RHETORICAL STRUCTURE MODELING FOR LECTURE SPEECH SUMMARIZATION

Pascale Fung, Justin Jian Zhang, Ricky Ho Yin Chan, Shilei Huang

Human Language Technology Center
Department of Electronic & Computer Engineering
University of Science & Technology (HKUST)
Clear Water Bay, Hong Kong

{pascale, zjustin, eehuangs}@ece.ust.hk, ricky@cs.ust.hk

ABSTRACT

We propose an extractive summarization system with a novel non-generative probabilistic framework for speech summarization. One of the most under-utilized features in extractive summarization is rhetorical information -semantically cohesive units that are hidden in spoken documents. We propose Rhetorical-State Hidden Markov Models (RSHMMs) to automatically decode this underlying structure in speech. We show that RSHMMs give a 68.67% ROUGE-L F-measure, a 6.44% absolute increase in lecture speech summarization performance compared to the baseline system without using RSHMM. We further propose an enhanced Rhetorical-State Hidden Markov Model (RSHMM++) for extracting hierarchical structural summaries from lecture speech. We show that RSHMM++ gives a 72.01% ROUGE-L F-measure, a 3.34% absolute increase in lecture speech summarization performance compared to the baseline system without using rhetorical information. We also propose Relaxed DTW for compiling reference summaries.

Index Terms— Rhetorical structure, speech summarization, lecture speech

1. INTRODUCTION

Automatic summarization of lecture speech is the process of recognition, distillation and presentation of spoken documents in a structural text form, to be presented to the user. Unlike written documents, on one hand, spoken documents or transcriptions produced by automatic speech recognition system, often lack explicit structure information, such as: titles, subtitles, paragraph/topic boundaries, fonts and so on to help interpret the underlying semantic information and produce summaries with hierarchical structure. On the other hand, other than linguistic features extracted from ASR transcriptions, acoustic/phonetic characteristics can be extracted from relevant speech data.

Extractive summarization is a common approach of speech summarization. There are many existing extractive summarization systems using acoustic and linguistic features [1, 2, 3, 4, 5, 6]. Nevertheless, those systems ignore one important under-utilized information—rhetorical structure—existing in the speech data. Lectures and presentations are planned semi-spontaneous speech. Like all planned speech, lecture speakers follow a relatively rigid rhetorical structure. According to rhetorical structure theory [7], a text plan is composed by several elements. We envision the text plan of lecture speech as illustrated in Figure 1. For a written document, rhetorical structure is the story flow of the document and consists of several rhetorical units which are represented by paragraphs or sub-paragraphs. Similarly in the spoken document and relevant speech data, rhetorical structure also exists.

Our previous work [6] and other researchers have suggested that rhetorical information exist also in spoken documents and efficient modeling of this information is helpful to the summarization task. [8] and [9] used the Hearst method [10] to segment documents and detect topics for text summarization and topic adaptation of speech recognition systems for long speech archives respectively.

Some summarization systems make use of the simplest type of rhetorical information, commonly known as discourse feature, such as sentence or noun position offset from the beginning of the text [11, 12, 2]. [13] applied a HMM generative framework to broadcast news speech summarization. This type of discourse feature works well for news reports, but not as well in other genres such as lecture presentations [6]. Our proposed work combines the idea of rhetorical structure information and HMM probabilistic framework into summarizing lecture speech presentations.

In our previous work, we have proposed Rhetorical-State HMMs (RSHMMs) to first segment lecture speech into rhetorical units , before segmental SVMs for the summarization step. The rhetorical information can help improve summarization process [14]. However, inaccurate rhetorical unit boundaries by RSHMM tend to be carried over to the summarization step, causing further errors. In this paper, we propose enhanced Rhetorical-State HMM (RHMM++), shown in Fig-

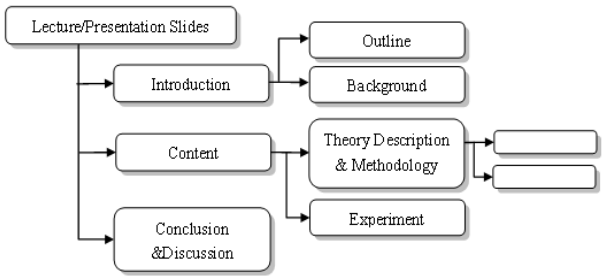

Fig. 1. Hierarchical text plan of lecture speech

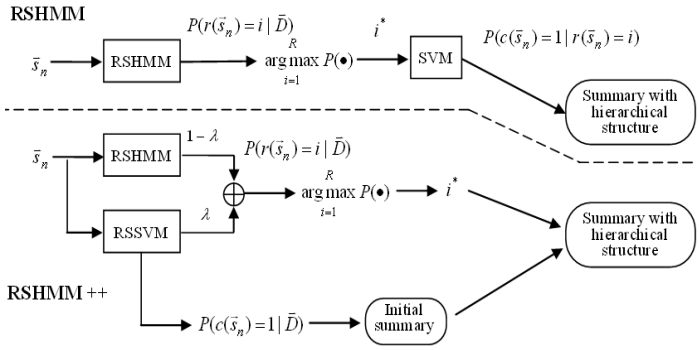

Fig. 2. Architecture of RSHMM++

ure 2. In RSHMM++, a Rhetorical-State SVM is used to extract an initial summary which is then transformed into a segmental structure extracted by a hybrid RSHMM/RSSVM. Besides, we compile reference summaries using Relaxed DTW between power point sentences and transcriptions. We describe the process of compiling reference summaries in Section 3.3. The agreement between automatically extracted reference summary and humans can reach 75%.

This paper is organized as follows: Section 2 depicts RSHMM, RSHMM++ and how to adopt them for extracting summaries with hierarchical structure. We then outline the acoustic/prosodic, and linguistic features for representing each sentence, depict how to compile reference summaries and perform our experiments in Section 3. We evaluate the results in Section 4. Our conclusion follows in Section 5.

2. METHODOLOGY

2.1. Approach 1: RSHMM for Lecture Speech Summarization

2.1.1. Extracting rhetorical structure by RSHMMs

The previous approach of segmental summarization showed us that rhetorical segments are indeed helpful. Looking further, as illustrated in Figure 1, rhetorical structure is in fact a hierarchical structure. In view of this, we propose a second approach of building Rhetorical State Hidden Markov Models

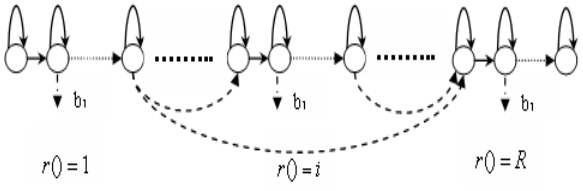

Fig. 3. Spoken document representation with RSHMMs

with state transitions that represent several kinds of rhetorical relations to better model this rhetorical structure.

In extractive summarization of lecture speech, for a transcribed document D with a sequence of N recognized sentences S_j from the ASR output: $D = \{S_1, S_2, \ldots, S_N\}$, j =1,2,...,N. We use RSHMMs to model the underlying rhetorical structure of the transcribed document. Figure 3 shows the concatenation of R RSHMMs to represent a spoken document.

Each RSHMM state contains a probability distribution $b_j()$ for the input feature vector $\mathbf{S}_n$ obtained from the acoustic and linguistic features for the sentence s_n. We use mixtures of multivariate Gaussian distribution as the probability distribution as in formula 1.

$$\mathrm{b_j}() = \sum_{m=1}^{M} \mathrm{c_{jm}} \mathrm{N}(\mathbf{S}_n; \mu_{jm}, \xi_{jm}) \quad (1)$$

where M is the number of mixture components in the state, c_{jm} is the weight of the m'th component and $\mathrm{N}(\mathbf{S}_n; \mu_{jm}, \xi_{jm})$ is a multivariate Gaussian with mean vector μ and covariance matrix ξ for the acoustic and linguistic features, as in formula 2.

$$\mathrm{N}(\mathbf{S}_n; \mu_{jm}, \xi_{jm}) = \frac{1}{\sqrt{(2\pi)^n |\xi|}} \mathrm{e}^{-\frac{1}{2}(\mathbf{S}_\mathrm{n}-\mu)^\mathrm{T} \xi^{-1} (\mathbf{S}_\mathrm{n}-\mu)} \quad (2)$$

Given that the spoken document used in this work are lecture presentations and assuming that these presentations consistently follow a rhetorical structure containing R sections, R HMMs (i.e. r_1, r_2,..., and r_R) are built to represent the respective sections. Each HMM is represented by three states, roughly corresponding to the beginning, the middle, and the ending part in a rhetorical "paragraph". Each of the states contain several Gaussian components. We trained each of the HMMs by performing Viterbi initialization, then followed by Baum-Welch re-estimation using the forward-backward algorithm.

We then place the trained HMMs into a sequential network structure of (r_1, r_2, ..., and r_R). We finally use the Viterbi algorithm to find the best rhetorical unit sequence for document D with N sentence represented by $\{S_1, S_2, \ldots, S_N\}$. This is equal to finding the best state sequence $Q^* = \{q_1, q_2, \ldots, q_N\}$in formula 3 and 4.

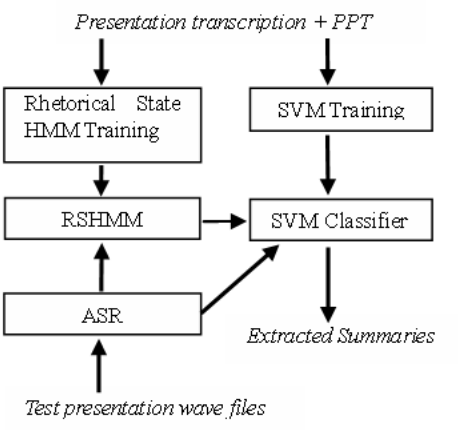

Fig. 4. Extractive Summarization of Lecture Speech Using RSHMMs

$$\mathrm{Q}^* = \underset{Q}{\operatorname{argmax}}\{P(S_1, S_2, \ldots, S_N | r_1, r_2, \ldots, r_R)\} \quad (3)$$

$$\mathrm{Q}^* = \underset{Q}{\operatorname{argmax}}\{a_{q(0)q(1)} \prod_{j=1}^{N} b_{q(j)}(s_j) a_{q(j)q(j+1)}\} \quad (4)$$

Finally, we annotate each sentence of the given document as i^* which approximately maximizes $P(r(s_k) = i|\overrightarrow{D})$.

$$i^* = \underset{i=1}{\overset{R}{\operatorname{argmax}}} P(r(s_k) = i|\overrightarrow{D}) \quad (5)$$

where $r()$ is a mapping function for the rhetorical unit, and we have a total of R rhetorical units in a single document. $\overrightarrow{D}$ is the feature vector representing the sentence sequence D.

2.1.2. Extractive summarization with shallow rhetorical structure

This step in our algorithm assigns each sentence to its place in a particular rhetorical unit, again roughly corresponding to a single power point slide in a presentation. Next we want to find M sentences to be classified as summary sentences by using the salient sentence classification function $c()$.

Based on the probabilistic framework, extractive summarization task is equal to estimating $P(c(S_j) = 1|\overrightarrow{D})$ of each sentence s_j.

We propose a novel probabilistic framework—RSHMM-enhanced SVM—for summarization process [14]. We approximate $P(c(S_j) = 1|\overrightarrow{D})$ in the following expression:

$$P(c(S_j) = 1|\overrightarrow{D}) \approx P(c(S_j) = 1|\overrightarrow{D}, r(S_j) = i^*) \quad (6)$$

where $c()$ is the salient sentence classification function; i^* can be obtained by equation (5). We then predict whether sentence s_j is a summary sentence or not by using a probability threshold. We set the probability $threshold(i^*)$ to be the compression ratio of rhetorical unit i^*.

$$P(c(S_j) = 1|\overrightarrow{D}, r(S_j) = i^*) > threshold(i^*) \quad (7)$$

We model $P(c(S_j) = 1|\overrightarrow{D}, r(S_j) = i^*)$ by SVM classifier with Radial Basis Function (RBF) kernel, as described in equation (8), SVM classifier as in [15]. One SVM classifier is trained for each rhetorical unit of the RSHMM network. All the HMMs in our experiments are trained by HTK [16]. The extractive summarization system with rhetorical information is described in Figure 4.

$$\mathrm{K}(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j \parallel^2), \gamma > 0 \quad (8)$$

2.2. Approach 2: RSHMM++ for Lecture Speech Summarization

The entire RSHMM++ system consists of two modules: Rhetorical-State SVM (RSSVM) and Rhetorical-State HMM (RSHMM), as shown in Figure 5.

2.2.1. Extractive Summarization Using RSSVM

In extractive summarization for spontaneous speech, for a transcribed document D, a recognized sentence sequence $\{s_1, s_2, \ldots, s_N\}$, we want to find the sentences to be classified as summary sentences by using the salient sentence classification function $c()$.

In the probabilistic framework, extractive summarization task is equal to estimating $P(c(\vec{s}_n) = 1|\vec{D})$ of each sentence s_n, where $\vec{s}_n$ obtains from the acoustic and linguistic features for the sentence s_n; $\vec{D}$, the sentence feature vector sequence $\{\vec{s}_1, \vec{s}_2, \ldots, \vec{s}_N\}$, represents the transcribed document D. Using total probability theory, we deduce that:

$$P(c(\vec{s}_n) = 1|\vec{D}) = \sum_{i=1}^{R} P(c(\vec{s}_n) = 1, r(\vec{s}_n) = i|\vec{D}) \quad (9)$$

We map one sentence to one rhetorical unit by using the mapping function $r()$. $r(\vec{s}_n) = i$ means that s_n is a sentence of rhetorical unit i.

Considering that the hierarchical text plan of lecture speech contains R rhetorical units in total, we adopt $2R$-class SVM classifier with RBF kernel as the summarization part of RSHMM++: R summary sentence classes $\{summarySent\_r_i\}$, represented by dark balls in Figure 5; R non summary sentence classes $\{non\_summarySent\_r_i\}$, represented by color balls in Figure 5, $i = 1, 2, ..., R$. We

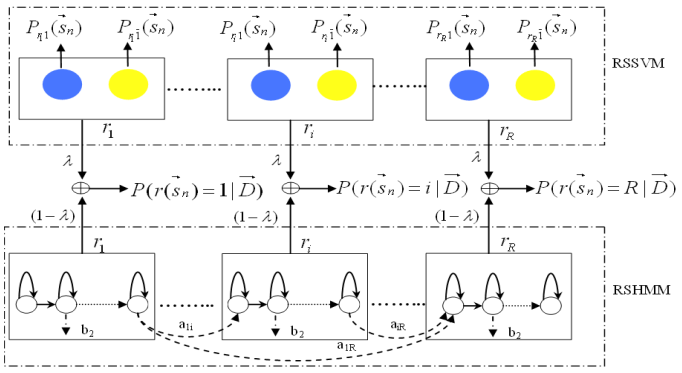

Fig. 5. Architecture of the hybrid RSHMM/RSSVM

estimate $P(c(\vec{s}_n) = 1, r(\vec{s}_n) = i|\vec{D})$ by output probability $P_{r_i1}(\vec{s}_n)$ of the class $summarySent\_r_i$.

We then classify those sentences which satisfy criterion 10 as summary sentences. We set prior probability $P(c(\vec{s}_n) = 1)$ to be the compression ratio of spoken documents.

$$\frac{P(c(\vec{s}_n) = 1|\vec{D})}{P(c(\vec{s}_n) = 1)} > 1 \tag{10}$$

2.2.2. *Rhetorical Structure Extraction Using Hybrid RSHMM/RSSVM*

According to criterion 10, we obtain an initial summary with flat structure. In this section, we describe how to extract rhetorical structure using hybrid RSHMM/RSSVM from speech data. Initial summaries are then transformed into hierarchically structured summaries.

We annotate each sentence of the given document as i^* which approximately maximizes $P(r(\vec{s}_n) = i|\vec{D})$.

$$i^* = \underset{i=1}{\overset{R}{\operatorname{argmax}}} P(r(\vec{s}_n) = i|\vec{D}) \tag{11}$$

We estimate $P(r(\vec{s}_n) = i|\vec{D})$ according to equation 12:

$$P(r(\vec{s}_n) = i|\vec{D}) = \lambda * P_1 + (1 - \lambda) * P_2 \tag{12}$$

where P_1 is the value of $P(r(\vec{s}_n) = i|\vec{D})$ estimated by RSSVM, shown in equation 13; P_2 is the value of $P(r(\vec{s}_n) = i|\vec{D})$ estimated by RSHMM, shown in equation 14; λ is combined factor, $\lambda \in [0, 1]$. We estimate this factor by cross validation training.

$$\begin{aligned} P_1 = P(c(\vec{s}_n) = 1, r(\vec{s}_n) = i|\vec{D}) \\ +P(c(\vec{s}_n) = 0, r(\vec{s}_n) = i|\vec{D}) \end{aligned} \tag{13}$$

According to the hierarchical text plan of lecture speech, shown in Figure 1, we train RSHMM by building R HMMs (i.e. r_1, r_2,..., and r_R) to represent respective rhetorical units. Each HMM is represented by three states and each of the state contains two Gaussian components. We trained each of the HMMs by performing Viterbi initialization and then followed by Baum-Welch re-estimation using the forward-backward algorithm.

Table 1. Acoustic/Prosodic Features

Feature Name	Feature Description
Duration	Duration of the sentence
Speaking Rate	Average syllable Duration
F0 (I-V)	F0 min,max,mean,slope, range
E (I-V)	Energy min, max, mean, slope, range

Table 2. Linguistic Features

Feature Name	Feature Description
Len (I-III)	Total no. of words in the current, previous and the next sentence
TFIDF, Sim	Total TFIDF and cosine similarity between the sentence and the entire document

$$P_2 = \sum_{k=1}^{3} P(s_n \in state_k \; of \; r_i|\vec{D}) \tag{14}$$

Finally, we annotate each summary sentence of the given document as i^* according to criterion 11 to produce a hierarchically structured summary.

3. EXPERIMENTAL SETUP

3.1. Acoustic and Linguistic Features

We represent each sentence s_n of the given speech by a feature vector $\vec{s}_n$ using acoustic and linguistic features same as in previous work [2, 3, 14]. The details of these features are listed in Table 1 and Table 2.

3.2. The Corpus

We collect the lecture speech corpus containing wave files of 111 presentations recorded from the NCMMSC2005 and NCMMSC2007 conferences. Power point slides and manual transcriptions are also collected. Each presentation lasts about 15 minutes on average. In our previous work, each presentation was automatically divided into on average 83 segment units. The ASR system runs in multiple passes and performs unsupervised acoustic model adaptation as well as unsupervised language model adaptation [14] with 70.3% recognition accuracy.

In our summarization experiments, we adopt 71 presentations from the lecture speech corpus. We use 62 presentations containing 5132 segment units as training data and the remaining 9 presentations of 736 segment units as test data.

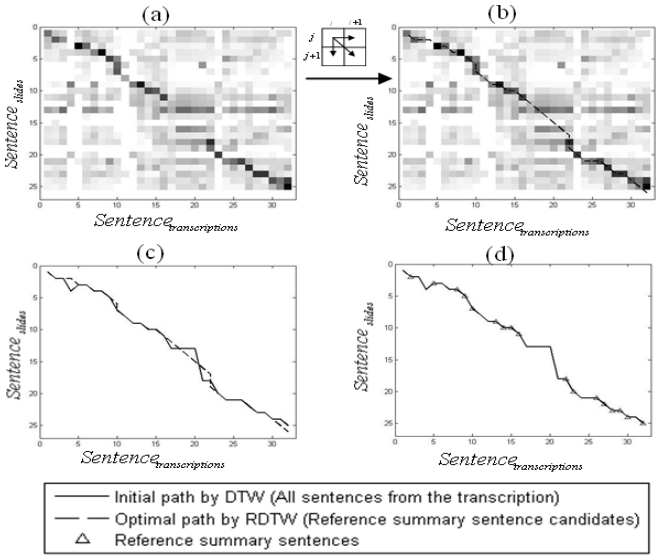

Fig. 6. Compiling an example reference summary

3.3. Reference Summaries

In this work, we compile reference summaries using Relaxed DTW between power point sentences and transcriptions. We assume that a good summary should consist of salient sentences from each of the rhetorical units (e.g. title, introduction, background, methodology, experiments, and conclusion sections in a conference presentation).

First we calculate the similarity scores matrix $Sim = (s_{ij})$, where $s_{ij} = similarity(Sent_{trans}[i], Sent_{slides}[j])$, between the sentences in the transcription and the sentences in the slides. We then obtain the distance matrix $Dist = (d_{ij})$, where $d_{ij} = 1 - s_{ij}$. We assume the number of sentence in the transcription is M (i=1,2,...M); and the number of sentences or segments in the slides is N (j=1,2,...,N), shown in Figure 6(a).

Next, we calculate the initial warp path which is the minimum-distance warp path $P = (p_1, p_2, ..., p_n, ..., p_N)$ by DTW algorithm, shown in Figure 6(b):

$$Dist(P) = \sum_{n=1}^{N} Dist(p_n) = \sum_{n=1}^{N} d_{i_n j_n} \tag{15}$$

Given that the speaker often does not follow the slide order strictly, we adopt Relaxed Dynamic Time Warp (RDTW) for finding the optimal path, according to the equation 16. The transcription sentences on this path are reference summary sentence candidates, shown in Figure 6(c).

$$\begin{cases} i_n^{opt} = i_n^{ini} \\ j_n^{opt} = \underset{j=j_n^{ini}-C}{\overset{j_n^{ini}+C}{\operatorname{argmin}}} d_{i_n^{opt},j} \end{cases} \tag{16}$$

We denote the initial path $(p_1^{ini}, p_2^{ini}, ..., p_n^{ini}, ..., p_N^{ini})$, where p_n^{ini} is represented by (i_n^{ini}, j_n^{ini}). We then obtain the optimal path $(p_1^{opt}, p_2^{opt}, ..., p_n^{opt}, ..., p_N^{opt})$, where p_n^{opt} is represented by (i_n^{opt}, j_n^{opt}). C is relaxation factor.

We then select the sentences i_n^{opt} of the transcription whose similarity scores of sentence pairs: (i_n^{opt}, j_n^{opt}) are higher than the pre-defined threshold as the final reference summary sentences, shown in Figure 6(d). Referred to Figure 1, we produce two versions of reference summaries with $R = 3$, which contains Introduction, Content, Conclusion&Discussion, and $R = 5$, which contains Outline, Background, Theory Description&Methodology, Experiment, Conclusion&Discussion.

4. EXPERIMENTAL RESULTS

We perform two sets of experiments: Experiment I for extractive summarization with $R = 3$ reference summaries and Experiment II for that with $R = 5$ reference summaries. We then build binary SVM classifier (one summary sentence class and one non-summary sentence class) without rhetorical information as our baseline system. We use ROUGE-L (summary-level Longest Common Subsequence) precision, recall and F-measure as main metrics in our experiments. The results are shown in Table 3.

We find that RSHMM++ achieves the best performance ROUGE-L F-measure 72.01%, a 9.78% absolute increase compared to the baseline when we use three-part-template reference summaries, and ROUGE-L F-measure 70.02%, a 2.27% absolute increase compared to the baseline when we use five-part-template reference summaries. Furthermore, we find that RSHMM++ consistently outperforms RSHMM. On average, the difference of summarization performance is absolute ROUGE-L F-measure 2.6%. That is to say, the inaccurate rhetorical information made by the HMM part of RSHMM++ is not carried over to the summarization process.

From Table 3, we also find that linguistic features are always more effective than acoustic features. The performance of the models which are created by only linguistic features cannot be improved much by adding acoustic information. This shows that, at least for lecture speech, what is said is more important than how it is said. This is probably due to the variable speaking styles of lecture speakers.

5. CONCLUSION

We have presented an enhanced Rhetorical-State Hidden Markov Model (RSHMM++) for extractive summarization of lecture speech. RSHMM++ can automatically decode the underlying rhetorical information in lecture speech and produce hierarchically structured summaries. Our RSHMM++ summarizer produced ROUGE-L F-measure of 72.01%, a 9.78% absolute increase in lecture speech summarization performance compared with the baseline system without using

Table 3. ROUGE-L F-measure of summarization performance on the manual sentence segmentation transcriptions using reference summaries

	Features	Baseline	RSHMM	RSHMM++
R=3	Li+Ac	.6223	.6867	***.7201***
	Li	.6223	.6823	.7160
	Ac	.5516	.5520	.5586
R=5	Li+Ac	.6775	.6815	***.7002***
	Li	.6770	.6810	.6946
	Ac	.5811	.5722	.5840

Ac: Acoustic features; Li: Linguistic features
Baseline: binary SVM classifier without rhetorical information;

rhetorical information. We also found that that RSHMM++ consistently outperforms RSHMM. Furthermore, we once again found that, at least for lecture speech, linguistic features are always more effective than acoustic features.

6. ACKNOWLEDGEMENT

This work was partially supported by CERG 612806 of the Hong Kong Research Grants Council.

7. REFERENCES

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

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

[3] J.J. Zhang, H.Y. Chan, P. Fung, and L. Cao, "A Comparative Study on Speech Summarization of Broadcast News and Lecture Speech," *Interspeech 2007 (Eurospeech)*, pp. 2781–2784, 2007.

[4] H.M. Wang Y.T. Chen, H.S. Chiu and B. Chen, "A Unified Probabilistic Generative Framework for Extractive Spoken Document Summarization," *Proc. Interspeech 2007*, pp. 2805–2808, 2007.

[5] C. Hori, S. Furui, R. Malkin, H. Yu, and A. Waibel, "Automatic speech summarization applied to English broadcast news speech," *Proc. ICASSP2002, Orlando, USA*, vol. 1, pp. 9–12, 2002.

[6] J.J. Zhang, H.Y. Chan, and P. Fung, "Improving lecture speech summarization using rhetorical information," *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, pp. 195–200, 2007.

[7] W.C. Mann and S.A. Thompson, *Rhetorical Structure Theory: A Theory of Text Organization*, University of Southern California, Information Sciences Institute, 1987.

[8] D. Tatar, E. Tamaianu-Morita, A. Mihis, and D. Lupsa, "Summarization by Logic Segmentation and Text Entailment," *Advances in Natural Language Processing and Applications*, pp. 15–26, 2008.

[9] Nemoto Y. AKITA, Y. and T. Kawahara, "PLSA-based topic detection in meetings for adaptation of lexicon and language model," *Proc. Interspeech 2007*, pp. 602–605, 2007.

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

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

[13] YT Chen et al., "Extractive Chinese Spoken Document Summarization Using Probabilistic Ranking Models," *Proc. ISCSLP*, 2006.

[14] P. Fung, R. Chan, and J.J Zhang, "Rhetorical-State Hidden Markov Models For Extractive Speech Summarization," *Acoustics, Speech, and Signal Processing, 2008. Proceedings.(ICASSP'08)*, pp. 4957–4960, 2008.

[15] 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.

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