ILP-BASED COMPRESSIVE SPEECH SUMMARIZATION WITH CONTENT WORD COVERAGE MAXIMIZATION AND ITS ORACLE PERFORMANCE ANALYSIS

Atsunori Ogawa, Tsutomu Hirao, Tomohiro Nakatani, and Masaaki Nagata

NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan

ABSTRACT

We propose an integer linear programming (ILP)-based compressive speech summarization method that maximizes the coverage of content words in a resultant summary. It is an unsupervised method and, under the designed constraints, it performs a single-step globally optimal summarization of a given long speech recording, which is decoded as a confusion network form of an automatic speech recognition (ASR) hypothesis sequence. It selects as many different content words as possible from the speech input that inevitably includes a high level of redundancy (e.g. the repetition of the same word) under a given length constraint. In experiments using a lecture speech corpus, we obtained higher summarization performance in terms of ROUGE scores than with a baseline extractive summarization method. We further conduct experimental analyses to obtain the oracle (upper bound) performance of the summarization methods. The analysis results show that the oracle performance is very high even though the ASR hypotheses include recognition errors. It is significantly higher than the system performance and, in addition, the oracle performance of the compressive method is significantly higher than that of the extractive method. These results confirm that our method is a promising approach.

Index Terms— Compressive speech summarization, integer linear programming (ILP), maximum coverage of content words, oracle (upper bound) performance

1. INTRODUCTION

Speech (speech-to-text) summarization is an essential technology for quickly reviewing the contents of a long speech recording, and it has been actively studied over many years [1–3]. In the basic framework of speech summarization, first automatic speech recognition (ASR) is performed on the target speech recording and then the obtained ASR hypotheses are summarized. Spontaneous speech inevitably includes high redundancy, e.g. the repetition of the same word. Therefore, the key issue in speech summarization is how to select as many different content words as possible from the speech input under a given length constraint [4, 5].

Many of the previous studies on speech summarization have focused on extractive methods, e.g. [6–16]. An extractive method, e.g. the maximum marginal relevance (MMR) method [17, 18], selects important utterances from the target speech recording and concatenates them to form a summary. It is basically a simple and thus robust method. However, it selects utterances in a greedy fashion at the utterance level, and thus it does not maximize the content word coverage in the resultant summary. This is one reason why methods other than extractive approaches should also be pursued [19].

As an alternative to the extractive approaches, each utterance is compressed in [20, 21]. With the method, a summarization score is defined, and words are selected from the target utterance using a dynamic programming (DP) framework so as to maximize the summarization score under the given length constraint. The selected word sequence corresponds to the compressed utterance. Each utterance can be accurately compressed with this method. However, a summary of the whole target speech recording, which achieves the maximum coverage of the content words, cannot be obtained by simply concatenating the compressed utterances.

As other examples, two-step summarization methods are employed in [19, 22]. With these methods, first the utterances are extracted and then each of the extracted utterances is compressed. An integer linear programming (ILP)-based compression framework [23] is employed in [19]. As with the DP-based framework [20–22], in the ILP-based framework, an objective function is defined and maximized under the designed constraints (including the given length constraint) to obtain a compressed utterance. With the ILP-based framework, an optimal summary can be obtained under the constraints, which is not always available with the DP-based framework [23]. A good summary can be obtained with these methods, however, it is not a globally optimal summary of the whole target speech recording that achieves maximum coverage of the content words, since the methods perform summarization in two separate steps.

In this paper, we apply an *ILP-based compressive speech summarization method that maximizes the content word coverage* to the ASR hypotheses. To the best of our knowledge, this is the first trial of compressive speech summarization. Compressive summarization has recently been actively studied in the written text domain, e.g. [24–26]. In contrast to the above described methods, a compressive method *jointly performs the extraction and the compression*. We derive an ILP formulation [23] of compressive summarization that maximizes the coverage of the content words [4, 5] and that is applicable to both the 1-best and confusion network (CN) [27] forms of ASR hypotheses. By applying our method to the whole target speech recording in a single step, we can obtain a globally optimal summary that achieves the maximum coverage of the content words under the given length constraint. We compared our method with an extractive method in terms of the ROUGE scores [28] using a lecture speech corpus. The experimental results confirm the effectiveness of our method. We also conducted experimental analyses to obtain the *oracle* (*upper bound*) *performance* of the speech summarization methods [29, 30]. This is the first trial to reveal the oracle summarization performance for ASR hypothesis inputs that include recognition errors. The analysis results confirm that our method is a promising approach.

2. RELATION TO PRIOR WORK

An ILP-based extractive method is applied to meeting recordings in [31] to obtain summaries that achieve the maximum coverage of the concepts (content words) with an extractive (i.e. utterance-level selection) framework. We employ this method as the baseline in our experiments. The CN form of the ASR hypotheses is used and their superior performance to the 1-best hypotheses is reported in [7,9,10]. We also compare the 1-best and CN forms of ASR hypotheses in our experiments.

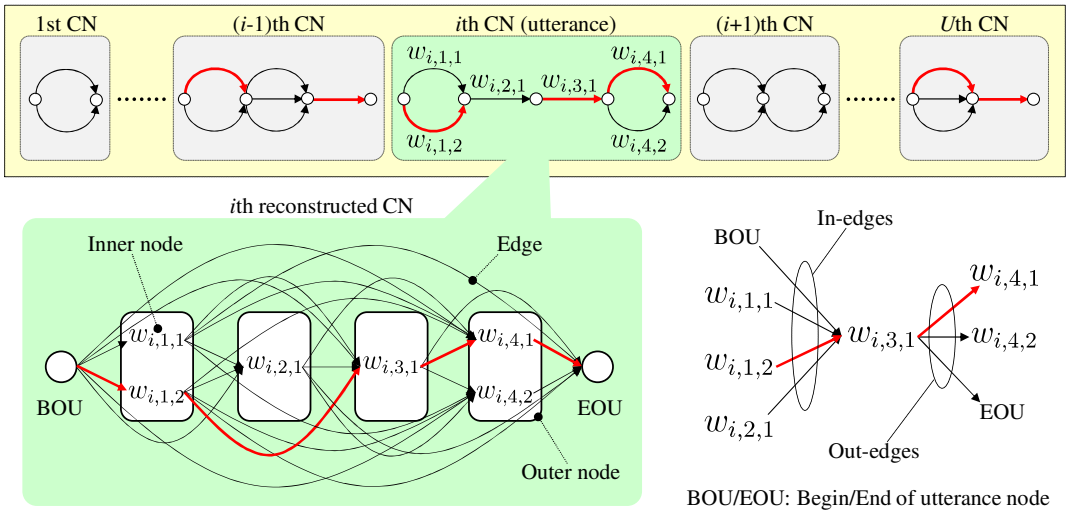

Fig. 1. (Top) CN form of ASR hypothesis sequence that is a decoding result of one whole (long) speech recording consisting of U utterances. One CN corresponds to the decoding result of one utterance. The red arcs (recognized words) in each CN are selected in the resultant summary. (Bottom left) The ith reconstructed CN. (Bottom right) In- and out-edges that are connected with the inner node (word) $w_{i,3,1}$.

In contrast to the above described unsupervised approaches, supervised approaches, especially neural network (NN)-based approaches, have begun to be studied, e.g. [32–36]. For example, recurrent NN (RNN) language models (LMs) [32], paragraph embedded features [33], and convolutional NNs (CNNs) [35], are used to select utterances in the extractive method. Abstractive summarization (headline generation) from short speech recordings is performed using RNNs [34] or connectionist temporal classification (CTC) model [36]. These methods seem to be promising. However, a huge amount of data is needed to train the NN models. In addition, it currently appears to be difficult to build an NN model that can receive a long speech recording (e.g. a whole lecture speech) and perform compressive summarization in a single step. Consequently, it is difficult for the NN model to maximize the coverage of the content words in the resultant summary. Therefore, in this paper, we focus on the ILP-based unsupervised summarization methods.

3. ILP-BASED COMPRESSIVE SPEECH SUMMARIZATION WITH COVERAGE MAXIMIZATION

We describe in detail our ILP-based compressive summarization method with coverage maximization for a CN form of ASR hypothesis sequence input. We also briefly mention our method for a 1-best sequence input, the baseline ILP-based extractive method [31], and methods for obtaining oracle summaries [29, 30].

3.1. Reconstruction of CN

As shown at the top of Fig. 1, our method receives a CN [27] form of ASR hypothesis sequence input, which is a decoding result of one whole (long) speech recording, and performs single-step compressive summarization for the given CN sequence while maximizing the coverage of the content words in the resultant summary. We reconstruct the structure of the given original ith CN as shown in the bottom left of Fig. 1. We redefine an arc (recognized word) in the original CN as an inner node (e.g. $w_{i,1,1}$) including the begin and end of utterance nodes (i.e. BOU and EOU) in the reconstructed CN. We also redefine a set of arcs in the same segment in the original CN as an outer node (e.g. $w_{i,1,1}$ and $w_{i,1,2}$) in the reconstructed CN. We then connect an inner node to all the other inner nodes with edges in a left-to-right fashion (without directly connecting BOU to EOU). Note that inner nodes included in the same outer node are not connected with each other. For example, $w_{i,1,1}$ is connected with edges to $w_{i,2,1}$, $w_{i,3,1}$, $w_{i,4,1}$, $w_{i,4,2}$, and EOU, but it is not connected with $w_{i,1,2}$. By repeating this procedure for each of the CN in the given original CN sequence, we can obtain a reconstructed CN sequence. In the following, we derive an ILP formulation [23] of our method using this reconstructed CN sequence as the input.

3.2. ILP Formulation for CN Sequence Input

Compressive speech summarization (joint extraction and compression) can be regarded as a path (a sequence of inner nodes and edges) selection problem in the given reconstructed CN sequence. We define the score of a path as the sum of the weights of inner nodes, those of edges, and the coverage of the content words included in the path. Then, we select a path that maximizes the score under a given length constraint as the best path, i.e. the resultant summary. This is a typical combinatorial optimization problem and an NP-hard problem. We derive an ILP formulation to solve this problem as shown in the Eqs. (1) to (12). In these equations, U is a set of CNs in the given reconstructed CN sequence, V_i is a set of outer nodes in the ith CN, $N_{i,j}$ is a set of inner nodes of the jth outer node in the ith CN, and W is a set of the content words in U.

Equation (1) is the objective function that we define to find the best path. $f_{i,j,k}$ is the significance score for the inner node $w_{i,j,k}$ (the kth inner node of the jth outer node in the ith CN) defined as

$$f_{i,j,k}=\text{conf}(w_{i,j,k})+\text{tfidf}(w_{i,j,k}),$$

where conf$(\cdot)$ returns the confidence score for an inner node (i.e. the reliability of a recognized word) and tfidf$(\cdot)$ returns the tfidf (= tf $\times$ idf) score for the inner node. The confidence score (probability) is attached to each inner node in the CN through consensus decoding [27]. We obtain the tf score (count) of an inner node by summing the confidence scores of the inner node in U as with [7, 9, 10]. We estimate the idf score of each inner node using text data (in our experiments, this is the data used to train ASR models). $n_{i,j,k}$ is a binary indicator, and $n_{i,j,k}=1$ denotes that $w_{i,j,k}$ is included in the best path. $g^{i,t,q}_{i,s,p}$ is the edge score between the pth inner node of the

sth outer node and the qth inner node of the tth outer node in the ith CN. We use the word bigram score as this edge score as

$$g_{i,s,p}^{i,t,q}=P(w_{i,t,q}|w_{i,s,p}).$$

We train a bigram LM employing the text data that is used to estimate the idf scores of inner nodes (i.e. the data used to train the ASR models). $e_{i,s,p}^{i,t,q}$ is a binary indicator, and $e_{i,s,p}^{i,t,q}=1$ denotes that the edge is included in the best path. α and β are parameters for balancing the above described significance and bigram scores. z_h is a binary indicator, and $z_h=1$ denotes that the hth content word in U is included in the best path, and $z_h=0$ denotes otherwise. With this last term of Eq. (1), our method attempts to include (cover) as many different content words (in our experiments, nouns, verbs, and adjectives) as possible in the best path. This term denotes the maximum coverage constraint.

Equations (2) to (12) represent the constraints. Equation (2) ensures that there are fewer than L characters in the best path where $l_{i,j,k}$ denotes the number of characters (length) of the inner node (word) $n_{i,j,k}$. In an experimental evaluation, L is usually determined based on the number of characters included in the reference transcription of the target speech recording and the given length constraint (or the compression ratio). Equation (3) ensures that at most one inner node can be selected from an outer node. Equation (4) represents the constraint between an inner node and in-edges that are connected to the inner node. Similarly, Eq. (5) represents a constraint between an inner node and out-edges that are connected to the inner node. These two equations ensure that, when an inner node is selected in the best path, one in-edge and one out-edge that are connected to the selected inner node must be selected. For example, as shown in the bottom right of Fig. 1, the inner node $w_{i,3,1}$ is selected and, consequently, one in-edge connected from $w_{i,2,1}$ and one out-edge connected to $w_{i,4,1}$ are selected. To avoid generating short and fragmented compressed utterances, Eqs. (6) to (8) ensure that the number of inner nodes included in a compressed utterance exceeds K.

We can solve the above defined ILP-based optimization problem exactly (i.e. we can find exactly the best path) using a solver, e.g. the CPLEX optimizer [37]. After solving the problem, we can obtain the best path, i.e. the resultant summary, by collecting the inner nodes according to $n_{i,j,k}=1$. For example, in Fig. 1, a sequence that consists of the three inner nodes (words), $w_{i,1,2}$, $w_{i,3,1}$, and $w_{i,4,1}$, is included in the resultant summary as a compressed utterance. The resultant summary is a globally optimal summary of the target speech recording under the above described constraints where the maximum coverage of the content words is achieved.

3.3. 1-best Input, Extractive Method, and Oracle Summaries

We can obtain an ILP formulation for a 1-best ASR hypothesis sequence input as a simplified version of that for the CN sequence input. The 1-best input can be regarded as the CN input in which all outer nodes have only one inner node, i.e. $|N_{i,j}|=1$ for all i and j. Therefore, we can obtain the ILP formulation for the 1-best input from Eqs. (1) to (12) by removing all the sum operations that relate to $|N_{i,*}|$ and, consequently, removing the indices k, p, and q from all the variables. The CN input clearly has an advantage over the 1-best input, since we can also select inner nodes (words) from words that are ranked below the 2nd-best in the CN.

For the baseline extractive method, we employ an ILP formulation similar to that described in [31]. We use the 1-best input for the extractive method. Note that, as with our method for the CN input, our method for 1-best input and the extractive method for the 1-best input also employ the maximum coverage constraint of the content words.

We obtain the extractive and compressive oracle summaries based on the methods proposed in [29] and [30], respectively, using ASR hypotheses and reference summaries. This is the first trial to reveal the oracle (upper bound) summarization performance for the ASR hypotheses that include recognition errors. The oracle summaries are obtained so as to maximize the ROUGE scores [28]. They are metrics that are widely-used to evaluate the quality of obtained system summaries by comparison with the reference summaries based on the co-occurrence statistics of word unigrams (ROUGE-1), bigrams (ROUGE-2), and skip-grams plus unigrams (ROUGE-SU, typically, ROUGE-SU4 that allows four word skips at maximum).

$$\begin{aligned}\text{maximize} \sum_{i=1}^{|U|}\Bigg(&\alpha\sum_{j=1}^{|V_i|}\sum_{k=1}^{|N_{i,j}|} f_{i,j,k}n_{i,j,k} \\ &+\beta\sum_{s=1}^{|V_i|-1}\sum_{p=1}^{|N_{i,s}|}\sum_{t=s+1}^{|V_i|}\sum_{q=1}^{|N_{i,t}|} g_{i,s,p}^{i,t,q}e_{i,s,p}^{i,t,q}\Bigg) \\ &+\sum_{h=1}^{|W|} z_h, \end{aligned} \tag{1}$$

$$\text{subject to} \sum_{i=1}^{|U|}\sum_{j=1}^{|V_u|}\sum_{k=1}^{|N_{i,j}|} \ell_{i,j,k}n_{i,j,k} \le L, \tag{2}$$

$$\forall i,j: \sum_{k=1}^{|N_{i,j}|} n_{i,j,k} \le 1, \tag{3}$$

$$\forall i,j,k,s,p: \sum_{s=1}^{j-1}\sum_{p=1}^{|N_{i,s}|} e_{i,s,p}^{i,j,k} - n_{i,j,k} = 0, \tag{4}$$

$$\forall i,j,k,t,q: \sum_{t=j+1}^{|V_i|}\sum_{q=1}^{|N_{i,t}|} e_{i,j,k}^{i,t,q} - n_{i,j,k} = 0, \tag{5}$$

$$\forall i: \frac{\sum_{j=1}^{|V_i|}\sum_{k=1}^{|N_{i,j}|} n_{i,j,k}}{K} \ge a_{i,1}, \tag{6}$$

$$\forall i: 1-\frac{\sum_{j=1}^{|V_i|}\sum_{k=1}^{|N_{i,j}|} n_{i,j,k}}{|V_i|} \ge a_{i,2}, \tag{7}$$

$$\forall i: a_{i,1}+a_{i,2}=1, \tag{8}$$

$$\forall i,j,k: n_{i,j,k} \in \{0,1\}, \tag{9}$$

$$\forall i,s,t,p,q: e_{i,s,p}^{i,t,q} \in \{0,1\}, \tag{10}$$

$$\forall h: z_h \in \{0,1\}, \tag{11}$$

$$\forall i: a_{i,*} \in \{0,1\}. \tag{12}$$

4. EXPERIMENTS

We conducted experiments to evaluate the ILP-based speech summarization methods with coverage maximization described in Section 3 using the corpus of spontaneous Japanese (CSJ) [38], which is a large scale speech corpus of academic lectures. Our compressive methods, which use the 1-best or CN form of the ASR hypothesis inputs (hereafter, referred to as *Comp. 1-best* and *Comp. CN*, respectively) are compared with the baseline extractive method, which uses the 1-best inputs (*Extr. 1-best*) [31]. To confirm the effectiveness of the coverage maximization of the content words, we also evaluated

Comp. CN, which does not use the coverage constraint, i.e. the last term in Eq. (1) (*Comp. CN w/o coverage*). In addition, we conducted experiments to reveal the oracle (upper bound) performance of the summarization methods [29,30] when they are applied to ASR hypotheses including recognition errors. We used the CPLEX optimizer [37] to perform summarization (i.e. to solve the ILP-based optimization problems).

4.1. Experimental Settings

The training data consisted of 250 hours of speech, which includes 198k utterances and 3.4M words. Using this data, we trained a CNN acoustic model and a trigram LM, which were used to perform ASR. We also trained a bigram LM and estimated idf scores for each word in the vocabulary, which were used to perform summarization. The vocabulary size was set at 31k. The evaluation data consisted of 100 lectures as shown in Table 1. We performed ASR on this data with a weighted finite state transducer (WFST)-based one-pass speech recognizer [39] using the CNN acoustic model and the trigram LM described above, and we obtained the 1-best and CN forms of the ASR hypotheses for each utterance in the data. We obtained the confidence and tf scores for each recognized word in the ASR hypotheses, which were used to perform summarization.

We performed single-step summarization with the four methods described above for each obtained ASR hypothesis sequence that corresponds to a whole lecture speech (12 min length on average) using the obtained features, i.e. the confidence, tfidf (= tf $\times$ idf), and bigram LM scores. We set the compression ratio at 10% at the character level by adjusting the number of characters L in Eq. (2). Based on the results of preliminary experiments, we set α and β in Eq. (1) at 0.6 and 1.5, respectively, and K in Eq. (6) at 20 so that one compressed utterance consisted of at least 20 words (about 30 characters). We evaluated the resultant system summaries using *ROUGE-1*, *-2*, and *-SU4* scores [28] (see Section 3.3). All types of words were taken into account when calculating these scores. As for ROUGE-1, we also employed the version that takes only the content words (the nouns, verbs, and adjectives) into account (referred to as *ROUGE-1CW*). Three or four reference summaries were made by different human subjects for each lecture with a 10% compression ratio at the character level. We thus calculated a ROUGE score for the resultant summary by averaging the ROUGE scores obtained using these reference summaries. We also counted how many different content words on average were included in a resultant summary.

We obtained oracle summaries based on the methods proposed in [29, 30] (see Section 3.3). We used the ROUGE-2 score in the objective function and thus the oracle summaries were obtained so as to maximize the ROUGE-2 scores.

4.2. Experimental Results

Table 2 shows the summarization performance of the four methods (systems). The absolute values of the obtained ROUGE scores are reasonable compared with those reported in the literature. Comparing the results of Extr. 1-best and Comp. 1-best, we can confirm the superiority of the compressive method over the extractive method. Comp. CN slightly but consistently outperforms Comp. 1-best. From these results, we can confirm the effectiveness of using CNs as the form of ASR hypotheses as reported in [7, 9, 10]. Analyzing the summaries obtained with Comp. CN, 18.6% of the words are selected from recognized words that are ranked below the 2nd-best in the input CNs. Comparing the results of Comp. CN and Comp. CN w/o coverage, we can confirm the effectiveness of introducing the maximum coverage constraint of the content words. Its effect is especially (reasonably) noticeable in the ROUGE-1CW score. Comp. CN includes 98.6 on average of different content words in the resultant summary, whereas Comp. CN w/o coverage only includes 75.8 different content words (see the following discussion).

Table 1. 100 lecture speech data used for evaluation.

Total / average length	20 hours / 12 min
Total / average number of utterances	13963 / 140
Total / average number of words	249537 / 2495
1-best word error rate	14.5%

Table 2. System performance in terms of the four ROUGE scores. #CW shows avg. # of different content words included in a summary.

Method	ROUGE-1	-1CW	-2	-SU4	#CW
Extr. 1-best	0.595	0.433	0.225	0.303	106.2
Comp. 1-best	0.621	0.451	0.267	0.356	96.8
Comp. CN	**0.624**	**0.459**	**0.270**	**0.360**	98.6
Comp. CN w/o coverage	0.616	0.429	0.263	0.357	75.8

Table 3. Oracle (upper bound) performance.

Method	ROUGE-1	-1CW	-2	-SU4	#CW
Extr. 1-best	0.754	0.637	0.424	0.460	83.0
Comp. 1-best	0.805	0.695	0.579	0.584	86.3
Comp. CN	**0.832**	**0.737**	**0.632**	**0.623**	87.4

Table 3 shows the oracle (upper bound) performance when the summarization methods are applied to the ASR hypothesis inputs. From these results, we can confirm that the oracle performance is very high even though the ASR hypotheses include recognition errors (the 1-best word error rate is 14.5%). It is significantly higher than the system performance. In addition, we can confirm again the superiority of the compressive method over the extractive method and that of Comp. CN over Comp.1-best. These results confirm that our methods, especially Comp. CN, are promising. However, they also indicate that we need to improve our methods (e.g. by introducing more effective features and constraints) since there is a large score gap between the system and oracle summaries.

A comparison of a system summary and the corresponding oracle summary shows that the system summary includes more different content words than the oracle summary. However, the ROUGE-1CW score of the system summary is lower than that of the oracle summary. The number of different content words included in an oracle summary shows the upper bound of the number of different content words that can be *correctly* included in a summary. This means that, even though the coverage constraint performs its role steadily as described above, *incorrect* content words are also included in a system summary (especially, in an Extr. 1-best summary). We also need to develop a method for selecting the content words more accurately.

5. CONCLUSION AND FUTURE WORK

We have applied ILP-based compressive summarization with content word coverage maximization to ASR hypotheses for the first time. We confirmed experimentally the superiority of the compressive method over the extractive method and the effectiveness of considering the maximum coverage of the content words. We also confirmed from oracle (upper bound) performance analyses that our method is promising but still has considerable room for improvement.

Future work will include the introduction of dependency structures to improve the grammatical correctness of the resultant summaries. We also plan to leverage the NN-based techniques. We will start with the simple utilization of the NN-based models and features in our framework as successfully utilized in [32, 33].

6. REFERENCES

[1] K. McKeown, J. Hirshberg, M. Galley, and S. Maskey, "From text to speech summarization," in *Proc. ICASSP*, 2005, pp. V997–V1000.

[2] S. Furui, "Recent advances in automatic speech summarization," in *Proc. SLT*, 2006, pp. 16–21.

[3] L.-S. Lee, J. Glass, H.-Y. Lee, and C.-A. Chan, "Spoken content retrieval - Beyond cascading speech recognition with text retrieval," *IEEE/ACM Trans. on ASLT*, vol. 23, no. 9, pp. 1389–1420, Sept. 2015.

[4] D. Gillick, B. Favre, and D. Hakkani-Tür, "The ICSI summarization system at TAC 2008," in *Proc. Text Analysis Conference (TAC)*, 2008.

[5] H. Takamura and M. Okumura, "Text summarization model based on maximum coverage problem and its variant," in *Proc. EACL*, 2009, pp. 781–789.

[6] N. Garg, B. Favre, K. Reidhammer, and D. Hakkani-Tür, "ClusterRank: A graph based method for meeting summarization," in *Proc. Interspeech*, 2009, pp. 1499–1502.

[7] S.-H. Lin and B. Chen, "Improved speech summarization with multiple-hypothesis representations and Kullback-Leibler divergence measures," in *Proc. Interspeech*, 2009, pp. 1847–1850.

[8] S. Xie, D. Hakkani-Tür, B. Favre, and Y. Liu, "Integrating prosodic features in extractive meeting summarization," in *Proc. ASRU*, 2009, pp. 387–391.

[9] S. Xie and Y. Liu, "Using confusion networks for speech summarization," in *Proc. NAACL*, 2010, pp. 46–54.

[10] S. Xie and Y. Liu, "Using n-best lists and confusion networks for meeting summarization," *IEEE Trans. on ASLP*, vol. 19, no. 5, pp. 1160–1169, July 2011.

[11] B. Chen and S.-H. Lin, "A risk-aware modeling framework for speech summarization," *IEEE Trans. on ASLP*, vol. 20, no. 1, pp. 211–222, Jan. 2012.

[12] B. Chen, H.-C. Chang, and K.-Y. Chen, "Sentence modeling for extractive speech summarization," in *Proc. ICME*, 2013.

[13] F. Koto, S. Sakti, G. Neubig, T. Toda, M. Adriani, and S. Nakamura, "The use of semantic and acoustic features for open-domain TED talk summarization," in *Proc. APSIPA*, 2014.

[14] S.-H. Liu, H.-S. Lee, H.-T. Hung, K.-Y. Chen, B. Chen, H.-M Wang, H.-C. Yen, and W.-L. Hsu, "Incorporating proximity information in relevance language modeling for extractive speech summarization," in *Proc. APSIPA*, 2015, pp. 401–407.

[15] K.-Y. Chen, S.-H. Liu, Berlin Chen, and H.-M. Wang, "Improved spoken document summarization with coverage modeling techniques," in *Proc. ICASSP*, 2016, pp. 6010–6014.

[16] M.H. Bokaei, H. Sameti, and Y. Liu, "Summarizing meeting transcripts based on functional segmentation," *IEEE/ACM Trans. on ASLP*, vol. 24, no. 10, pp. 1831–1841, Oct. 2016.

[17] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proc. SIGIR*, 1998, pp. 335–336.

[18] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell, "Summarizing text documents: Sentence selection and evaluation metrics," in *Proc. SIGIR*, 1999, pp. 121–128.

[19] F. Liu and Y. Liu, "Towards abstractive speech summarization: Exploring unsupervised and supervised approaches for spoken utterance compression," *IEEE Trans. on ASLP*, vol. 21, no. 7, pp. 1469–1480, July 2013.

[20] C. Hori and S. Furui, "Automatic speech summarization based on word significance and linguistic likelihood," in *Proc. ICASSP*, 2000, pp. 1579–1582.

[21] C. Hori and S. Furui, "Speech summarization: An approach through word extraction and a method for evaluation," *IEICE Trans. on Inf. & Syst.*, vol. E87-D, no. 1, pp. 15–25, Jan. 2004.

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

[23] J. Clarke and M. Lapata, "Global inference for sentence compression an integer linear programming approach," *Journal of Artificial Intelligence Research*, vol. 31, pp. 399–429, 2008.

[24] A.F.T. Martins and N.A. Smith, "Summarization with a joint model for sentence extraction and compression," in *Proc. NAACL HLT Workshop on Integer Linear Programming for NLP*, 2009, pp. 1–9.

[25] D. Gillick and B. Favre, "A scalable global model for summarization," in *Proc. NAACL HLT Workshop on Integer Linear Programming for NLP*, 2009, pp. 10–18.

[26] T. Berg-Kirkpatrick, D. Gillick, and D. Klein, "Jointly learning to extract and compress," in *Proc. HLT*, 2011, pp. 481–490.

[27] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," *Computer Speech and Language*, vol. 14, no. 4, pp. 373–400, Oct. 2000.

[28] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Workshop on Text Summarization Branches Out*, 2004, pp. 74–81.

[29] T. Hirao, M. Nishino, J. Suzuki, and M. Nagata, "Enumeration of extractive oracle summaries," in *Proc. EACL*, 2017, pp. 386–396.

[30] T. Hirao, M. Nishino, and M. Nagata, "Oracle summaries of compressive summarization," in *Proc. ACL*, 2017, pp. 275–280.

[31] D. Gillick, K. Riedhammer, B. Favre, and D. Hakkani-Tür, "A global optimization framework for meeting summarization," in *Proc. ICASSP*, 2009, pp. 4769–4772.

[32] K.-Y. Chen, S.-H. Liu, B. Chen, H.-M. Wang, E.-E. Jan, W.-L. Hsu, and H.-H. Chen, "Extractive broadcast news summarization leveraging recurent neural network language modeling techniques," *IEEE/ACM Trans. on ASLP*, vol. 23, no. 8, pp. 1322–1334, Aug. 2015.

[33] K.-Y. Chen, K.-W. Shih, S.-H. Liu, B. Chen, and H.-M. Wang, "Incorporating paragraph embeddings and density peaks clustering for spoken document summarization," in *Proc. ASRU*, 2015, pp. 207–214.

[34] L.-C. Yu, H.-Y. Lee, and L.-S. Lee, "Abstractive headline generation for spoken content by attentive recurrent neural networks with ASR error modeling," in *Proc. SLT*, 2016, pp. 151–157.

[35] C.-I. Tsai, H.-T. Hung, K.-Y. Chen, and B. Chen, "Extractive speech summarization leveraging convolutional neural network techniques," in *Proc. SLT*, 2016, pp. 158–164.

[36] B.-R. Lu, F. Shyu, Y.-N. Chen, H.-Y. Lee, and L.-S. Lee, "Order-preserving abstractive summarization for spoken content based on connectionist temporal classification," in *Proc. Interspeech*, 2017, pp. 2899–2903.

[37] IBM, "CPLEX optimizer," https://www.ibm.com/analytics/cplex-optimizer.

[38] K. Maekawa, "Corpus of spontaneous Japanese: its design and evaluation," in *Proc. Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, 2003, pp. 7–12.

[39] T. Hori, C. Hori, Y. Minami, and A. Nakamura, "Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition," *IEEE Trans. on ASLP*, vol. 15, no. 4, pp. 1352–1365, May 2007.