A Comparative Study on Extractive Speech Summarization of Broadcast News and Parliamentary Meeting Speech

Jian Zhang
School of Computer Science
Dongguan University of Technology
Dongguan, China
zjian03@gmail.com

Huaqiang Yuan
School of Computer Science
Dongguan University of Technology
Dongguan, China
hyuan66@163.com

***Abstract*—We carry out a comprehensive study of acoustic/prosodic, linguistic and structural features for speech summarization, contrasting two genres of speech, namely Mandarin Broadcast News and Cantonese Parliamentary Speech. We find that structural features are superior to acoustic and lexical features when summarizing broadcast news because of the fact that in the same Mandarin broadcast program, the distribution and flow of summary utterances are relatively consistent. We use different machine learning algorithms to construct the binary-class summarizers to select the best features for extractive summarization, and obtain state-of-the-art performances: ROUGE-L F-measure of 0.682 for Mandarin Broadcast News, and 0.737 for Cantonese Parliamentary Meeting Speech. In the case of Parliamentary Meeting Speech summarization, we show that our summarizer performed surprisingly well ROUGE-L F-measure of 0.729 by using ASR transcription despite the character error rate of 27%. We also discover that the different choices of algorithms almost do not affect the consistency of our findings.**

***Keywords*-Feature Comparison; Extractive Speech Summarization; Meeting Speech; Broadcast News**

I. Introduction

Speech summarization, a technique of extracting important information and removing irrelevant information from a spoken document or audio document, has become a new area of study in the last few years. Many text-based features and speech-based features have been proposed in speech summarization systems for summarizing different genres of speech data [1], [2], [3], [4], [5]. [1] proposed a method that calculates the maximum summarization score of a set of words extracted from an ASR utterance, according to a target summarization ratio. The summarization score consists of word significance measure and linguistic likelihood which are all text-based features and extracted from transcriptions. [3] focused on how to use acoustic information alone for speech summarization. [4] proposed the use of probabilistic latent topical information for extractive summarization of Mandarin Broadcast news. [5] adapted the notion of risk minimization for extractive speech summarization by formulating the selection of summary sentences as a decision-making problem. [6]used regression and sampling to make the classes more balanced. [2] performed an empirical study of the usefulness of different types of features—acoustic, structural, and lexical features—in selecting summary utterances for English broadcast news. They train a Bayesian Network classifier as their summarizer. They find that the structural features are superior to other features as predictors of summary utterances.

In this paper, considering that there has not been an empirical study investigating the relative contribution of different feature combinations—acoustic, structural, and lexical features—as predictors for summarizing different genres of Mandarin or Cantonese speech data, we perform a thorough investigation on the performance of our summarizer for the two genres of speech: broadcast news and Parliamentary meeting speech with these features. We also investigate whether different machine learning methods as the summarizer affect the relative contribution of different feature combinations.

II. Features and Summarization Methods

A. Acoustic/Prosodic Features

Researchers commonly use acoustic/prosodic variation – changes in pitch, intensity, speaking rate – and duration of pause extracted from speech signals, for tagging the important contents of their speeches [7]. We also investigate these features for their efficiency in predicting summary utterances on Mandarin broadcast news and Parliamentary meeting speech. Our acoustic feature set contains twelve features described as follows. ***Duration***: time duration of the utterance; ***SpeakingRate***: average syllable duration; ***F0I-IV***: the minimum/maximum/mean/slop of F0 value; ***F0V***: the difference between *F0II* and *F0I*; ***EI-IV***: the minimum/maximum/mean/slop of Energy value; ***EV***: the difference between *EII* and *EI*.

We calculate *Duration* from the annotated manual transcriptions that align the audio documents. We then obtain *SpeakingRate* by phonetic forced alignment by HTK [8]. we extract F0 features and energy features from audio data by using Praat [9].

B. Structural/Discourse Features

We use the structural/discourse feature PoissonNoun proposed in Justin's work [10] which is based on the following assumptions: first, if a sentence contains new noun words, it probably contains new information. The noun word's Poisson score varies according to its position [11]. We use Poisson distribution to approximate the variation. Second, if a noun word occurs frequently, it is likely to be more important than other noun words, and the sentence

978-1-4799-5330-1/14/$31.00 ©2014 IEEE

Table I
LEXICAL FEATURES

Feature Name	Feature Description
LenI-III	the number of words in the current/previous /next utterance
NEI	the number of Named Entities in the current utterance
NEII	the number of Named Entities which appear in the utterance at the first time in a story
NEIII	the ratio of the number of unique Named Entities to the number of all Named Entities
TFIDF	tf*idf of each word in the utterance
Cosine	cosine similarity score between two utterance vectors

with these high frequency noun words should be included in a summary.

Normally, the broadcast news stories have similar structure in the same program. Each news starts with an anchor, followed by the formal report of the story by other reporters or interviewees. Based on this finding, we define four structural features for broadcast news: ***Position***, ***TurnI***, ***TurnII*** and ***TurnIII***. We calculate these structural features from the annotated information of Mandarin broadcast news corpus.

- *Position*: one news has *k* utterances, then we set $(1-(0/k))$ as *Position* value of the first utterance in the news, and set $(1-((i-1)/k))$ as *Position* value of the i^{th} utterance.
- *TurnI*: one news has *m* turns, then we set $(1-(0/m))$ as *TurnI* value of the utterances which belong to the first turn's content, and set $(1-((j-1)/m))$ as *TurnI* values of the utterances which belong to the j^{th} turn's content.
- *TurnII* and *TurnIII*: the previous/next turn's *TurnI* value.

We use ***Position***, ***TurnI***, ***TurnII***, ***TurnIII***, and ***Poisson Noun*** as structural feature set of broadcast news. Considering that one parliamentary meeting always has only one turn, we use ***Poisson Noun*** as structural feature of meeting speech.

C. Lexical Features

Our lexical feature set contains eight features: ***LenI***, ***LenII***, ***LenIII***, ***NEI***, ***NEII***, ***NEIII***, ***TFIDF*** and ***Cosine***, described in Table I.

All lexical features are extracted from the manual transcriptions or ASR transcriptions. For calculating length features, we segment Chinese words of the broadcast and meeting transcriptions. We use an off-the-shelf Chinese lexical analysis system, the open source ICTCLAS [12], which labels Chinese words using a set of 39 tags, to segment and POS tag our corpora. We use an in-house Named Entity Recognition (NER) system for extracting Named Entities.

D. Summarization Methods

We consider the extractive summarization as a binary classification problem. We predict whether each utterance of the broadcast news or meeting speech should be in a summary or not.

We investigate three different classification algorithms as the summarizer: (1)binary SVM; (2)Hidden Markov SVM (HMSVM) [13]; (3) Conditional Random Field (CRF) [14]. We use Radial Basis Function(RBF) kernel for constructing binary SVM classifier [15].

Considering that HMSVM combines the advantages of maximum margin classifier and kernels with the elegance and efficiency of HMMs, and can effectively handle the dependency between neighboring utterances, we train a binary classifier.

Considering that CRF provides a probabilistic framework for calculating the probability of the optimal summary label sequence globally conditioned on the feature vector sequence, we build a CRF classifier for selecting the salient utterances from the input utterances.

III. THE CORPORA AND REFERENCE SUMMARIES

We use a portion of the 1997 Hub4 Mandarin corpus available via LDC as experiment data. The related audio data were recorded from China Central Television(CCTV) International News programs, including 23-day broadcast from 14th January, 1997 to 21st April, 1997. Each broadcast lasts approximately 32 minutes, and has been hand-segmented into speaker turns. We evaluate our summarizer on the several-turns news stories each of which is presented by more than one reporter. The corpus has 347 news which contain 4748 utterances in total. For evaluation, we manually annotated these broadcast news, and extracted segments as reference summaries at compression rate(CR) 20%.

Our parliamentary meeting speech corpus is collected from the Hong Kong Legislative Council. We use all 70 Ordinary Session meeting data from the year 2008 and the year 2009, including audio files, Hansards, and minutes. All wave files are segmented into several utterance units by three human annotators. We use our in-house spontaneous speech recognition system to produce an automatic transcription with manual utterance segmentation. After adding noise and garbage models to the lexicon, the system performs at 73% accuracy, or 27% character error rate. The reference summaries of CR 20% are generated by three human annotators based on the content of the minutes and Hansards.

IV. EXPERIMENTS AND EVALUATION

A. Experiment Settings and Evaluation Metrics

We perform two sets of experiments: Experiment I for Mandarin Broadcast New summarization and Experiment II for Cantonese Parliamentary Meeting Speech summarization. In Experiment I, we use 70% of the broadcast corpus consisting of 3,294 utterances as training set and the remaining 1,454 utterances as held-out test set, upon which our summarizer is tested. We use these reference summaries with different compression rate for training different summarizer. In Experiment II, we use 70% of the meeting corpus: 49 meetings of 147,294 utterances as training set and remaining 11 meetings of 33,066 utterances as held-out test set.

Table II
EVALUATION BY ROUGE-L F-MEASURE IN EXPERIMENT I

Feature Set (Methods)	Le	St	Ac	Le +St	Ac +St	Ac +Le	All
(single SVM)	.543	.59	.311	.635	.609	.529	.637
(HMSVM)	.575	.621	.392	.672	.63	.562	.676
(CRF)	.581	.63	.453	.664	**.64**	.595	**.682**

Ac: Acoustic; St: Structural; Le: Lexical; All: all features

Table III
EVALUATION BY ROUGE-L F-MEASURE IN EXPERIMENT II

Feature Set (Methods)	Le	St	Ac	Le +St	Ac +St	Ac +Le	All
(single SVM)-H	.652	.603	.592	.68	.692	.698	.70
(HMSVM)-H	.681	.629	.621	.714	.72	.726	.729
(CRF)-H	.69	.637	.633	.72	**.728**	.733	**.737**
(single SVM)-A	.631	.58	.592	.664	.689	.682	.69
(HMSVM)-A	.66	.612	.621	.696	.71	.704	.717
(CRF)-A	.667	.628	.633	.705	.723	.715	.729

H: Hansard Transcriptions; A: ASR Transcriptions
Ac: Acoustic; St: Structural; Le: Lexical; All: all features

We evaluate our summarizer's performance by the metric ROUGE (Recall Oriented Understudy for Gisting Evaluation)which can measure overlap units between automatic summaries and reference summaries.(We also measured the performance by F-measure, but the results are not included here due to space limitations.)

We use ROUGE-L (summary-level Longest Common Subsequence) precision, recall and F-measure, which are described by equation (1,2,3) [16].

$$\mathrm{P}_{lcs} = \frac{\sum_{i=1}^{u} \mathrm{LCS}_{\mathrm{U}}(r_i, C)}{n} \tag{1}$$

$$\mathrm{R}_{lcs} = \frac{\sum_{i=1}^{u} \mathrm{LCS}_{\mathrm{U}}(r_i, C)}{m} \tag{2}$$

$$\mathrm{Fmeasure}_{lcs} = \frac{2 \times \mathrm{P}_{lcs} \times \mathrm{R}_{lcs}}{\mathrm{P}_{lcs} + \mathrm{R}_{lcs}} \tag{3}$$

Given a reference summary of *u* utterances containing a total of *m* words and a candidate summary of *v* utterances containing a total of *n* words, $LCS_U(r_i, C)$ is the LCS score of the union longest common subsequence between reference utterance r_i and candidate summary C.

B. Summarization Performance

Firstly, from Table II and III, we can see that by using acoustic and structural features, our summarizer yields the best performance: ROUGE-L F-measure of 0.682 for summarizing Mandarin broadcast news, while by using acoustic, lexical, and structural features, our summarizer yields the best performance: ROUGE-L F-measure of 0.737 for summarizing Cantonese Parliamentary meeting speech.

Table II also shows that when we trained a CRF classifier as our summarizer by using structural features, the performance is superior to that by using other features: ROUGE-L F-measure of 0.63, 4.9% higher than the ROUGE-L F-measure produced by using lexical features and 17.7% higher than the ROUGE-L F-measure produced by using acoustic features. Furthermore, we find that structural features especially *Position* are the most useful predictors for extractive summarization. This result is in contrast to the finding that structural features are less important than lexical features in Table III. This is due to the fact that in the same Mandarin broadcast program, the distribution and flow of summary utterances are relatively consistent. Therefore structural features play a key role in speech summarization for Mandarin broadcast news. From Table II and Table III, we can see that the different choices of summarization methods almost do not affect the consistency of this finding.

Table II shows that by using the combination of acoustic and structural features, our summarizer produces good performance at ROUGE-L F-measure of 0.64 which is 3.8% lower than the performance by using all features and 4.5% higher than the performance by using acoustic features and lexical features, while from Table III we find that our summarizer yields F-measure of 0.728 which is only 0.9% lower than the performance by using all features and 0.5% higher than the performance by using acoustic features and lexical features. That is to say, lexical features play a more important role in speech summarization for Mandarin broadcast news, compared with that for meeting speech. This is due to the fact that the speaking styles and keyword lists of anchors and reporters are relatively consistent in the broadcast news, while the speaking styles and keyword lists of meeting speakers always variable.

From Table III we make a surprising discovery that summarization performance is very high: ROUGE-L F-measure of 0.729 by using all features extracted from ASR transcriptions, even when the ASR accuracy is only 73%, which is only 0.8% lower than that by using all features from Hansard transcriptions. Upon error analysis, we find that 92% of all mis-recognized words are single characters, which in Chinese often do not bear any content. As such, the effect of recognition errors on extractive summarization results is minimal. This finding suggests that it is possible to summarize Cantonese meeting speech data without placing a stringent demand on speech recognition accuracy.

Besides, from Table II and Table III, we also can see that the different choices of summarization methods almost do not affect the consistency of our findings. We also find that CRF based summarizer performs consistently better than HMSVM based summarizer and SVM based summarizer. This is because CRF algorithm can produce the optimal summary label sequence globally conditioned on the feature vector sequence.

V. CONCLUSION

In this paper, we have presented a first known empirical study on speech summarization with acoustic, structural, and lexical features, contrasting two genres of speech data: Mandarin broadcast news and Cantonese parliamentary meeting speech. We found that structural features are

superior to acoustic and lexical features when summarizing broadcast news because of the fact that in the same Mandarin broadcast program, the distribution and flow of summary utterances are relatively consistent. Furthermore, we have shown that, compared with Cantonese parliamentary meeting speech summarization, lexical features play a more important role in speech summarization for Mandarin broadcast news because of the speaking styles and keyword lists of anchors and reporters are relatively consistent in the broadcast news, while the speaking styles and keyword lists of meeting speakers always variable.

Meanwhile, our CRF based summarizer yielded state-of-the-art performance: ROUGE-L F-measure of 0.682 for Mandarin broadcast and ROUGE-L F-measure of 0.737 for Cantonese parliamentary meeting speech. Moreover, we have shown that our summarizer performed surprisingly well ROUGE-L F-measure of 0.729 by using ASR transcription despite the character error rate of 27%. This finding also suggested that high quality speech summarization can be achieved without stringent requirement on speech recognition accuracy. We also found that the different choices of summarization methods almost do not affect the consistency of our findings.

ACKNOWLEDGMENT

This work was partially supported by the Natural Science Foundation of China (Grant No.61300197), the Foundation of Guangdong Educational Committee (Grant No.2012KJCX0099), and the Natural Science Foundation of Guangdong Province of China (Grant No.S2012040007560). The authors would like to thank Pascale FUNG for sharing her experiment resources.

REFERENCES

[1] C. Hori and S. Furui, "Advances in automatic speech summarization," *Proc. EUROSPEECH2001*, vol. 3, pp. 1771–1774, 2001.

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

[3] S. R. Maskey and J. Hirschberg, "Summarizing Speech Without Text Using Hidden Markov Models," *Proc. NAACL*, 2006.

[4] B. Chen, Y. Yeh, Y. Huang, and Y. Chen, "Chinese Spoken Document Summarization Using Probabilistic Latent Topical Information," *Proc. ICASSP*, 2006.

[5] B. Chen and S.-H. Lin, "A risk-aware modeling framework for speech summarization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 211–222, 2012.

[6] S. Xie and Y. Liu, "Improving supervised learning for meeting summarization using sampling and regression," *Computer Speech & Language*, vol. 24, no. 3, pp. 495–514, 2010.

[7] J. Hirschberg, "Communication and prosody: Functional aspects of prosody," *Speech Communication*, vol. 36, no. 1, pp. 31–43, 2002.

[8] P. Woodland, C. Leggetter, J. Odell, V. Valtchev, and S. Young, "The development of the 1994 HTK large vocabulary speech recognition system," *Proceedings ARPA Workshop on Spoken Language Systems Technology*, pp. 104–109, 1995.

[9] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer, version 3.4," *Institute of Phonetic Sciences of the University of Amsterdam, Report*, vol. 132, p. 182, 1996.

[10] J. Zhang, S. Huang, and P. Fung, "RSHMM++ for extractive lecture speech summarization," in *IEEE Spoken Language Technology Workshop, 2008. SLT 2008*, 2008, pp. 161–164.

[11] K. Church and W. Gale, "Poisson mixtures," *Natural Language Engineering*, vol. 1, no. 2, pp. 163–190, 1995.

[12] K. Zhang and Q. Liu, "ICTCLAS," *Institute of Computing Technology, Chinese Academy of Sciences: http://www. ict. ac. cn/freeware/003_ictclas. asp*, 2002.

[13] Y. Altun, I. Tsochantaridis, T. Hofmann *et al.*, "Hidden markov support vector machines," in *ICML*, vol. 3, 2003, pp. 3–10.

[14] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[15] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[16] C. Lin, "Rouge: A Package for Automatic Evaluation of Summaries," *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, pp. 25–26, 2004.