization Approaches

The task we investigate in this paper is extractive meeting summarization, where important DA units[3] in the transcripts are selected to form a summary according to a predefined summarization ratio. We represent extractive summarization as a binary classification problem and use supervised approaches for this task. A maximum entropy classifier[4] is trained using the annotated data and assigns posterior probabilities for each DA during testing. The higher ranked DAs are selected in summary. We make use of length, structural, and similarity related cues

[1] Meeting Bed002 was dropped due to poor transcription quality.

[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[3] When there is no ambiguity, we will just call these units 'DAs' in the rest of the paper.

[4] http://www.cs.utah.edu/~hal/megam/

that have been proved to be very effective and form competitive baselines for this task [3, 16]. First we perform an in-depth analysis of the discriminative power of features normalized differently based on speaker attributes and speaking style, and then evaluate the incorporation of speaker-sensitive features in the supervised system.

In a preprocessing step, we first eliminate some DA candidates in the original transcripts that are not likely to be summary DAs. We construct a list of stopwords consisting of 250 and 200 words respectively for human transcripts and ASR output. These words have the lowest IDF values (inverse DA frequency). DAs that only consist of stopwords and functional words are then filtered out. This step has several benefits: (1) remove unlikely summary candidates: we found on the training set that this filtering process removes 56.36% of the non-summary DA candidates, while only 9.15% of the summary DAs are removed; (2) improve the balance between the two classes for model training: the ratio between summary and non-summary DAs is reduced from the original 1:14.45 to 1:6.94; (3) better form a speaker turn: the average number of turns in each meeting decreases dramatically, from 833.65 to 264.10. This shows that the prevalent existence of backchannels (e.g., "uh-huh") tends to break the conversation discourse into small pieces, hence we believe a filtering process can keep the original speaker turn and will be more useful for our study of speaker turn related features.

### 3.1. Basic Features in Supervised Summarization

- Length and location features (Len + Loc (mt)):
  (A) utterance length, measured by number of words, or seconds (2 features).
  (B) location of the utterance, represented using the portion of utterances before and after the current DA. This can be measured by number of words, DAs, or seconds (6 features).
  We normalize the above base features to $[0, 1]$ by dividing each of them by the maximum value obtained at the meeting scale: length features are normalized using the longest DA in the meeting; location features are divided by all the utterances in the meeting. We refer to this meeting scale normalization as "mt".
- Similarity features (Cos-sim):
  This is the cosine similarity between a DA ($S_i$) and the entire meeting ($S_j$) under the vector space model:

$$Sim(S_i, S_j) = \frac{\sum_k w_{i,k} \times w_{j,k}}{\sqrt{\sum_k w_{i,k}^2} \times \sqrt{\sum_k w_{j,k}^2}}$$

  The term weight for a word $w_{i,k}$ is determined by $\sqrt{TF} \times IDF$, where $TF$ is its term frequency in text segment $S_i$, and $IDF$ is the inverse document frequency. We use only content words and non-stopwords to form the word vectors.

### 3.2. Accounting for Speaker Characteristics

Meetings often have multiple participants. They alternatively present their ideas or thoughts, resulting in an interleaved discourse. We expect a thorough analysis of the speakers' speaking style would help us better understand whether speaker related characteristics have an impact on summarization and how we can effectively utilize speaker information for meeting summarization. Based on data examination, we came up with the following speculative hypotheses that we expect would affect summarization systems.

- Hyp1: Sentence length information may be adjusted based on speakers for summarization. Utterance length is likely to be affected by a speaker's style (verbose or not) or the person's knowledge in a topic discussion.
- Hyp2: Sentence location features can also take into consideration speaker information. Important issues in a meeting might be brought up by different speakers, but for each individual speaker, s/he tends to pose the important DAs at the beginning and conclude at the end of all of his/her utterances within the entire meeting, or with respect to a specific turn of this speaker.

In order to test our hypotheses and investigate if speaker specific information can indeed help summarization, we introduce two different normalization methods: (1) speaker meeting-level normalization (denoted by "spkr"): location features are calculated based on all utterances from the same speaker in the meeting; for each of the length, location, and similarity related features, we divide its feature value by the maximum obtainable feature value among all the utterances by this speaker; (2) speaker turn-level normalization (denoted by "turn"): location features are calculated based on utterances from the current speaker turn; each feature value is divided by the maximum available value within this turn. This normalization aims to model the speaker's behavior within a particular speaker turn, thus it is a more local normalization, whereas the first method considers more global behavior from the entire meeting.

We first evaluate the discriminating power of the above normalized features on the training set using two criteria: (1) "AbsDiff": For each feature, this is the absolute difference between the average value for summary and non-summary DAs. A larger "AbsDiff" value therefore corresponds to better class separability of this feature; (2) Fisher's discriminant ratio $C$: For a specific feature variable, Fisher's ratio $C$ across two classes $i$ (summary DAs) and $j$ (non-summary DAs) is defined as:

$$C = \frac{(\mu_i - \mu_j)^2}{\sigma_i^2 + \sigma_j^2} \quad (1)$$

where $\mu_i$, $\mu_j$ and $\sigma_i^2$, $\sigma_j^2$ are means and variances of classes $i$ and $j$. Larger Fisher's ratio means stronger discriminating power of this feature.

Table 1 shows the results on the training set using these two metrics for all of the features, normalized based on the meeting, speaker, and speaker turns (denoted as "mt", "spkr", "turn" respectively). For length, location, and similarity related features, we observed an improved discriminating power after performing speaker level normalization. For location features, using speaker-turn level normalization also increases their ability to separate the summary DAs from the non-summary DAs. The speaker normalized length features promote the DAs from unverbose speakers; while the normalized location features emphasize on the relative location within the utterances of the same speaker or within the current speaker turn. These results verified our hypotheses above.

The last hypothesis we have is that speaker biographic attributes might affect meeting summarization. The factors we consider are male or female speakers, native or non-native speakers, and different roles (professors or non-professors in our corpus). We performed an analysis to examine the difference of the summary DA percentage corresponding to these factors, in order to see if these factors affect the chance of an utterance being selected into the summary. Table 2 shows the results

| Normalization | | AbsDiff | | | Fisher Ratio | | |
|---|---|---|---|---|---|---|---|
| | | mt | spkr | turn | mt | spkr | turn |
| Len | second | .082 | **.112** | .067 | .142 | **.153** | .022 |
| | word | .085 | **.113** | .068 | .131 | **.150** | .025 |
| Loc (begin) | second | .054 | **.063** | **.061** | .018 | **.023** | **.020** |
| | DA | .052 | **.059** | .039 | .017 | **.021** | .008 |
| | word | .054 | **.062** | **.059** | .018 | **.023** | **.019** |
| Loc (end) | second | .054 | **.058** | **.071** | .017 | **.020** | **.025** |
| | DA | .052 | **.058** | **.085** | .017 | **.021** | **.039** |
| | word | .054 | **.058** | **.073** | .018 | **.020** | **.027** |
| Sim | Cos-sim | .080 | **.090** | .043 | .097 | **.099** | .014 |

Table 1: Fisher ratio and AbsDiff of average scores between summary and non-summary DAs for different features, using three different normalization methods.

of the summary percentage difference using different measurements: words, DAs, and seconds. For example, suppose we use the number of DAs as the measurement and consider gender factor. Among all the utterances by male speakers, 13.0% of the DAs are included in the summary, while 11.5% of the DAs from female speakers are included in the summary. Their difference is 1.5%, which is presented in the table corresponding to the "Male/Female" row and "DA" column. We observe that overall the difference with respect to these factors is rather small. The fact that the faculty role shows little effect might be because that the participants in the ICSI meeting corpus included many non-faculty researchers. This finding may be corpus dependent. The gender information yields slightly larger difference in summary utterance selection than other attributes. Using the number of DAs as the measurement results in larger difference compared to using words or seconds.

| Difference | | second | DA | word |
|---|---|---|---|---|
| Speaker Attributes | Male/Female | 1.2 | **1.5** | 1.3 |
| | Native/Non-native | 0.1 | **0.5** | 0.1 |
| | Faculty/Non-faculty | 0.1 | **0.2** | 0.1 |

Table 2: Difference of summary percentage (in %) for different speaker attributes, measured using seconds, DAs, and words.

# 4. Experiments

## 4.1. Evaluation Metrics

We evaluate meeting summarization performance using the following three metrics:

- ROUGE [17]. This has been widely used in prior studies on meeting summarization task. ROUGE scores measure the n-gram overlap between the system summary and a set of human reference summaries. We use ROUGE-1 and ROUGE-2 F-scores to make our results comparable to other previous research.
- DA F-score. This one compares the system extracted summary DAs to human annotated ones, and calculates the DA-level precision and recall scores. The DA-level F-measure is then calculated as the harmonic mean of the precision and recall values with equal weights.
- Pyramid approach [18]. Following [4], we use a location-restricted Pyramid score to measure summarization performance. In this approach, a summary content unit (SCU) is defined as a word and its location in the document (index of DA), thus the same word appearing in different locations are discriminated. For example, ("voice", 16) and ("voice", 27) are considered as two different SCUs. The score of each SCU is defined as the total number of times it appears in the human reference summaries. For each system generated summary, we compute a score $D$ by adding up all its SCU scores. A maximum score $D^*$ is calculated as the maximum obtainable SCU scores given the summary length constraint. The Pyramid score is defined as $P = D/D^*$.

## 4.2. Results on Development Set

In this experiment, we compare different normalization methods, at the meeting, speaker, or speaker turn level, and also investigate whether combining features with different normalization levels can result in further improvement. The features are length, location, and similarity related features, as described in Section 3.1. All these experiments are performed on the 20-meeting development set. Based on the results in Table 1, for length and similarity features, we compare the "mt" and "spkr" level normalization and their combination ("turn" level normalization is not included due to its lower discriminating power). For location features, we use the three normalization methods by themselves, and various combination of them. Table 3 shows the ROUGE-1 results using 15% compression ratio, with best results for each feature category shown in bold.

| | Normalization | Human | ASR |
|---|---|---|---|
| Len | mt | 69.11 | **65.56** |
| | spkr | 69.32 | 65.15 |
| | mt + spkr | **69.47** | 65.36 |
| Len + Loc | mt | 69.67 | 65.86 |
| | spkr | 69.73 | 65.77 |
| | turn | 70.17 | 65.78 |
| | mt + spkr | 69.52 | 65.96 |
| | mt + turn | 70.33 | 66.25 |
| | spkr + turn | **70.42** | 66.22 |
| | mt + spkr + turn | 70.32 | **66.25** |
| Len + Sim | mt | 69.18 | 65.53 |
| | spkr | **69.19** | **65.59** |
| | mt + spkr | 69.02 | 65.52 |

Table 3: Summarization results (ROUGE-1 F-measure) on development set, using features normalized at the meeting, speaker, or speaker turn level, or various combination.

For length related features, combining both "mt" and "spkr" level normalization achieves the best performance on human transcripts, while "mt" only normalization results in best performance on ASR output. We then use these two setups as base features for human and ASR transcripts respectively, and add location or similarity features with different normalization levels. For location features, on human transcripts, combining "spkr" and "turn" level normalization works best, and "turn" normalization has the best performance by itself; for ASR output, utilizing all the three levels of normalization or "mt+turn" yields the best results. For similarity features, the best performance is from "spkr" normalization on both human transcripts and ASR output. Combining with "mt" does not yield additional gain. We also experimented with adding binary features indicating whether the DA is from male/female, native/non-native, or faculty/non-faculty speakers, but did not observe performance gain over the base features.

## 4.3. Results on Test Set

Table 4 shows the results on the test set using different metrics: ROUGE-1, ROUGE-2, Pyramid, and the DA-level F-measure.

| | | R-1 F-score | | | | R-2 F-score | | | | | Pyramid | | DA F-Score | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Word Ratio | | 14% | 15% | 16% | 17% | 21% | 22% | 23% | 24% | 25% | 16% | 23% | 16% | 23% |
| Human | Oracle | 76.39 | 76.95 | 77.10 | 76.89 | 60.98 | 61.54 | 62.03 | 62.45 | 62.80 | 85.87 | 97.41 | 72.13 | 89.62 |
| | Supervised | 70.38 | 70.59 | 70.54 | 70.25 | 43.82 | 44.25 | 44.31 | 44.12 | 44.06 | 55.70 | 59.53 | 28.55 | 35.93 |
| | + Spkr-norm | 71.20 | 71.58 | **71.81** | 71.56 | 45.21 | 45.09 | **45.53** | 45.12 | 45.14 | **57.94** | **62.04** | **29.80** | **38.71** |
| | Other work | 71.23 in [7], 71.5 in [4] | | | | 39.37 in [7], 44.2 in [4] | | | | | 55.4 in [4] | | - | |
| ASR | Oracle | 70.79 | 71.11 | 71.07 | 70.74 | 43.63 | 43.93 | 44.18 | 44.37 | 44.53 | 86.80 | 98.04 | 74.09 | 91.50 |
| | Supervised | 66.87 | 66.97 | 66.91 | 66.72 | 33.95 | 33.93 | 33.66 | 33.56 | 33.27 | 54.08 | 59.85 | 27.93 | 37.02 |
| | + Spkr-norm | 67.27 | 67.33 | **67.41** | 66.85 | 33.59 | 33.94 | 33.96 | **34.07** | 33.82 | **56.16** | **60.56** | **29.50** | **37.88** |
| | Other work | 66.72 in [7], 71.4 in [4] | | | | 31.41 in [7], 42 in [4] | | | | | 50.4 in [4] | | - | |

Table 4: Summarization results on the test set using human transcripts and ASR output.

For ROUGE, we present results for a range of word compression ratios. For Pyramid and DA-level F-measure, due to space limit we only present results for two compression ratios (chosen corresponding to the best compression ratios based on ROUGE results). The "Supervised" method uses the meeting-level normalization, while "+ Speaker-norm" uses the best combination of different normalization levels, determined based on results in Table 3. "Other work" lists the best results reported in other previous studies for the same task for a comparison (these results may not be directly comparable to ours due to different subtle experimental setups). We also include an "Oracle" result in this test. For human transcript, this was generated by randomly selecting DAs from the pool of reference summary DAs, until the word compression ratio is reached. This random selection process is repeated 1000 times and the average score was reported. The "Oracle" result for ASR output was from a similar procedure, but using the ASR words with aligned DAs.

From Table 4, we can see that "+ Speaker-norm" consistently outperforms "Supervised" across different metrics. When using ROUGE measures, there is a bigger improvement using human transcripts than ASR condition (the improvement on human transcripts is statistically significant with $p < 0.01$ for R-1 and R-2; improvement on ASR output is significant with $p < 0.05$ for R-1, measured by paired t-test). The performance gap between automatic summarization and the "Oracle" score is still quite large, especially when using ROUGE-2 scores. By counting bigram match (rather than unigram in ROUGE-1), ROUGE-2 is more strict, requiring the system to extract more similar sentences to the reference summary. The performance gain of "+ Speaker-norm" over "Supervised" using DA-level F-measure and Pyramid metrics is larger than that of the ROUGE scores. In addition, the performance difference between human transcripts and ASR condition is also much smaller using these two metrics than ROUGE. This shows that these metrics are less sensitive to ASR errors (in particular, DA-level F-measure does not consider words at all). Overall, our results are comparable to (or better than) those in previous studies. In addition, our approach shows consistent improvement over the baseline normalization method across different testing conditions (transcripts, compression ratios, evaluation metrics), which suggests the robustness of our approach.

## 5. Conclusions and Future Work

In this paper, we explored using speaker characteristics to help extractive meeting summarization. We evaluated different speaker normalization methods and accordingly selected the combination of various speaker normalized features for the supervised summarization system. Our experiments were conducted using the ICSI meeting corpus with different transcripts: human transcripts and ASR output. We show that sentence length and location information is more effective when normalized based on speaker information. We obtain consistent improvement for different test conditions using our methods incorporating meeting characteristics. This study of meeting-specific characteristics for summarization suggests that more investigation is needed in this direction.

## 6. Acknowledgments

This work is supported by NSF award IIS-0845484. Any opinions expressed in this work are those of the authors and do not necessarily reflect the views of NSF.

## 7. References

[1] X. Zhu, G. Penn, and F. Rudzicz, "Summarizing multiple spoken documents: Finding evidence from untranscribed audio," in *Proc. of ACL-IJCNLP*, 2009.

[2] S. Maskey and J. Hirschberg, "Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization," in *Proc. of Eurospeech*, 2005.

[3] G. Murray, S. Renals, J. Carletta, and J. Moore, "Evaluating automatic summaries of meeting recordings," in *Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation*, 2005.

[4] M. Galley, "A skip-chain conditional random field for ranking meeting utterances by importance," in *Proc. of EMNLP*, 2006.

[5] J. Zhang, H. Y. Chan, P. Fung, and L. Cuo, "A comparative study on speech summarization of broadcast news and lecture speech," in *Proc. of Interspeech*, 2007.

[6] G. Murray and G. Carenini, "Summarizing spoken and written conversations," in *Proc. of EMNLP*, 2008.

[7] S. Xie, D. Hakkani-Tur, B. Favre, and Y. Liu, "Integrating prosodic features in extractive meeting summarization," in *Proc. of ASRU*, 2009.

[8] G. Murray, S. Renals, J. Carletta, and J. Moore, "Incorporating speaker and discourse features into speech summarization," in *Proc. of HLT-NAACL*, 2006.

[9] J. Zhang, R. H. Y. Chan, and P. Fung, "Extractive speech summarization using shallow rhetorical structure modeling," *IEEE Trans. on Audio, Speech, and Language Processing*, 2009.

[10] C. Hori and S. Furui, "A new approach to automatic speech summarization," *IEEE Transactions on Multimedia*, vol. 5, no. 3, pp. 368–378, 2003.

[11] S.-H. Lin and B. Chen, "Improved speech summarization with multiple-hypothesis representations and kullback-leibler divergence measures," in *Proc. of Interspeech*, 2009.

[12] 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 ICASSP*, 2003.

[13] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proc. of 5th SIGDIAL Workshop*, 2004.

[14] A. Stolcke, B. Chen, H. Franco, and et al., "Recent innovations in speech-to-text transcription at SRI-ICSI-UW," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1729–1744, 2006.

[15] T. Brants, "TnT – a statistical part-of-speech tagger," in *Proc. of the 6th Applied NLP Conference*, 2000.

[16] G. Penn and X. Zhu, "A critical reassessment of evaluation baselines for speech summarization," in *Proc. of ACL-HLT*, 2008.

[17] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *the Workshop on Text Summarization Branches Out*, 2004.

[18] A. Nenkova and R. Passonneau, "Evaluating content selection in summarization: The pyramid method," in *Proc. of HLT-NAACL*, 2004.