IMPROVING LECTURE SPEECH SUMMARIZATION USING RHETORICAL INFORMATION

Justin Jian Zhang, Ho Yin Chan, Pascale Fung

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

ABSTRACT

We propose a novel method of extractive summarization of lecture speech based on unsupervised learning of its rhetorical structure. We present empirical evidence showing that rhetorical structure is the underlying semantics which is then rendered in linguistic and acoustic/prosodic forms in lecture speech. We present a first thorough investigation of the relative contribution of linguistic versus acoustic features and show that, at least for lecture speech, what is said is more important than how it is said. We base our experiments on conference speeches and corresponding presentation slides as the latter is a faithful description of the rhetorical structure of the former. We find that discourse features from broadcast news are not applicable to lecture speech. By using rhetorical structure information in our summarizer, its performance reaches 67.87% ROUGE-L F-measure at 30% compression, surpassing all previously reported results. The performance is also superior to the 66.47% ROUGE-L F-measure of baseline summarization performance without rhetorical information. We also show that, despite a 29.7% character error rate in speech recognition, extractive summarization performs relatively well, underlining the fact that spontaneity in lecture speech does not affect the central meaning of lecture speech.

Index Terms— Speech Summarization, Lecture Speech, Rhetorical Information

1. INTRODUCTION

Automatic summarization of lecture speech is a useful tool for education, research, and personal interests. Research efforts in summarizing lecture speech are still limited today [1, 2], while broadcast news summarization has been a major focus in this area[3, 4, 5, 6, 7, 8]. We argue that since lecture speech is different from broadcast news stylistically, its rhetorical structure is also different, as is manifested in the outlines of presentation slides associated with lecture speech. It has been shown in previous work that news speech summarization relies more on structure features than on lexical features[9]. It remains as an open question whether systems trained to summarize broadcast news are directly applicable to lecture speech. There has been an extensive amount of work in modeling discourse structures by using prosodic and acoustic features[10, 11]. Indeed, it has been argued that while rhetorical structure is the underlying message in lecture speech, both acoustic speech and lexical linguistics are representations of this message. We seek to show that lexical linguistics and acoustic speech are intertwined and are direct renditions of rhetorical structure. We propose that lecture summarization performance can benefit from modeling rhetorical structure explicitly. We also note that lecture speech is planned but also contains a certain amount of spontaneity. It is interesting to know whether this places a particular stringent demand on speaker-independent automatic speech recognition systems.

This paper is organized as follows: Section 2 describes our motivation and methodology in modeling rhetorical structures in lecture speech. Section 3 outlines the acoustic/prosodic, lexical, discourse characteristics of lecture speech, and depicts our extractive summarizer. In section 4 we describe the corpus and our lecture speech recognition system for automatic transcription, and then describe the experiments and results. We finally conclude in Section 5.

2. RHETORICAL STRUCTURE CHARACTERISTICS IN LECTURE SPEECH

Unlike conversational speech, lectures and presentations are planned speech. Like all planned speech, lecture speakers follow a relatively rigid rhetorical structure: s/he starts with an overview of the topic to be presented, followed by the actual content with more detailed descriptions, and then concludes at the end. According to rhetorical structure theory[12], these are elements of a text plan. In this paper, we envision the text plan of lecture speech to be as illustrated in Figure 1.

In 2003, we first propose using Hidden Markov models to model rhetorical structure in a text summarization task[13, 14]. Considering that humans tend to segment presentations into introduction, content, and conclusion sections, [15] proposes a summarization method based on this rhetorical struc-

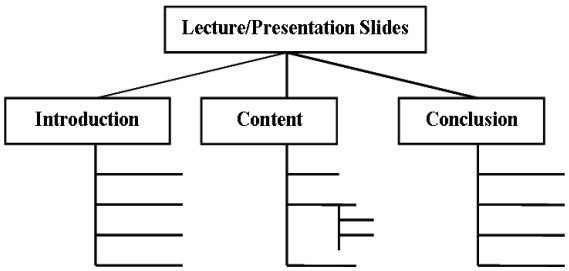

Fig. 1. Text plan of lecture speech

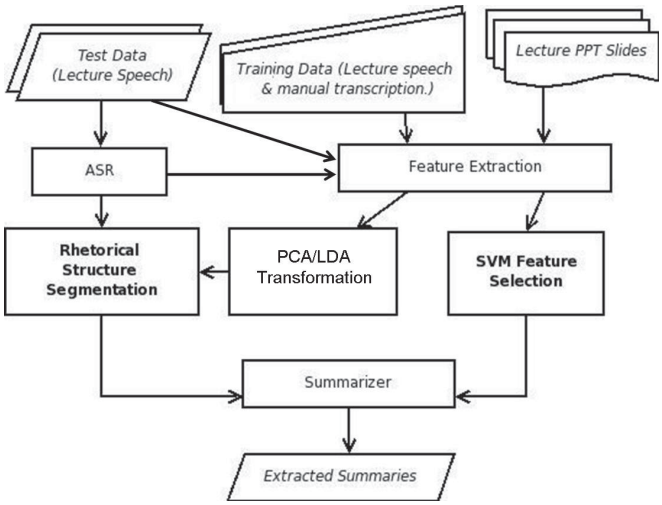

Fig. 2. Extractive Summarization of Lecture Speech with Rhetorical Structure Modeling

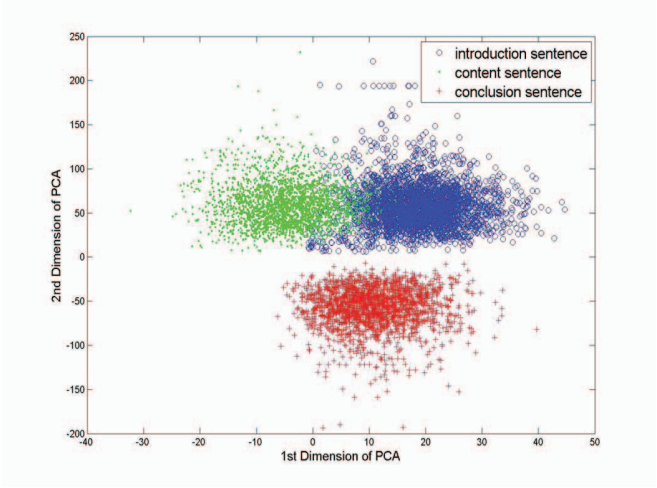

Fig. 3. All feature vectors in the training data and *conclusion*

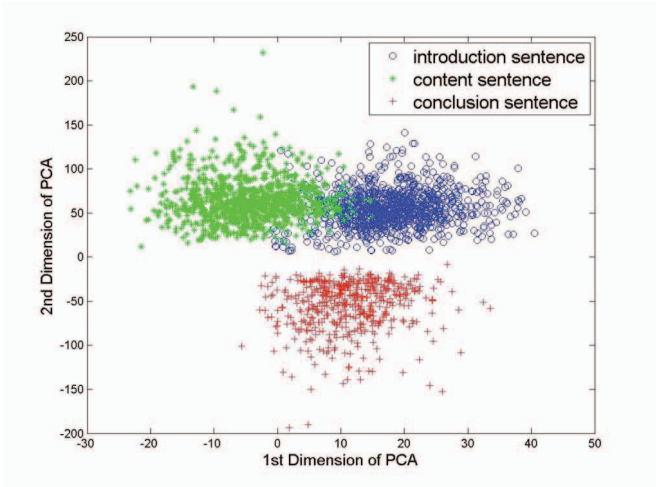

Fig. 4. All feature vectors in the test data and *conclusion*

ture characteristic. They estimate the introduction and conclusion section boundaries based on the Hearst method[16].

Many linguists believe that speech acoustics contribute to rhetorical and discourse structure. [11] provide empirical evidence that discourses can be segmented reliably, and that acoustic features are used by speakers to convey linguistic structure at the discourse level in English domain. Likewise, in our work, we assume that correlation between acoustic features and discourse structure exists in Mandarin lecture speech, and that we may use the acoustic features for extracting the discourse structure.

Since lecture speeches are mostly based on presentation slides with main gisting points, rather than read from a script, the content and format of the presentation slides is a faithful representation of the rhetorical structure of lecture speech. In our work, we use PCA projection of all acoustic/phonetic, lexical, and discourse features of lecture speech render the underlying rhetorical structure. PCA reduces the multidimensional feature vectors to two dimensions by finding the orthogonal vectors that best represent all the features. The PCA transformation is given by equation 1. Figure 3 and Figure 4 are visualizations of the rhetorical structure of lecture speech.

$$Y^T = X^T W = V \sum W^T \quad (1)$$

Where $\sum \mathrm{W}^T$ is the singular value decomposition (SVD) of X^T.

3. EXTRACTIVE SUMMARIZATION OF LECTURE SPEECH

We consider extractive summarization as a binary classification problem; that is to say, we predict whether each sentence

of the lecture speech should be in a summary or not. We use Radial Basis Function(RBF) kernel for constructing SVM classifier as in [17].

3.1. Acoustic/Phonetic Features of Lecture Speech

There is a large amount of previous work seeking to demonstrate that acoustic prosodic profile of speech closely models its discourse or rhetorical structure[18, 19, 10, 11]. [20] suggested that acoustic features are useful for extracting salient sentences from Broadcast News. [21] also use acoustic features such as F0 and Energy features for speech summarization of spontaneous speech.

We argue that acoustic/prosodic features in lecture speech may not be as meaningful as those in either news speech or conversational speech.

First, lecture speech differs greatly from Broadcast News due to speaker variability. Most of Broadcast News consists of speech by anchors and reporters who are professionally trained to use prosody to emphasize important points[7]. On the other hand, lecture speakers have a wider range of speaking style as many are not trained speakers.

Second, lecture speech is planned, and is less spontaneous than conversational speech. A typical lecture speaker (in a class, at a conference), facing a receptive audience, often sounds dull and monotonic, compared to in a conversation. Unlike conversational speech, there are often long sentences in lecture speech delimited by only a short pause[22].

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 [23]. We also investigate these features for their efficiency in predicting summary sentences of lecture presentation.

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 [24]. Next, we extract F0 features and energy features from audio data by using Praat [25].

By using feature selection on these acoustic/prosodic features, we find that DurationI and DurationII are extremely discriminatory, while other features make little contribution to predicting summary sentences. It is probably caused by the sharp variation of acoustic features in spontaneous speech. Besides, different speaking styles may add relative insignificance of these acoustic features.

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

3.2. Lexical and Discourse Features of Lecture Speech

[26] showed that prosodic models outperform language models in speech tasks. Previous work on Broadcast News summarization have even shown that salient sentences can be found based on their acoustic and structural features alone, without lexical features[20, 9].

However, we argue that since lecture speech is prosodically less stylistic than Broadcast News, the relative contribution of lexical features might be more important in summarization.

In fact, one approach to speech summarization is simply to extract salient sentences from transcriptions of speech. This approach, however, would seem to place a stringent demand on the accuracy of automatic transcriptions. Indeed, [22] suggest that speaker adaptation might be necessary for lecture speech transcription.

Similar to text summarization, lexical information can help us predict the summary sentences. We extract eight lexical features from transcriptions: ***LenI***, ***LenII***, ***LenIII***, ***TFIDF*** and ***Cosine***. We describe these features in Table 2.

$$\mathrm{tf} = \frac{n_i}{\sum_k n_k} \tag{2}$$

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 presentation.

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

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

We extract all lexical features from the manual and ASR transcriptions respectively. For calculating length features, we segment Chinese words in these transcriptions. We use an

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
TFIDF	tf*idf; tf and idf defined as equation(2,3)
Cosine	cosine similarity measure between two sentence vectors

off-the-shelf Chinese lexical analysis system, the open source HIT IR Lab Chinese Segmenter [27] to segment and part of speech tag our corpora.

We extract the discourse feature–Poisson Noun[9] described in equation(4), which contains rhetorical and discourse structure information, based on the section boundaries of each presentation.

$$Poisson\ Noun_j(i) = \frac{\sum_{k=1}^{N_i} \text{ppois}(p, \lambda) \times TF(k)}{N_i} \quad (4)$$

In equation (4), N_i is the number of noun words in sentence i, which belongs to section j; $TF(k)$ is the frequency of word k in news j; p means that word k appeared in the p^{th} time within section j.

3.3. Extraction of Rhetorical Structure

In this paper, we extract the rhetorical structure of the presentation by using slides or unsupervised learning from feature vectors of each sentence. The extraction process is described in Algorithm 1.

Based on Algorithm 1, we segment the transcriptions into 3 sections. And then we train three summarization models for different sections. We call them the introduction summarizer, the content summarizer, and the conclusion summarizer. We then extract the summaries of each section by using the corresponding summarizer. Finally we combine the three summaries as the single summary. We call this kind of summarizer as segmental summarizer, different from the baseline whole summarizer that do not use rhetorical information.

Algorithm 1 Extracting rhetorical structure of lecture speech

For training data

S1 Split each presentation slides into three sections: introduction; content; conclusion; then use three content vectors for representing the slides.

S2 Extract one content vector for each sentence of the transcriptions and calculate cosine similarity between transcription's content vectors and slides' content vectors; and then find the section boundaries based on the cosine similarity distribution.

For test data

T1 Initialize section boundaries as follows: 30% of introduction, 40% of content, and 30% of conclusion.

T2 Extract acoustic and lexical features from each sentence of the transcriptions, and extract the discourse feature based on the initialized section boundaries.

T3 Use one vector containing all features for representing one sentence and project each sentence vector into two dimension form by using Principle Component Analysis (PCA).

T4 Use K-means to cluster the two-dimensional vectors into three groups and then produce new section boundaries.

T5 Recalculate discourse feature based on the new section boundaries and go to **T3** until the section boundaries remain the same.

4. EXPERIMENTS AND EVALUATION

4.1. The Corpus

Our lecture speech corpus contains wave files of 60 presentations recorded from the NCMMSC2005 conference, together with power point files, and manual transcriptions. Each presentation contains about 222 units and lasts approximately 15 minutes. Here we use 40 of the 60 presentations with well organized power point for our experiments. Besides the manual sentence segmentation and transcriptions, we automatically segment transcriptions into sentence units and produce the transcriptions by in house lecture speech recognition system[28].

4.2. Lecture Speech Recognition System

4.2.1. Sentence Boundary Detection

For sentence boundary detection in lecture presentations, we first trained Gaussian Mixture Models (GMM) for the silence, noise, Mandarin initial speech, Mandarin final speech and non English word speech events using the convention EM algorithm, where each of the GMMs contains 256 components and is represented by 3-7 HMM states. A grammar based Viterbi decoder is then used to find the GMM sequences with time boundaries. The GMM sequences are then relabeled to speech/non-speech labels. The final speech boundaries are obtained by merging the speech labels which are nearby (0.2s) and padding silence (0.1s) in either end of the merged speech segments. The average length of the automatic created speech segments is 2.2s, which is shorter compare to the 3.9s average length of the manual segments.

In addition, we adopt rule-based segment merging model for sentence boundary adaptation. We then obtain longer sentences with an average length of 3.75s, which is only 0.15s shorter than average length of the manual segmented sentence.

4.2.2. Performances of the ASR system

Our ASR system runs in multiple passes. In the first pass, a decoder performs time-synchronous Viterbi beam search through a lexical tree to produce 1-best result and a lattice, where context dependent cross word triphone HMMs and word based bigram language model are applied. Lattice re-scoring is then performed using trigram language model to obtain another 1-best result. A bigram branch and a trigram branch are obtained and unsupervised acoustic model adaptation with the MLLR approach is applied on each branch. Lattice rescoring is then performed on each branch using the adapted acoustic model, and produces new recognized results. The recognized text from the braches are then mixed and used for unsupervised trigram language model adaptation. A final re-decoding is done by using the adapted acoustic model and the adapted trigram language model. We obtained 69.7% and 70.3% accuracy for manual and automatic segmented sentences respectively in our system.

4.3. Experiment Settings

In our experiment, we use 40 presentations of the corpus described in Section 4.1. We use 85% of the lecture corpus consisting of 34 presentations that contain 6049 sentences as training set and the remaining 6 presentations that contain 1116 sentences by automatic sentence segmentation or 1033 sentences by manual sentence segmentation as held-out test set, upon which our summarizer is tested.

4.4. Evaluation Metrics

We use ROUGE-L(summary-level Longest Common Subsequence) precision and recall, which are described by equation (5,6), as evaluation metrics [9]. We then calculate ROUGE-L F-measure by using them.

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

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

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.

4.5. Summarization Performances

By using ASR transcriptions, we perform several sets of experiments on the segmental summarizer and the whole summarizer. We obtain our reference summaries of summarization ratio (SR) 30% based on cosine similarity between the content of power point and transcriptions. Based on a combination of acoustic, lexical, and discourse features we obtain several versions of revised section boundaries(between introduction and content section or between content and conclusion section of each presentation). We use these revised boundaries in our segmental summarizer evaluation and compare the performance of the segmental summarizer with that of the whole summarizer described in Table 3.

Table 3. Evaluation by ROUGE-L F-measure

Features	S1	S2	S3
Le	.6491	**.6787**	**.6756**
Ac	.6195	.6095	.5823
Di	.6178	.2343	.1500
Ac+Le	.6600	.6709	.6438
Le+Di	.6555	.5750	.5441
Ac+Di	.6335	.4370	.4257
Ac+Le+Di	**.6647**	.5984	.5984

S1: Whole Summarizer on Manual sentence boundaries trans;
S2: Segmental Summarizer on Manual sentence boundaries trans;
S3: Segmental Summarizer on Auto sentence boundaries trans;
Ac: Acoustic; Di: Discourse; Le: Lexical

Firstly, Table 3 shows that by using lexical features, our segmental summarizer yields the best performance: ROUGE-L F-measure of 0.6787, 1.4% higher than the best performance produced by the whole summarizer. This clearly shows that the contribution of rhetorical structure in the lecture speech.

Table 3 shows that lexical features rank higher than acoustic features in all experiments. This shows that, at least for lecture speech, what is said is more important than how it is said. This is due to the speaking styles of lecture speakers variable.

From Table 3, we also can see that the contribution of discourse feature is even less important in the segmental summarizer than in the whole summarizer. This clearly shows that discourse feature from broadcast news are not applicable to lecture speech as they are based on sentence position.

Despite the fact that ASR accuracy is only 70.3%, our segmental summarizer produces very high performance: ROUGE-L F-measure of 0.6787 under manual sentence segmentation and ROUGE-L F-measure of 0.6756 under automatic sentence segmentation. Upon error analysis, we find that 91.76% of all misrecognized units, which are generated by substitution, insertion or deletion errors, are meaningless characters or words. These units often do not bear the core content of Mandarin presentations.

5. CONCLUSIONS

We present a novel method of extraction-based lecture speech summarization by using rhetorical structure of presentation. Using rhetorical structure, we improve summarization performance from 0.6647 to 0.6787 ROUGE-L F-measure for 30% compression, which is higher than reported in all pre-

vious works. We then show that the contribution of lexical features is more than that of acoustic features, which shows that what is said is more important than how it is said for lecture speech summarization. We also discovered that the discourse features from broadcast news are not useful in lecture speech. While extractive summarization relies on finding salient sentences from automatic transcriptions of lecture speech, we find that summarization performance is still very good despite a 29.7% character error rate. This is because the misrecognized words and characters are mostly function words, stop words and filled pauses, which are not pertinent to the central message of the lecture speech.

6. REFERENCES

[1] S. Furui and T. Kawahara, "Transcription and Distillation of Spontaneous Speech," *Springer Handbook on Speech Processing and Speech Communication. Springer Press, Germany*, pp. 1–27, 2007.

[2] C. Hori and A. Waible, "Spontaneous Speech Consolidation for Spoken Language Applications," *Proc. of Interspeech 2005*.

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

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

[5] K. Koumpis and S. Renals, "Automatic summarization of voicemail messages using lexical and prosodic features," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 2, no. 1, pp. 1–24, 2005.

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

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

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

[9] Jian Zhang and Pascale Fung, "Speech summarization without lexical features for Mandarin broadcast news," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume*, Rochester, New York, April 2007, pp. 213–216, Association for Computational Linguistics.

[10] J. Hirschberg and C. Nakatani, "A prosodic analysis of discourse segments in direction-giving monologues," *Proc. ACL*, pp. 286–293, 1996.

[11] C.H. Nakatani, J. Hirschberg, and B.J. Grosz, "Discourse structure in spoken language: Studies on speech corpora," *AAAI 1995 Spring Symposium Series: Empirical Methods in Discourse Interpretation and Generation*, pp. 106–112, 1995.

[12] WC Mann and SA Thompson, "Rhetorical Structure Theory: A Theory of Text Organization, USC," *Information Sciences Institute Research Report RR-87-190*, 1987.

[13] P. Fung, G. Ngai, and C.S. Cheung, "Combining Optimal Clustering and Hidden Markov Models for Extractive Summarization," *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12*, pp. 21–28, 2003.

[14] P. Fung and G. Ngai, "One story, one flow: Hidden Markov Story Models for multilingual multidocument summarization," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 3, no. 2, pp. 1–16, 2006.

[15] M. Hirohata, Y. Shinnaka, K. Iwano, and S. Furui, "Sentence extraction-based presentation summarization techniques and evaluation metrics," *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 1, 2005.

[16] M.A. Hearst, "TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages," *Computational Linguistics*, vol. 23, no. 1, pp. 33–64, 1997.

[17] C.C. Chang and C.J. 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.

[18] M.A.K. Halliday, *Intonation and grammar in British English*, Mouton, 1967.

[19] D.R. Ladd, *Intonational Phonology*, Cambridge University Press, 1996.

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

[21] S. Furui, T. Kikuchi, Y. Shinnaka, and C. Hori, "Speech-to-text and speech-to-speech summarization of spontaneous speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 4, pp. 401–408, 2004.

[22] T. Kawahara, H. Nanjo, and S. Furui, "Automatic transcription of spontaneous lecture speech," *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*, pp. 186–189, 2001.

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

[24] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK Book (for HTK Version 3.0)," *Cambridge University*, 2000.

[25] 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, pp. 182, 1996.

[26] E. Shriberg and A. Stolcke, "Prosody modeling for automatic speech recognition and understanding," *Mathematical Foundations of Speech and Language Processing*, pp. 105–114, 2004.

[27] H.Z.T.L.J. Ma and X. Liao, "Chinese Word Segmentation with Multiple Postprocessors in HIT-IRLab," *SIGHAN*.

[28] Pascale Fung Lu Cao Ho Yin Chan, Justin Jian Zhang, "A MANDARIN LECTURE SPEECH TRANSCRIPTION SYSTEM FOR SPEECH SUMMARIZATION," *IEEE Automatic Speech Recognition and Understanding Workshop*, 2007.