A Comparative Study on Speech Summarization of Broadcast News and Lecture Speech

Jian Zhang[1], *Ho Yin Chan*[2], *Pascale Fung*[3], *Lu Cao*[4]

Human Language Technology Center
Electronic and Computer Engineering
University of Science and Technology
Clear Water Bay,Hong Kong

zjustin@ust.hk,ricky@cs.ust.hk,
pascale@ee.ust.hk,ee_clx@stu.ust.hk

10.21437/Interspeech.2007-717

Abstract

We carry out a comprehensive study of acoustic/prosodic, linguistic and structural features for speech summarization, contrasting two genres of speech, namely Broadcast News and Lecture Speech. We find that acoustic and structural features are more important for Broadcast News summarization due to the speaking styles of anchors and reporters, as well as typical news story flow. Due to the relatively small contribution of lexical features, Broadcast News summarization does not depend heavily on ASR accuracies. We use SVM based summarizer to select the best features for extractive summarization, and obtain state-of-the-art performances: ROUGE-L F-measure of 0.64 for Mandarin Broadcast News, and 0.65 for Mandarin Lecture Speech. In the case of Lecture Speech summarization where lexical features are more important, we make the surprising discovery that summarization performance is very high (0.63 ROUGE-L F-measure) even when the ASR accuracy is low (21% CER).

Index Terms: speech summarization

1. 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 English or Japanese speech data [1, 2, 3, 4, 5, 6]. [1] proposes a method that calculates the maximum summarization score of a set of words extracted from an ASR sentence, 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] proposes a multi-stage compaction approach based on lexical features such as TFIDF scores and Named Entity frequency. [5] uses statistical methods to identify words to be included in a summary, based on linguistic and acoustic/prosodic features of the Japanese broadcast news transcriptions. [6] extracts structural features from audio documents to help summarization. [2] focuses on how to use acoustic information alone for speech summarization. [7] also uses acoustic features for speech summarization of spontaneous speech. [4] performs an empirical study of the usefulness of different types of features—acoustic, structural, and lexical features—in selecting summary sentences for English broadcast news. They find that the structural features are superior to other features as predictors of summary sentences.

There have also been some research efforts on summarizing Mandarin speech data [8, 9]. [8] proposes the use of probabilistic latent topical information for extractive summarization of Mandarin spoken documents. [9] presents spoken document summarization scheme using acoustic, prosodic, and linguistic information. However, there has not been an empirical study investigating the relative contribution of different feature combinations—acoustic, structural, and lexical features—as predictors for summarizing Mandarin broadcast news and lecture speech. In this paper, we perform a thorough investigation on the performance of our summarizer for the two genres of Mandarin speech data with these features.

In Section 2 we describe our summarizer and the features used in experiments. We describe the Mandarin broadcast corpus and lecture speech corpus on which our system operates in Section 3. In Section 4 we perform our experiments and evaluate the results. Our conclusion follows in Section 5.

2. Features and Methodology

In this section, we propose acoustic, structural, and lexical features that we use to predict whether the sentence should be included in a summary or not. We then go on to describe our summarizer.

2.1. Acoustic/Prosodic Features

Acoustic/prosodic features in speech summarization system are usually extracted from audio data. Researchers commonly use acoustic/prosodic variation – changes in pitch, intensity, speaking rate – and duration of pause for tagging the important contents of their speeches [10]. We also investigate these features for their efficiency in predicting summary sentences on Mandarin speech data.

Our acoustic feature set contains thirteen features: ***DurationI***, ***DurationII***, ***SpeakingRate***, ***F0I***, ***F0II***, ***F0III***, ***F0IV***, ***F0V***, ***EI***, ***EII***, ***EIII***, ***EIV*** and ***EV***. We describe these features in Table 1.

We calculate *DurationI* from the annotated manual transcriptions that align the audio documents. We then obtain *DurationII* and *SpeakingRate* by phonetic forced alignment by HTK. Next, we extract F0 features and energy features from audio

Table 1: Acoustic/Prosodic Features

Feature Name	Feature Description
DurationI	time duration of the sentence
DurationII	the average phoneme duration
SpeakingRate	average syllable duration
F0I	F0's minimum value
F0II	F0's maximum value
F0III	the difference between *F0II* and *F0I*
F0IV	the mean of F0 value
F0V	F0 slope
EI	minimum energy value
EII	maximum energy value
EIII	the difference between *EII* and *EI*
EIV	the mean of energy value
EV	energy slope

data by using Praat [11].

2.2. Structural Features

The probability distributions of words in texts can be adequately estimated by Poisson mixture [12]. Noun words that may be primitive organizers of written text also follow Poisson distribution [13]. Based on these findings, we define a structural and discourse feature, called ***Poisson Noun*** as in equation(1).

$$Poisson\ Noun_j(i) = \frac{\sum_{k=1}^{N_i} \text{ppois}(p, \lambda) \times TF(k)}{N_i} \quad (1)$$

In equation (1), N_i is the number of noun words in sentence i, which belongs to story or presentation j; $TF(k)$ is the frequency of word k in story or presentation j; p means that word k appeared in the p^{th} time within story or presentation j.

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

Table 2: Lexical Features

Feature Name	Feature Description
LenI	the number of words in the sentence
LenII	the previous sentence's *LenI* value
LenIII	the next sentence's *LenI* value
NEI	the number of Named Entities in the sentence
NEII	the number of Named Entities which appear in the sentence at the first time in a news
NEIII	the ratio of the number of unique Named Entities to the number of all Named Entities
TFIDF	tf*idf; tf and idf defined as equation(2,3)
Cosine	cosine similarity measure between two sentence vectors

- *TurnIII*: the next turn's *TurnI* value.

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

2.3. Lexical Features

Our lexical feature set contains eight features: ***LenI***, ***LenII***, ***LenIII***, ***NEI***, ***NEII***, ***NEIII***, ***TFIDF*** and ***Cosine***. These features are described in Table 2.

$$\text{tf} = \frac{n_i}{\sum_k n_k} \quad (2)$$

with n_i being the number of occurrences of the considered word, and the denominator is the number of occurrences of all words in a story or presentation.

$$\text{idf} = \log \frac{|D|}{|(d_i \supset t_i)|} \quad (3)$$

$|D|$ is the total number of sentences in the considered story or presentation. $|(d_i \supset t_i)|$ is the number of sentences where the word t_i appears.

All lexical features are extracted from the manual transcriptions or ASR transcriptions. For calculating length features, we segment Chinese words of the broadcast and lecture transcriptions. We use an off-the-shelf Chinese lexical analysis system, the open source ICTCLAS [14], which labels Chinese words using a set of 39 tags, to segment and POS tag our corpora.

We use a Named Entity Recognition (NER) system for extracting Named Entities. The NER system introduced boosting, a promising and theoretically well-founded machine learning method, to Chinese named entity identification [15].

2.4. Summarizer

We consider the extractive summarization as a binary classification problem; that is to say, we predict whether each sentence of the broadcast news should be in a summary or not. Our summarizer contains the preprocessing stage and the estimating stage. The preprocessing stage extracts features and normalizes all features by equation (4) in advance.

$$N_j = \frac{w_j - mean(w_j)}{dev(w_j)} \quad (4)$$

Here, w_j is the original value of feature j which is used to describe sentence i; $mean(w_j)$ is the mean value of feature j in our training set or test set; $dev(w_j)$ is the standard deviation value of feature j in our training set or test set.

The estimating stage predicts whether each sentence of the broadcast news is in a summary or not. We use Radial Basis Function(RBF) kernel for constructing SVM classifier as in [16] which provides LIBSVM, a library for support vector machines. We adopt the principle of cross-validation in the pre-training process. First, we divide the training set into v subsets of equal size. Sequentially one subset is tested using the classifier trained on the remaining $(v-1)$ subsets. Thus, each instance of the whole training set is predicted once. In the i^{th} training cycle, we can get one $w_2(i)$ parameter, which can avoid the effect resulted from unbalance between *summary-sent class* and *other-sent class*.

After pre-training process, we set the mean of $w_2(i)$ as the w_2 parameter of the estimating part of our summarizer. Then we build the summarizer on the whole training set and predict the sentences of test set.

3. 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. They include 23-day broadcast from 14th January, 1997 to 21st April, 1997, which contain 593 stories and weather forecasts. 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 sentences in total. For evaluation, we manually annotated these broadcast news, and extracted segments as reference summaries at compression rate(CR): 10%, 15% and 20%. We build three baselines referring to different versions of reference summaries. When using CR 10% summaries, we build the baseline by choosing the first 10% of sentences from each story. Our baseline results in F-measure score are given in Table 3.

Our lecture speech corpus contains wave files of 60 presentations in Mandarin Chinese, the presentation slides (power point), and manual transcriptions. Each presentation contains about 222 units and lasts approximately 15 minutes. All wave files are segmented into several sentence units by human judges. We use our in house spontaneous speech recognition system to produce an automatic transcription. After adding noise and garbage models to the lexicon, the system performs at 78.2% accuracy, or 21.8% character error rate. We generate our reference summaries of CR 30% and 20% based on the content of the presentation slides and manual transcriptions.

4. Experiments and Evaluation

4.1. Experiment Settings and Evaluation Metrics

We perform two sets of experiments: Experiment I for Mandarin Broadcast New summarization and Experiment II for Mandarin Lecture Speech summarization. In Experiment I, we use 70% of the broadcast corpus consisting of 3294 sentences as training set and the remaining 1454 sentences 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 lecture corpus: 28 presentations of 5342 sentences as training set and remaining 12 presentations of 1740 sentences as held-out test set.

Table 3: Evaluation by ROUGE-L F-measure in Experiment I

Feature Set	CR10%	CR15%	CR20%	Ave
Le	.5734	.5156	.5432	.5441
St	**.587**	**.6066**	**.5967**	**.5968**
Ac	.2522	.4063	.3113	.3238
Le+St	.5919	.6859	.6354	.6377
Ac+St	**.617**	**.601**	**.609**	**.609**
Ac+Le	.5708	.4928	.5289	.5308
Ac+Le+St	.5989	.621	.6098	.6099
Baseline	.21	.32	.43	.32

Ac: Acoustic; St: Structural; Le: Lexical;
CR:Compression Rate

Table 4: Evaluation by ROUGE-L F-measure in Experiment II

Feature Set	Prec	Recall	F-meas
Le(M)	.5458	.7261	.6232
Le(A)	.4785	.8643	.6159
St(M)	**.4763**	**.721**	**.5736**
St(A)	**.4477**	**.8436**	**.5849**
Ac(M)	.4095	.9232	.5673
Ac(A)	.4091	.9227	.5669
Le+St(M)	.5613	.7649	.6475
Le+St(A)	.5002	.832	.6248
Ac+St(M)	**.4835**	**.8841**	**.6251**
Ac+St(A)	**.4463**	**.9104**	**.5989**
Ac+Le(M)	.5213	.7897	.628
Ac+Le(A)	.4772	.8691	.6161
Ac+Le+St(M)	.5531	.7798	.6472
Ac+Le+St(A)	.4896	.8742	.6277

(M): Manual Transcriptions; (A): ASR Transcriptions
Ac: Acoustic; St: Structural; Le: Lexical;

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 (5,6,7) [17].

$$\mathrm{P}_{lcs} = \frac{\sum_{i=1}^{u} \mathrm{LCS_U}(r_i, C)}{n} \tag{5}$$

$$\mathrm{R}_{lcs} = \frac{\sum_{i=1}^{u} \mathrm{LCS_U}(r_i, C)}{m} \tag{6}$$

$$\mathrm{Fmeasure}_{lcs} = \frac{2 \times \mathrm{P}_{lcs} \times \mathrm{R}_{lcs}}{\mathrm{P}_{lcs} + \mathrm{R}_{lcs}} \tag{7}$$

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

4.2. Summarization Performance

Firstly, from Table 3 and 4, we can see that by using lexical and structural features, our summarizer yields ROUGE-L F-

measure of 0.6377 for Mandarin broadcast and 0.6475 for Mandarin lecture speech.

Table 3 also shows that structural features are superior to acoustic and lexical features: ROUGE-L F-measure of 0.5968, 27.68% higher than the baseline and 5.27% higher than the average ROUGE-L F-measure produced by using lexical 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 4. This is due to the fact that in the same Mandarin broadcast program, the distribution and flow of summary sentences are relatively consistent. Therefore structural features play a key role in speech summarization for Mandarin broadcast news.

Furthermore, we find that in comparison with lecture speech summarization, acoustic and structural features are more important for Broadcast News summarization, and the contribution of lexical features is relatively small. Table 3 shows that by using the combination of acoustic and structural features, our summarizer produces good performance at ROUGE-L F-measure of 0.609 which is only is 0.09% lower than the performance by using all features and 6.49% higher than the performance by using lexical features, while in Table 4 we find that our summarizer yields F-measure of 0.6251 which is only 0.19% higher than the performance by using lexical features. This is due to the fact that the speaking styles of anchors and reporters are relatively consistent in the broadcast news, while the speaking styles of lecture speakers always variable.

Besides, from Table 4 we make a surprising discovery that summarization performance is very high: ROUGE-L F-measure of 0.6277 by using all features even when the ASR accuracy is only 78.2%. Upon error analysis, we find that 95.2% of all misrecognized words are single characters, which in Chinese oftern 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 Mandarine speech data without placing a stringent demand on speech recognition accuracy.

5. 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 Mandarin speech data: broadcast news and lecture speech. We found that structural features are superior to acoustic and lexical features when summarizing broadcast news. In particular, we have shown that, compared with lecture speech summarization, acoustic and structural features make more important contribution to Mandarin broadcast news summarization because of the relatively consistent speaking styles of anchors and reports, as well as distribution and flow of summary sentences in the same broadcast program. We found that our summarizer performs surprisingly well at the average F-measure of 0.609 only by using acoustic and structural features.

Meanwhile, our SVM based summarizer yielded state-of-the-art performance: ROUGE-L F-measure of 0.6377 for Mandarin broadcast and ROUGE-L F-measure of 0.6472 for Mandarin lecture speech. Moreover, we have shown that our summarizer performed surprisingly well ROUGE-L F-measure of 0.6277 by using ASR transcription despite the character error rate of 21%. This finding also suggested that high quality speech summarization can be achieved without stringent requirement on speech recognition accuracy.

6. 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, "Summarizing Speech Without Text Using Hidden Markov Models," *Proc. NAACL*, 2006.

[3] B. Kolluru, Y. Gotoh, and H. Christensen, "Multi-stage compaction approach to broadcast news summarisation," *Proc. of Interspeech 2005, Lisbon, Portugal*, 2005.

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

[5] A. Inoue, T. Mikami, and Y. Yamashita, "Improvement of Speech Summarization Using Prosodic Information," *Proc. of Speech Prosody*, 2004.

[6] S. Maskey and J. Hirschberg, "Automatic summarization of broadcast news using structural features," *Proceedings of Eurospeech 2003*, 2003.

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

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

[9] C. Huang, C. Hsieh, and C. Wu, "Spoken Document Summarization Using Acoustic, Prosodic and Semantic Information," *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pp. 434–437, 2005.

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

[11] 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.

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

[13] A. Badalamenti, "Speech Parts as Poisson Processes," *Journal of Psycholinguistic Research*, vol. 30, no. 5, pp. 497–527, 2001.

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

[15] X. Yu, M. Carpuat, and D. Wu, "Boosting for chinese named entity recognition," in *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*. Sydney, Australia: Association for Computational Linguistics, July 2006, pp. 150–153.

[16] C. Chang and C. Lin, "LIBSVM: a library for support vector machines," *Software available at http://www. csie. ntu. edu. tw/cjlin/libsvm*, vol. 80, pp. 604–611, 2001.

[17] 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.