Integrating Prosodic Features in Extractive Meeting Summarization

Shasha Xie [1,2], Dilek Hakkani-Tür [2], Benoit Favre [2], Yang Liu [1]

[1] *Department of Computer Science, The University of Texas at Dallas, Richardson, TX*
{shasha,yangl}@hlt.utdallas.edu

[2] *International Computer Science Institute, Berkeley, CA*
{dilek,favre}@icsi.berkeley.edu

***Abstract*—Speech contains additional information than text that can be valuable for automatic speech summarization. In this paper, we evaluate how to effectively use acoustic/prosodic features for extractive meeting summarization, and how to integrate prosodic features with lexical and structural information for further improvement. To properly represent prosodic features, we propose different normalization methods based on speaker, topic, or local context information. Our experimental results show that using only the prosodic features we achieve better performance than using the non-prosodic information on both the human transcripts and recognition output. In addition, a decision-level combination of the prosodic and non-prosodic features yields further gain, outperforming the individual models.**

I. Introduction

Extractive speech summarization selects the most representative segments from speech (transcripts or audio) to form a generic summary. Compared to text summarization that relies on lexical, syntactic, positional and structural information, speech summarization can leverage the additional sources of information contained in speech, such as speaker and acoustic/prosodic information. These represent how the document is said other than what is said, and may provide important information for summarization.

Several recent studies have evaluated the effect of traditional textual features and speech-specific acoustic/prosodic features using classifier-based methods in speech summarization [1], [2], [3], [4], [5]. For the broadcast news domain, [1], [2] showed that the best performance was obtained by combining acoustic features with lexical, structural and discourse features; however, when speech transcription is not available, using only acoustic and structural features can achieve good performance. Similar findings are also presented in [3] using acoustic and structural features for Mandarin broadcast news. In contrast, [4] showed different patterns for lecture summarization than broadcast news domain. The acoustic and structural features are less important due to the fact that the speaking styles of anchors and reporters are relatively consistent in broadcast news, whereas the speaking styles of lecture speakers vary a lot. In addition, in [5] the authors showed some negative results — using acoustic features can not outperform a very simple baseline that selects the longest sentences to construct the summary for both broadcast news and lecture speech.

Compared to broadcast news and lecture domains, less analysis regarding prosodic features has been conducted for meeting summarization. Most of meeting summarization research using supervised learning has focused on lexical and structural features, such as [6], [7]. [8] included prosodic features with a large number of lexical, structure, and discourse features, and showed that some prosodic features were selected as the top features, such as speech rate, mean pitch (F0) of last word of current utterance. In [9], the authors compared some unsupervised methods with feature-based approaches that included prosodic features, and showed that human judges favor the feature-based approaches. However, the results they presented were obtained by using all these different types of features, and they did not evaluate the impact of using only prosodic features on the system performance.

In this paper, we focus on the effect of prosodic features on meeting summarization, aiming to answer the following questions:

- How can prosodic features be effectively represented and used for meeting summarization?
- Is it possible to construct a good text-independent summarizer by only using prosodic features?
- Can we combine prosodic with non-prosodic features for better performance?

Extractive meeting summarization is a more challenging task compared to summarization of other speech genres because of its more spontaneous style, presence of multiple speakers, less coherence and often high speech recognition error rate. These may have a significant impact on the effectiveness of features. In this study, we propose different ways to normalize the prosodic features. In a meeting recording, the speaking styles may change across different speakers, topics, or possible latent subtopics. We thus introduce different normalization methods to represent such changes of speaking styles. We also evaluate different ways to combine prosodic and non-prosodic information, at both the feature and decision level. Our experimental results on both the human transcripts and recognition output show that using only prosodic features can outperform using non-prosodic information, and their combination at the decision level yields better performance than using a single information source.

978-1-4244-5480-8/09/$26.00 © 2009 IEEE

Authorized licensed use limited to: International Computer Science Inst (ICSI). Downloaded on March 04,2010 at 14:06:47 EST from IEEE Xplore. Restrictions apply.

The rest of this paper is organized as follows. In Section II, we describe the data we used. In Section III, we introduce the acoustic/prosodic features we extracted, and a brief introduction of non-prosodic features we use to construct the baseline. The experimental results are shown in Section IV, including the results of different normalization methods, and the combination of acoustic/prosodic and non-prosodic features. Conclusion and future work are given in Section V.

II. Corpus and Experimental Setup

We use the ICSI meeting corpus [10], which contains 75 recordings from natural meetings. Each meeting is about an hour long with over 1000 sentences and has multiple speakers. These meetings have been manually transcribed and annotated with dialog acts (DA) [11], topic segments, and extractive summaries [12]. The automatic speech recognition (ASR) output for this corpus is obtained from the SRI conversational telephone speech system [13], with a word error rate of about 38.2% on the entire corpus. We align the human transcripts and ASR output, then map the human annotated DA boundaries and topic boundaries to the ASR words, such that we have human annotation of DA and topic segmentation for the ASR output. In our experiments, we use human annotated DA segments as the sentence units and manual topic segmentation for both human transcripts and ASR output, in order to focus on evaluating the effectiveness of prosodic features and avoid the interference of other factors.

The same 6 meetings as in [9] are used as the test set. We arbitrarily selected 6 other meetings from the corpus as the development set to evaluate the proposed methods and optimize parameters. We use three reference summaries from different annotators for each meeting in the test set. For the development set, we only have one reference summary for each meeting. The lengths of the reference summaries are not fixed and vary across annotators and meetings. The average word compression ratio, defined as the ratio of the number of words in the summary and the original meeting, is 14.3%, with a standard deviation of 2.9% for the test set.

To evaluate summarization performance, we use ROUGE [14], which has been used in previous studies of speech summarization. ROUGE compares the system-generated summary with reference summaries (there can be more than one reference summary), and measures different matches, such as N-gram, longest common sequence, and skip bigrams. In this paper, we mainly use ROUGE F-measures to make our research comparable with previous work.

III. Features for Extractive Meeting Summarization

The extractive summarization task can be considered as a binary classification problem and solved using supervised learning approaches. Each training and testing instance (i.e., a sentence) is represented by a set of indicative features, and positive or negative labels are used to indicate whether this sentence is in the summary or not. In this paper, we use support vector machines (SVM) (the LibSVM implementation [15]) as the classifier because of its superior performance in many binary classification tasks.

A. Acoustic/Prosodic Features

Following previous research on meeting understanding, we first extract 13 original features using Praat [16]. We have five F0 related features representing the minimum, maximum, median, mean value of F0, and the range of F0 for each instance. Similarly, we extract five energy features for the minimum, maximum, median, mean value of energy, and the range of energy of each sample. We include a duration feature which is the length of the sentence in seconds. Two speaking rate features are used, which are the sentence duration divided by the number of words and the number of letters in each sentence, respectively.

For these prosodic features, in addition to the raw values, we investigate different normalization methods based on various information.

- **Speaker-based normalization**
  Meeting recordings have multiple speakers. In general, speakers have different pitch, energy and speaking rates. In this normalization measure, each of the feature values is normalized using the mean and variance values of that feature for each speaker.
- **Topic-based normalization**
  Meetings can often be divided into several parts, each with its own topic based on the discussion. We assume that a speaker may have different prosodic behaviors for different topics according to his/her interest, roles, or conversation partners in a topic discussion. Therefore in this method, all the feature values are normalized using the mean and variance values for a topic. Note that this normalization is performed for each speaker.
- **Local window-based normalization**
  This method does not rely on content information like the topic-based normalization or use only the information from the speaker himself. We expect that the speakers are affected by other participants and may adjust their speaking rates, pitch, or energy according to who they are talking to in a local context. We simply use the previous and the following N instances to normalize the feature values.

Following the idea of local window normalization, we expect that the differences between the current sentence and its neighbors in terms of prosodic cues can indicate the importance of the current sentence, therefore we propose to include prosodic delta features — the difference between the current instance's feature values and its previous M and next M instances. The idea of computing delta features has been widely used in tasks such as speech and speaker recognition to represent dynamic information.

B. Non-Prosodic Features

The non-prosodic features we use are described in details in [7], including lexical, discourse, structural and topic-related information. The lexical features include sentence length, the

Authorized licensed use limited to: International Computer Science Inst (ICSI). Downloaded on March 04,2010 at 14:06:47 EST from IEEE Xplore. Restrictions apply.

number of words in each sentence after removing stop words, the number of frequent words and bigrams, and the number of nouns or pronouns that appear for the first time in a sentence. In addition, we derive various TF (term frequency) and IDF (inverse document frequency) related features (e.g., max, mean, sum). The cosine similarity between the sentence and the entire meeting transcript is also included in the feature set. We compute some topic-related features to capture the characteristics of different topics within a meeting. Furthermore, because the meeting corpus has multiple participants, we create some features to indicate speaker information, such as whether the sentence is said by main speakers (measured by the words they speak in the meeting), whether there is a speaker change compared to the previous sentence, and how term usage varies across speakers in a meeting. In total, there are 57 features in this category of non-prosodic information.

IV. Experiments

We first present experimental results for the development set to evaluate various factors, including prosodic feature normalization methods, effect of prosodic delta features, and combination of prosodic and non-prosodic features. Then we demonstrate the final results on the test set.

A. Baseline Results

The baseline in our experiments is using all the non-prosodic features we described in Section III-B. Table I shows the ROUGE-1 (unigram match) F-measure scores for the human transcripts and ASR output. For the ASR condition, all the non-prosodic features are extracted from the ASR transcripts for both training and testing. Since the length of the human annotated summary varies for different documents, it is hard to pre-define a proper compression ratio for the summarization system. Moreover, the performance of the system, evaluated by ROUGE scores, is affected by the length of the system-generated summary. Therefore we show results for a few different word compression ratios. We can see that the results are consistently better on human transcripts than ASR output for different compression ratios, which is expected. Comparing with the previous work [8], [12], our baseline results are very competitive. Contrary to the observation in [5], we found that the "longest sentences" baseline yields a best ROUGE-1 F-measure of 59.23% on REF (when compression ratio is 18%), and does not outperform our non-prosodic baseline on meeting data; therefore, in this study, we evaluate our proposed approaches against the baseline using non-prosodic features.

TABLE I
THE BASELINE RESULTS (%) USING NON-PROSODIC FEATURES ON DEVELOPMENT SET.

compression ratio	13%	14%	15%	16%	17%	18%
REF	67.25	**67.80**	67.76	67.56	67.22	66.86
ASR	61.78	63.23	64.35	64.73	65.11	**65.15**

B. Results of Using Acoustic/Prosodic Features

Tables II and III show the ROUGE results for human transcripts and ASR output respectively when using only the acoustic/prosodic features described in Section III-A. We present results using the raw values of the prosodic features, as well as adding different normalized features. For a comparison, results using non-prosodic features are also included in the tables (baseline column in the table).

We can see that using the raw prosodic features underperforms the baseline, with more difference on ASR output. It is worth pointing out that the prosodic features and the output confidence scores are the same for the two conditions: human transcripts and ASR output, since they only rely on speech signals. When selecting summary sentences according to a predefined compression ratio, different transcripts (human vs. ASR output) are used to select the segments for these two conditions. The degraded performance on the ASR condition is mainly due to the high WER.

TABLE II
RESULTS (%) USING RAW VALUES AND DIFFERENT NORMALIZATION METHODS OF ACOUSTIC/PROSODIC FEATURES ON DEVELOPMENT SET FOR HUMAN TRANSCRIPTS.

ratio	*baseline*	*prosodic features*			
		raw value	*with normalization*		
			speaker	*speaker & topic*	*window*
13%	67.25	65.74	65.75	67.53	67.65
14%	**67.80**	66.13	66.35	68.08	68.25
15%	67.76	**66.42**	66.84	**68.40**	69.01
16%	67.56	66.14	66.91	68.25	**69.03**
17%	67.22	65.78	**67.09**	68.20	68.63
18%	66.86	65.42	66.96	67.71	68.35

TABLE III
RESULTS (%) USING RAW VALUES AND DIFFERENT NORMALIZATION METHODS OF ACOUSTIC/PROSODIC FEATURES ON DEVELOPMENT SET FOR ASR OUTPUT.

ratio	*baseline*	*prosodic features*			
		raw value	*with normalization*		
			speaker	*speaker & topic*	*window*
13%	61.78	59.54	60.07	59.62	62.51
14%	63.23	60.75	61.61	60.78	63.58
15%	64.35	61.59	62.55	61.73	64.46
16%	64.73	62.12	63.39	62.19	65.02
17%	65.11	62.37	63.55	**62.31**	65.62
18%	**65.15**	**62.56**	**63.84**	62.22	**65.63**

Results in Tables II and III show that in general there is a consistent improvement when using feature normalization. Adding speaker normalized prosodic features performs better than raw values on both human transcripts and ASR output, which is consistent with the findings in the domain of broadcast news summarization [1]. Adding topic normalization we can further improve the performance on human transcripts. Note that after speaker and topic normalization, the performance on human transcripts has already outperformed the baseline of using non-prosodic features on human transcripts. However, for the ASR condition, adding topic information degrades the ROUGE scores. This might be because recognition performance is different for each topic. The local window-based normalization is the most effective one among these

Authorized licensed use limited to: International Computer Science Inst (ICSI). Downloaded on March 04,2010 at 14:06:47 EST from IEEE Xplore. Restrictions apply.

three normalization methods for both human transcripts and ASR output. We tried different window length for human transcripts and ASR output respectively. The best window size for human transcripts is about half of the document length, but for ASR output a smaller window is preferred (1/9 of the document size). This normalization method yields a significantly better score than the baseline (69.03 vs. 67.56 for compression ratio of 16%) on human transcripts, but the difference on ASR output is much less.

Although using the raw values of prosodic features does not perform as well as using non-prosodic features, we have shown that with proper normalization, we obtain better performance than the baseline using non-prosodic features. This shows the feasibility of extracting the summary without text information. However, it is worth pointing out that the current extractive summarization system is based on reference DA boundaries, and for automatic DA segmentation, its performance is much better in the presence of word transcripts. Because the worse performance on ASR output is mainly caused by the word errors, the performance drop when using ASR output indicates that the selected summary sentences have some recognition errors. [9] showed that summary sentences have lower WER than the average WER. But we can still see that the WER has a great influence on summarization performance for the meeting domain.

The results of adding the delta features are shown in Table IV using the best normalization setup (local window normalization). The delta features are the difference between the current instance's feature values and its previous and next M instances. We tried different M values, and the best one is 4 for human transcripts and 5 for ASR output. We notice that adding the prosodic differences substantially improves performance, with more gain on human transcripts than ASR output. Comparing with the results presented in previous work using a large set of features including lexical, structural, discourse and prosodic features [7], [8], [9], we obtain state-of-the-art results by only using acoustic/prosodic features.

TABLE IV
RESULTS (%) OF ADDING DELTA FEATURES ON DEVELOPMENT SET.

compression ratio		13%	14%	15%	16%	17%	18%
REF	*window norm*	67.65	68.25	69.01	**69.03**	68.63	68.35
	+delta	69.9	70.8	**71.18**	**71.18**	70.69	70.49
ASR	*window norm*	62.51	63.58	64.46	65.02	65.62	**65.63**
	+delta	63.42	64.65	65.78	66.07	**66.31**	**66.31**

To evaluate the effect of different prosodic features, we performed remove-one feature evaluation using human transcripts. In order to better understand the impact of the basic prosodic features, in this experiment, we only used the raw prosodic features and their local window normalized values. Furthermore, since the prosodic modeling part for human transcripts and ASR output is the same, we use results on the human transcripts to avoid the confounding effect of ASR errors in calculating the ROUGE scores. We list the five most and least effective features together with their performance loss in Table V. The ranking of the features is obtained based on the performance change when removing the feature from the entire feature set.

TABLE V
THE FIVE MOST AND LEAST EFFECTIVE PROSODIC FEATURES EVALUATED USING HUMAN TRANSCRIPTS ON DEVELOPMENT SET.

Most Effective Features	**Less Effective Features**
Energy range (5.1%)	*f0 median* (-0.3%)
Normalized f0 median (4.7%)	*normalized energy mean* (0.7%)
energy mean (4.3%)	*speaking rate (letter)* (0.7%)
energy median (3.8%)	*normalized energy maximum* (1.0%)
f0 maximum (3.1%)	*normalized f0 mean* (1.1%)

C. Combination with Non-prosodic Features

Finally we investigate if these prosodic features combine well with the non-prosodic features to further improve the system performance. We use two different combination methods. First is the feature level combination. We combine the non-prosodic features with the prosodic feature set that yielded the best results (basic acoustic/prosodic features with local window normalization and delta features) in one large feature set. The second one is a decision level combination. We train separate models for these two information sources and then for each test instance linearly combine the confidence scores from the two models. The final summary is constructed by selecting the instances with higher combined confidence scores. The experimental results are presented in Table VI, along with the individual results using prosodic or non-prosodic information only. For the decision level combination method, we varied the combination weights for human transcripts and ASR output respectively, and show the best results here.

From the results, we can see that feature level combination hurts the summarization performance compared to using one information source only, and there is more degradation on human transcripts. However, we observe performance improvement using decision level combination for both human transcripts and ASR output. Interestingly, we notice that for human transcripts, a higher weight was given to the prosodic model (0.7 for prosodic and 0.3 for non-prosodic). This is consistent with the individual model performance — the results of prosodic features are much better than the non-prosodic ones. For ASR condition, equal weights (0.5 and 0.5) were used for the two models, which can be explained in part by the fact that the two systems have similar performance.

TABLE VI
RESULTS (%) OF INTEGRATING PROSODIC AND NON-PROSODIC INFORMATION, IN COMPARISON WITH USING ONLY ONE INFORMATION SOURCE ON DEVELOPMENT SET.

compression ratio		13%	14%	15%	16%	17%	18%
REF	*non-prosodic*	67.25	**67.80**	67.76	67.56	67.22	66.86
	prosodic	69.9	70.8	**71.18**	**71.18**	70.69	70.49
	feature combine	66.10	**66.70**	66.61	66.51	66.40	65.84
	decision combine	70.50	70.92	**71.40**	70.91	70.67	70.18
ASR	*non-prosodic*	61.78	63.23	64.35	64.73	65.11	**65.15**
	prosodic	63.42	64.65	65.78	66.07	**66.31**	**66.31**
	feature combine	63.06	64.16	64.75	65.40	**65.44**	65.23
	decision combine	64.15	65.36	66.12	66.44	**67.10**	67.02

Authorized licensed use limited to: International Computer Science Inst (ICSI). Downloaded on March 04,2010 at 14:06:47 EST from IEEE Xplore. Restrictions apply.

TABLE VII
RESULTS (%) ON TEST SET.

		ROUGE-1 F-measure						ROUGE-2 F-measure					
compression ratio		13%	14%	15%	16%	17%	18%	13%	14%	15%	16%	17%	18%
REF	non-prosodic	68.29	69.15	69.71	69.83	**69.91**	69.78	33.29	34.05	34.51	35.10	35.61	**36.14**
	prosodic	68.87	69.85	70.28	**70.56**	70.36	69.86	36.14	37.51	38.07	39.03	39.10	**39.23**
	combined	69.64	70.54	71.04	**71.23**	71.00	70.62	37.34	37.90	38.73	39.11	39.20	**39.37**
ASR	non-prosodic	58.71	60.04	61.00	61.58	62.15	**62.26**	25.41	25.92	26.47	27.14	27.70	**28.02**
	prosodic	64.55	65.18	**65.51**	**65.51**	65.41	65.07	27.68	28.34	29.07	29.43	29.43	**29.86**
	combined	65.14	65.91	66.53	**66.72**	66.63	66.19	27.84	28.63	29.58	30.30	31.00	**31.41**

D. Results on Test Set

The results on the test set are provided in Table VII for both human transcripts and ASR output, using only non-prosodic or prosodic information, and their combination. ROUGE-2 scores (bigram matching) are also included in order to provide more information for comparison. We selected the best setup based on the results on the development set and applied it to the test set. The prosodic feature set includes local window normalization and delta features. The combined system is based on a decision level combination of the prosodic and non-prosodic models. We observe similar trends as on the development set. Using only the prosodic features we obtain better performance than non-prosodic information, and the combination of the models yields further improvement. These results are consistent across human transcripts and ASR output, and ROUGE-1 and ROUGE-2 scores. We also verified that the results are significantly better than the baseline according to a paired t-test ($p < 0.05$).

V. CONCLUSION

Compared to text summarization, more information can be exploited for meeting summarization, such as speaker and acoustic/prosodic information. In this paper, we evaluated how to effectively represent acoustic/prosodic features, and how to integrate them with traditional textual features for meeting summarization. For prosodic information, we adopted three normalization methods based on different sources of information: speaker, topic, and local context information. Our experimental results on the ICSI meeting corpus showed that these normalization methods improve the performance compared with only using the raw values. We also demonstrated additional gain when adding the delta features to represent how a sentence is different from its neighbors. When using only the prosodic features, we were able to outperform the baseline of using the non-prosodic features. In addition, we evaluated different approaches to integrate prosodic and non-prosodic information, and showed that a decision level combination can improve summarization performance upon that of the individual models.

For future work, we will perform a more rigorous feature selection. We used remove-one feature evaluation on a subset of features in this study. We plan to use other feature evaluation methods, such as forward feature selection, to select effective features and better understand the impact of the combination of different features. We will also use a large feature set including all the prosodic and non-prosodic features. In addition, we will evaluate methods to more effectively combine prosodic and textual features that are very different in nature. Furthermore, we used human annotated dialog acts and topic segmentation on human transcripts or aligned them to ASR output in this work. We will investigate the effect of automatic DA and topic segmentation on acoustic/prosodic features in our future study.

ACKNOWLEDGMENT

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811), and NSF grant IIS-0845484.

REFERENCES

[1] S. Maskey and J. Hirschberg, "Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization," in *Proc. of Interspeech*, 2005.
[2] ——, "Summarizing speech without text using Hidden Markov Models," in *Proc. of HLT-NAACL*, 2006.
[3] J. Zhang and P. Fung, "Speech summarization without lexical features for mandarin broadcast news," in *Proc. of HLT-NAACL*, 2007.
[4] J. Zhang, H. Y. Chan, P. Fung, and L. Cao, "A comparative study on speech summarization of broadcast news and lecture speech," in *Proc. of ICASSP*, 2007.
[5] G. Penn and X. Zhu, "A critical reassessment of evaluation baselines for speech summarization," in *Proc. of ACL*, 2008.
[6] A. H. Buist, W. Kraaij, and S. Raaijmakers, "Automatic summarization of meeting data: A feasibility study," in *Proc. of CLIN*, 2005.
[7] S. Xie, Y. Liu, and H. Lin, "Evaluating the effectiveness of features and sampling in extractive meeting summarization," in *Proc. of IEEE Spoken Language Technology (SLT)*, 2008.
[8] M. Galley, "A skip-chain conditional random field for ranking meeting utterances by importance," in *Proc. of EMNLP*, 2006.
[9] G. Murray, S. Renals, and J. Carletta, "Extractive summarization of meeting recordings," in *Proc. of Interspeech*, 2005.
[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 ICASSP*, 2003.
[11] 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.
[12] 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.
[13] Q. Zhu, A. Stolcke, B. Chen, and N. Morgan, "Using MLP features in SRI's conversational speech recognition system," in *Proc. of Interspeech*, 2005.
[14] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *the Workshop on Text Summarization Branches Out*, 2004.
[15] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," 2001, software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.
[16] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," http://www.praat.org/, 2006.

Authorized licensed use limited to: International Computer Science Inst (ICSI). Downloaded on March 04,2010 at 14:06:47 EST from IEEE Xplore. Restrictions apply.