Recent Advances in Automatic Speech Summarization.

Conference Paper · January 2007

DOI: 10.1109/SLT.2006.326806 · Source: DBLP

Sadaoki Furui
Tokyo Institute of Technology

Recent Advances in Automatic Speech Summarization

Sadaoki Furui
Department of Computer Science
Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
furui@cs.titech.ac.jp

Abstract

Speech summarization technology, which extracts important information and removes irrelevant information from speech, is expected to play an important role in building speech archives and improving the efficiency of spoken document retrieval. However, speech summarization has a number of significant challenges that distinguish it from general text summarization. Fundamental problems with speech summarization include speech recognition errors, disfluencies, and difficulties of sentence segmentation. Typical speech summarization systems consist of speech recognition, sentence segmentation, sentence extraction, and sentence compaction components. Most research up to now has focused on sentence extraction, using LSA (Latent Semantic Analysis), MMR (Maximal Marginal Relevance), or feature-based approaches, among which no decisive method has yet been found. Proper sentence segmentation is also essential to achieve good summarization performance. How to objectively evaluate speech summarization results is also an important issue. Several measures, including families of SumACCY and ROUGE measures, have been proposed, and correlation analyses between subjective and objective evaluation scores have been performed. Although these measures are useful for ranking various summarization methods, they do not correlate well with human evaluations, especially when spontaneous speech is targeted.

1. Introduction

Spoken document retrieval is one of the most important applications of automatic speech recognition (ASR) technology. Effective speech summarization is expected to reduce the time required to review speech documents and to improve the efficiency of document retrieval. Thus, speech summarization technology is expected to play an important role in building various speech archives, including those for broadcast news, lectures, presentations, and interviews. Spoken document summarization results can be presented as either text or speech (Furui et al., 2004).

Compared to speech read from a text, such as in broadcast news utterances, ASR accuracy for spontaneous speech is still limited (Shinozaki & Furui, 2004). Recognition errors cause transcriptions obtained from spontaneous speech to include irrelevant or incorrect information. In addition, spontaneous speech is ill-formed and usually includes redundant information such as disfluencies, fillers, repetitions, repairs, and word fragments. Direct transcriptions are therefore not always useful, and processes for extracting important information and removing incorrect information are necessary for transcribing spontaneous speech for useful purposes. Automatic speech summarization is one approach to accomplishing this goal.

Summarization and question answering (QA) perform similar tasks, in that they both map an abundance of information to a (much) smaller unit, which is then returned to the user (Zechner, 2003). Therefore, speech summarization research will help the advancement of QA systems targeting speech documents. By condensing important points from long presentations and lectures, and presenting them in a summary speech, systems can provide the listener with a valuable means of absorbing more information in a much shorter period of time.

This is a revised version of the paper with the same title, published in the proceedings of the IEEE/ACL 2006 Workshop on Spoken Language Technology held on December 10-13, 2006, in Aruba.

Although there is considerable research activity in text summarization, there has been less work done in speech summarization. Speech summarization poses a number of significant challenges that distinguish it from general text summarization. Applying text-based technologies (Mani & Maybury, 1999) to speech is not always viable and often the systems are not equipped to capture speech specific phenomena (Christensen et al., 2003; Kolluru et al., 2003). One fundamental problem with speech summarization is that target documents contain speech recognition errors and disfluencies (Murray et al., 2005a; Murray et al., 2005b; Valenza et al., 1999). Summarizing spontaneous speech is thus substantially different from text summarization. Christensen et al. (2004) provide evidence that more spontaneous parts of broadcast news (e.g. interviews) are less amenable to standard text summarization techniques.

2. Speech-to-text & speech-to-speech summarization

Speech summarization results can be presented as either text or speech. The former method has advantages in that: a) the documents can be easily looked through; b) those parts of the documents which are interesting for users can be easily extracted; and c) information extraction and retrieval techniques can be easily applied to the documents. However, presenting summarization results in text format has disadvantages in that: a) wrong information due to speech recognition errors cannot be avoided; and b) emotional content and other prosodic information present in speech is very difficult to represent in a non-acoustic format. On the other hand, the latter method does not have such disadvantages and it can preserve all the acoustic information included in the original speech.

Methods for presenting summaries by speech can be classified into two categories; a) simply presenting concatenated speech segments that are extracted from original speech, or b) synthesizing summarized text using a speech synthesizer. Since state-of-the-art speech synthesizers still cannot produce completely natural speech, the former method can easily produce better quality summarizations, and it does not have the problem of synthesizing wrong messages due to speech recognition errors. The major problem in using extracted speech segments is how to avoid unnatural noisy sound caused by the concatenation.

Since most of the research being conducted are targeting speech-to-text summarization, this paper focuses on this category.

3. Summarization methods

3.1. Sentence extraction-based methods

Recent work on spoken language summarization in unrestricted domains has focused almost exclusively on Broadcast News (Garofolo et al., 1999; Valenza et al., 1999). Koumpis and Renals (2000) have investigated the transcription and summarization of voice mail speech. Summarization of spontaneous speech in face-to-face situations using a mobile translation system has been attempted by Alexandersson and Poller (1998). Zechner and Waibel (2000) have investigated how the accuracy of summaries changes when methods for word error rate reduction are applied in summarizing conversations from television shows. Murray et al. (2005a; 2005b) have investigated summarization of meeting utterances. Among various techniques investigated for text summarization, most of the previous research on spoken language summarization have relied on extractive approaches, using relatively long units, such as sentences or speaker turns, as minimal units for summarization.

3.1.1. LSA-based method – 1 (Original method)

Sentence extraction using Latent Semantic Analysis (LSA), based on the Singular Value Decomposition (SVD), is one potential technique for text summarization (Gong & Liu, 2001). The SVD, one of the vector-space approaches, semantically clusters content words and sentences, and thus derives a latent semantic structure. The original $m \times n$ word-sentence matrix $\boldsymbol{A}$ (where without loss of generality it can be assumed that $m \geq n$), whose elements A_{ij} represent the (weighted) term frequency of word i in sentence j, is projected to a reduced dimensional representation. The word-sentence matrix is decomposed as follows:

$$\boldsymbol{A} = \boldsymbol{USV}^{\mathrm{T}} \tag{1}$$

where $\boldsymbol{U}$ is an $m \times n$ matrix of left-singular vectors, $\boldsymbol{S}$ is an $n \times n$ diagonal matrix of singular values sorted in descending order, and $\boldsymbol{V}$ is an $n \times n$ matrix of right-singular vectors.

Each singular vector, a row of $\boldsymbol{V}^{\mathrm{T}}$, represents a salient topic, with the columns representing sentences from the document. The singular vector with the largest corresponding singular value represents the topic that is hypothesized to be the most salient in the speech document. Therefore, a fixed number of singular vectors having relatively large singular values are selected for summarization. For each singular vector, that is each row in $\boldsymbol{V}^{\mathrm{T}}$, the sentence having the largest score is extracted as an important sentence. In this way, extracted sentences best describe the topics represented by the singular vectors and are semantically different from each other. When a desired summary length is given, the singular vectors having relatively large singular values are incrementally selected until the target summary length is reached.

3.1.2. LSA-based method – 2 (Revised method)

Two drawbacks to the previous method are that dimensionality is tied to summary length and that good sentence candidates may not be chosen if they do not “win” in any dimension (Steinberger & Jezek, 2004). In addition, when the singular vectors are selected incrementally, as the number of vectors being selected increases, the chances that non-relevant topics get included in a summary also increases. To address these problems, sentence extraction using dimension reduction based on LSA has been proposed (Hirohata et al., 2003; Hirohata et al., 2006; Murray et al., 2005a; Murray et al., 2005b; Steinberger & Jezek, 2004). In this method, a fixed number of sentences having relatively large sentence scores in the reduced dimensional space are selected. Hirohata et al. (2003; 2006) proposed the following LSA sentence score, which corresponds to the length of each sentence vector weighted by those singular values which correspond to its component parts:

$$Sc^{LSA}(i) = \sqrt{\sum_{k=1}^{n} v(i,k)^2 * \sigma(k)^2}, \tag{2}$$

where $v(i,k)$ is the kth element word of the ith sentence vector, $\sigma(k)$ is the corresponding singular value, and n is the number of dimensions of the new space. Using this method, extracted sentences not only describe the significant topics but also have a latent relationship between each other.

Hirohata et al. evaluated this method by applying it to the task of making abstracts from spontaneous presentations. Sentence location information, which has been used for text summarization, was combined to extract important sentences from the introduction and conclusion segments of each presentation. Locations of the introduction and conclusion segments were estimated based on the Hearst method (Hearst, 1997) using sentence cohesiveness. They also investigated the combination of confidence measures and linguistic likelihood to effectively extract sentences with fewer recognition errors. Experimental results showed that the dimension-reduction-based method incorporating sentence location information, the confidence measure, and linguistic likelihood achieved the best automatic speech summarization performance when requiring a 10% summarization ratio. The summarization ratio is defined by the ratio of the number of words in the summary to that in the original speech.

Murray et al. (2005a; 2005b) have shown the effectiveness of this dimension-reduction-based method in summarizing meeting recordings.

3.1.3. MMR-based method

The Maximal Marginal Relevance (MMR) method (Carbonell & Goldstein, 1998) uses the vector-space model of text retrieval and is particularly applicable to query-based and multi-document summarization. The MMR algorithm chooses sentences via a weighted combination of their relevance to a query (or for generic summaries, their general relevance) and their redundancy with sentences that have already been extracted, both derived using cosine similarity. The MMR score Sc^{MMR} (i) for a given sentence S_i in the document is given by

$$Sc^{MMR}(i) = \lambda(Sim(S_i, D)) - (1-\lambda)(Sim(S_i, Summ)), \tag{3}$$

where D is the average document vector, *Summ* is the average vector from the set of sentences already selected, and λ trades off between relevance and redundancy. *Sim* is the cosine similarity between two documents. In this implementation of MMR, the weight λ is annealed, so that relevance is emphasized when the summary is still short, and as the summary grows longer the emphasis is increasingly shifted towards minimizing redundancy.

Zechner (2003) reported experiments on the summarization of spoken multiparty dialogues, using an approach based on MMR, with the addition of automatic speech disfluency removal, sentence boundary marking, and question-answer pair detection.

3.1.4. Feature-based method

Feature-based classification approaches have been widely used in text and speech summarization. Kupiec et al. (1995) combined textual and prosodic features, using Gaussian mixture models for the extracted and non-extracted classes. The prosodic features were the mean and standard deviation of F0, energy, and duration, all estimated and normalized at the word-level, then averaged over the utterance. The two lexical features were the average and the maximum TF-IDF score for the utterance.

Steinberger et al. (2004) combined the LSA sentence scores described in Subsection 3.1.2 to complement the six features used by Kupiec et al., and showed that the LSA sentence score is beneficial in determining sentence importance.

Kong et al. (2006) proposed using topic significance scores and term entropy obtained through Probabilistic Latent Semantic Analysis (PLSA) to determine important sentences, and showed the effectiveness in Chinese broadcast news summarization.

Sameer et al. (2005) evaluated the usefulness of lexical, prosodic, structural and discourse features in selecting extractive summaries from news broadcasts. The lexical features include counts of named entities (person, organization and place names) for each sentence. The acoustic/prosodic features include speaking rate, F0 features (minimum, maximum, mean, range, and slope), log-energy features (minimum, maximum, mean, and slope), and sentence duration. Normalized features were produced by dividing each feature by the average of the feature's values for each speaker. The structural features include normalized/sentence position and speaker type (reporter or not). The discourse features include the number of new noun stems in each sentence, showing 'newness'. Experimental results showed that a summarization system that used a combination of these feature sets produced the most accurate summaries, and that a combination of acoustic/prosodic and structural features were enough to build a 'good' summarizer when a speech transcription is not available. They found that duration, minimum energy, and maximum energy were particularly discriminatory, while pitch features were among the least useful of the acoustic features.

3.2. Sentence compaction-based method

Hori et al. (2001) proposed a sentence compaction-based method, in which a set of words maximizing a summarization score is extracted from an automatically transcribed sentence, according to a target compression ratio. The extracted set of words is then connected to build a summary. The summarization score consists of:

- Word significance measure: amount of information conveyed by each word,
- Confidence measure: *a posteriori* probability of each word indicating the reliability of speech recognition result,
- Linguistic likelihood: n-gram probability of the word sequence, and
- Word concatenation probability: determined by the dependency structure in the original speech as obtained by a Stochastic Dependency Context Free Grammar (SD-CFG).

The proposed method was further extended to summarize multiple utterances (sentences), which results in a process of combining sentence extraction and compaction.

Hori et al. (2003) have developed an integrated speech summarization approach, based on finite state transducers, in which the recognition and summarization components are combined into a single Weighted Finite State Transducer (WFST). The summarization component consists of paraphrasing and sentence compaction processes. This approach enables the decoder to employ all the knowledge sources in a one-pass search strategy, and therefore reduces the search errors. Another advantage is that the target summary can be derived almost in real time, since the speech can be directly translated into the target sentences frame by frame using a Viterbi search in the integrated network. Experimental results for presentation speech recognition and summarization task showed improvements in both recognition and summarization accuracy over a conventional two-step method.

Kolluru et al. (2005) proposed a multi-stage compaction approach to broadcast news summarization. It employs a network of multi-layer perceptrons (MLP) to remove incorrectly transcribed words based on confidence scores, and to select significant chunks (phrases) at multiple stages based on TF-IDF scores and named entity frequency. The experimental results show that this approach can produce summaries with good information content in comparison to a simple sentence extraction method.

3.3. Combination of sentence extraction and sentence compaction

Kikuchi et al. (2003) proposed a two-stage summarization method consisting of important sentence extraction and word-based sentence compaction, as shown in Fig.1. In this method, after removing all the fillers based on speech recognition results, a set of relatively important sentences is extracted, and sentence compaction is applied to the set of extracted sentences. The sentence extraction and compaction ratios are controlled according to a summarization ratio initially determined by the user. Sentence and word units are extracted from the speech recognition results and concatenated to produce summaries using the method described in Subsection 3.2 which was originally proposed for sentence compaction by Hori, et al. (2001). Thus, the sentence and word units are extracted so that they maximize the weighted sum of the linguistic likelihood, amount of information, confidence measure, and grammatical likelihood of concatenated units. The proposed method has been applied to the summarization of broadcast news utterances, as well as to unrestricted-domain spontaneous presentations, and has been evaluated by objective and subjective measures. It has been confirmed that the proposed method is effective for both English and Japanese speech summarization. It was found that sentence extraction plays a more important role than sentence compaction in improving summarization performance, especially when the summarization ratio is relatively low such as 10%.

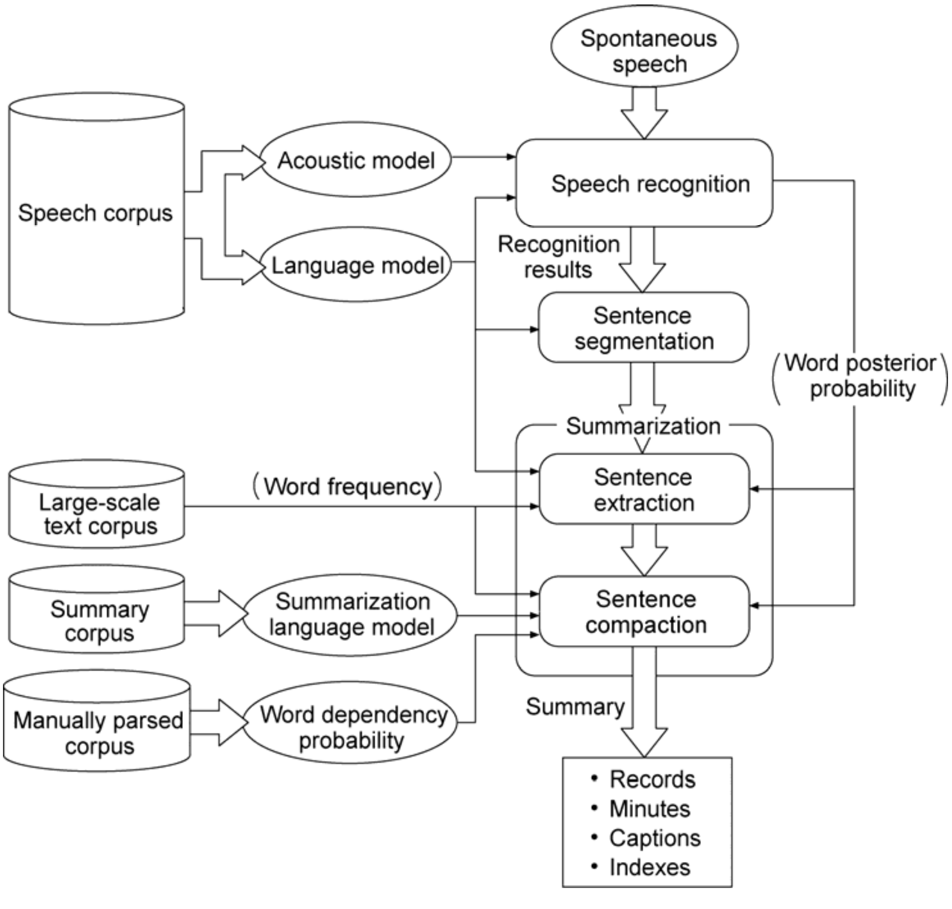

Fig. 1 – The two-stage automatic speech summarization process consisting of sentence extraction and compaction.

3.4. Sentence segmentation

Speech recognition results have no punctuation or proper segmentation, and the readability and usability of such data can be significantly improved by segmenting text into logical units such as sentences. Furthermore, automatic speech summarization can then be used to remove redundancies and erroneous parts, and to extract the important parts of data. It has also been shown that the segmentation has a significant effect on the further processing of the speech, such as information extraction and topic detection (e.g. Shriberg et al. (2000)). Research has shown that methods developed for segmenting written text are insufficient when processing speech, due to the poor grammatical structure, disfluencies, incorrectly recognized words and other characteristics of speech. Even the definition of a sentence in speech is unclear.

Christensen et al. (2005) investigated using a maximum entropy (ME) approach to build statistical models for both utterance (sentence) and topic segmentation in the framework of summarizing ABC news broadcasts from the TDT-2 broadcast news corpus. The experimental work addressed the effect on performance of the topic boundary detector of three factors: the types of feature being used, the quality of the ASR transcripts, and the quality of the utterance boundary detector. Cue word, prosodic and N-gram features were employed, after a feature selection algorithm was used to reduce the number of features to a manageable size. A positive effect was found from combining information extracted directly from the audio stream (i.e. pause duration) with content information obtained from the ASR transcripts. Results of overall experiments showed that the topic segmentation was not affected severely by transcripts errors, whereas errors in the utterance segmentation had more devastating effects.

Mrozinski et al. (2006) investigated an automatic sentence segmentation method based on combining word- and class-based statistical language models to predict sentence and non-sentence boundaries from the viewpoint of its effect on summarization accuracy. The segmentation is done by modeling the probability of a sentence boundary given a certain word history with language models trained on transcriptions and texts from several sources. The resulting segmented data was used as the input to the summarization system proposed by Kikuchi et al. (2003). Experiments were conducted with broadcast news and spontaneous presentation speech. The results showed that proper sentence segmentation is essential to achieving good performance with an automatic summarization system for both broadcast news and presentation speech.

4. Evaluation schemes

4.1. Extrinsic and intrinsic evaluations

Summaries are inherently hard to evaluate because the quality of a summary depends both on the use for which it is intended and on a number of other, subjective, human factors, such as how readable an individual finds a summary or what information an individual thinks should be included in a summary. Although extrinsic evaluation, in which summarization results are assessed in a task-based setting and their usefulness is determined as part of an information browsing and access interface, is ideal, it is also time-consuming and expensive. Therefore, intrinsic evaluation, in which summarization results are assessed in a task-independent setting, is normally employed. Since it is impossible to conduct human evaluation of automatic summarization results every time methods and parameters are changed, it is essential to develop objective evaluation metrics. It is still ideal to use manual summaries as targets of the automatic summary. However, not only do manual summaries vary according to human subjects, but sentence boundaries produced by automatic processes are also variable. Therefore, it is important to develop methods for coping effectively with this problem.

4.2. SumACCY

Hori et al. (2001; 2004) proposed merging all the human summaries into a single word network which is considered to approximately cover all possible correct summaries. Word accuracy of the automatic summary is then measured as a summarization accuracy, SumACCY, by comparing the word sequence with the closest one extracted from the word network.

This metric works reasonably well for relatively easy summarization tasks, but runs into problems when the variation between manual summaries is large, since the network may accept inappropriate summaries which consist of words extracted from different subjects. Therefore, Hirohata et al. (2006) have proposed using word accuracy obtained by using the manual summaries individually, SumACCY-E. In this metric, the largest score of SumACCY-E among human summaries, SumACCY-E/max, or the average score of SumACCY-E, SumACCY-E/ave, is used.

Hori et al. (2003) have shown that an automatic metric WSumACCY which rewarded consensus matches performed better and was more stable than two other metrics (SumACCY and BLEU) that did not take advantage of the consensus matches.

4.3. Rouge

ROUGE-N, which was originally proposed to evaluate written text summarization, is an N-gram recall between an automatic summary and a set of reference (manual) summaries (2004). ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation, and the number of co-occurrences of 1-grams, 2-grams, or 3-grams in the reference summary and the automatic summary is usually used. ROUGE-N is computed as follows:

$$\frac{\sum_{S \in \{Reference\ \ Summaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{Reference\ \ Summaries\}} \sum_{gram_n \in S} Count(gram_n)} \tag{4}$$

where n stands for the length of the n-gram, $gram_n$, and $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in an automatic summary and a set of reference summaries. Note that the number of n-grams in the denominator of the ROUGE-N formula increases as we add more references. Also note that the numerator sums over all reference summaries. This effectively gives more weight to matching n-grams occurring in multiple references. Therefore an automatic summary that contains n-grams shared by more references is favored by the ROUGE-N measure. ROUGE-L (Longest common subsequence) and ROUGE-W (Weighted longest common subsequence) are measures of common subsequences shared between two summaries, with ROUGE-W favoring contiguous common subsequences.

4.4. Experimental results

Lin (2004) showed that ROUGE-1, ROUGE-2, and ROUGE-W correlate well with human judgments for text summarization in the large-scale evaluations at the Document Understanding Conference (DUC) supported by NIST. It was shown that ROUGE-1 correlates particularly well with human judgments of informativeness.

Hirohata et al. (2006) investigated and evaluated various objective evaluation metrics in the framework of sentence extraction-based speech summarization for Japanese presentations under the condition of 10% summarization ratio. In the subjective evaluation, the summaries were evaluated in terms of ease of understanding and appropriateness as summaries on five levels: 1-very bad; 2-bad; 3-normal; 4-good; and 5-very good. The subjective evaluation results were converted into factor scores using factor analysis in order to normalize subjective differences. Correlation analysis between subjective and objective evaluation scores confirmed that SumACCY-E, ROUGE-2, and ROUGE-3 were effective evaluation metrics.

Murray et al. (2005b) compared feature-based approaches using prosodic and lexical features with MMR and LSA-based approaches to automatic speech summarization of multiparty meetings, using the ICSI Meetings Corpus (Janin et al., 2003). All of the automatic summaries were 10% of the original document length. The quality of the summaries was evaluated by using ROUGE-N. Human summaries were used for evaluation and for training the feature-based approaches. Experimental results show that, of the four summarization approaches, the LSA method performed the best on every meeting in terms of ROUGE-1, ROUGE-2, and ROUGE-L measures. The LSA approach was significantly better than either feature-based approach, but was not a significant improvement over MMR. It was observed that, for every meeting, the WER (word error rate) of the summaries was lower than the WER of the meeting as a whole, similar to the observations in the case of broadcast news summarization (Valenza et al., 1999; Zechner & Waibel, 2000).

Murray et al. (2005a) carried out objective and subjective evaluations of automatic summaries of business meetings, with the central interest being whether or not the two types of evaluations correlate with each other. Three basic approaches including MMR, LSA and feature-based classification were evaluated. The human judges were presented with 12 questions at the end of each summary; 6 of the questions regarded informativeness and 6 involved readability and coherence. The evaluations used a Likert scale based on agreement or disagreement with statements, such as the following informativeness statements:

1. The important points of the meeting are represented in the summary.
2. The summary avoids redundancy.
3. The summary sentences seem to be relevant on average.
4. The relationship between the importance of each topic and the amount of summary space given to that topic seems appropriate.
5. The summary is repetitive.
6. The summary contains unnecessary information.

Experimental results show that, in general, ROUGE did not correlate well with the human evaluations for this data. Although the MMR and LSA approaches were deemed to be significantly better than the feature-based approaches according to ROUGE, these findings were reversed according to the human evaluations. ROUGE has been shown to correlate well with human evaluations in DUC, when used on news corpora, but the summarization of spontaneous speech from meetings is quite different from summarizing news articles. Contradictory results concerning ROUGE, obtained by Hirohata et al. (2006) and Murray et al. (2005a), probably mean that ROUGE can be used reliably only when a large number of test points are available.

Valenza et al. (1999) used the information retrieval indexing and search software produced by the CUED HTK group for the 1998 TREC-7 conference to compare the IR performance of the overall decoded text to that of the summaries for broadcast news programs. The summaries, consisting of a keyword list, N-grams (N consecutive words), and sentences, were made by combining acoustic confidence measures with inverse frequency-based measures. Experimental results showed that in the majority of the cases, key information was retained in the summaries; in fact, in some instances precision increased from the full-text to the summaries.

5. Conclusions

Although various automatic speech summarization techniques have been proposed and tested, their performance is still much worse than that of manual summarization. In order to build really useful speech summarization systems applicable to real applications, we definitely need more efficient and speech-focused techniques, including sentence (utterance) segmentation methods. It remains to be determined through further experiments by researchers using various corpora whether or not the objective evaluation measures that have been proposed correlate well with human judgments. There still exists large room for improvement in the objective measures. It is also necessary to evaluate summaries extrinsically within the context of applications, instead of only using intrinsic evaluation methods.

6. References

Alexandersson, J. & Poller, P. (1998). Towards multilingual protocol generation for spontaneous dialogues. In *Proc. INLG-98*, Niagara-on-the-lake, Canada.

Carbonell, J. & Goldstein, J. (1998). The use of MMR, density-based reranking for reordering documents and producing summaries. In *Proc. ACM SIGIR* (pp. 335--336).

Christensen, H., Gotoh, Y., Kolluru, B. & Renals, S. (2003). Are extractive text summarization techniques portable to broadcast news. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 489--494). St. Thomas.

Christensen, H., Kolluru, B., Gotoh, Y. & Renals, S. (2004). From text summarization to style-specific summarization for broadcast news. In *Proc. ECIR-2004*.

Christensen, H., Kolluru, B., Gotoh, Y. & Renals, S. (2005). Maximum entropy segmentation of broadcast news. In *Proc. ICASSP 2005* (pp. I-1029--1032). Philadelphia, PA.

Furui, S., Kikuchi, T., Shinnaka, Y. & Hori C. (2004). Speech-to-text and speech-to-speech summarization of spontaneous speech. In *IEEE Trans. Speech & Audio Proc.* Vol. 12. No. 4. (pp. 401--408).

Garofolo, J. S., Voorhees, E. M., Auzanne, C. G. P. & Stanford, V. M. (1999). Spoken document retrieval: 1998 evaluation and investigation of new metrics. In *Proc. ESCA Workshop on Accessing Information in Spoken Audio* (pp. 1--7). Cambridge.

Gong, Y. & Liu, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. In *Proc. ACM Special Interest Group on Information Retrieval,* (pp. 19--25). New Orleans.

Hearst, M. A. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*. Vol. 23. No. 1. (pp. 33--64).

Hirohata, M., Shinnaka, Y. & Furui, S. (2003). A study on important sentence extraction methods using SVD for automatic speech summarization. In *Proc. 2003 Autumn Meeting of the Acoustical Society of Japan*, Vol. 1. (pp. 93-94). (in Japanese).

Hirohata, M., Shinnaka, Y., Iwano, K. & Furui, S. (2006). Sentence-extractive automatic speech summarization and evaluation techniques. *Speech Communication*. Vol. 48. No. 9. (pp. 1151—1161).

Hori, C. & Furui, S. (2001). Advances in automatic speech summarization. In *Proc. Eurospeech 2001*. (pp. 1771--1774).

Hori, C., Hori, T. & Furui, S. (2003). Evaluation methods for automatic speech summarization. In *Proc. Eurospeech 2003*. (pp. 2825--2828).

Hori, C., Hirao, T. & Isozaki, H. (2004). Evaluation measures considering sentence concatenation for automatic summarization by sentence or word extraction. *Proc. Association for Computational Linguistics*. (pp. 82--88). Barcelona.

Hori, T., Hori, C. & Minami, Y. (2003). Speech summarization using weighted finite-state transducers. In *Proc. Eurospeech 2003*. (pp. 2817--2820).

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A. & Wooters, C. (2003). The ICSI meeting corpus. In *Proc. ICASSP 2003*.

Kikuchi. T., Furui, S. and Hori, C. (2003). Two-stage automatic speech summarization by sentence extraction and compaction. In *Proc. ISCA-IEEE Workshop on Spontaneous Speech Processing and Recognition*. TAP10. Tokyo.

Kolluru, B., Christensen, H. and Gotoh, Y. (2005). Multi-stage compaction approach to broadcast news summarization. In *Proc. Interspeech 2005*. (pp. 69--72).

Kolluru, B., Christensen, H., Gotoh, Y. & Renals, S. (2003). Exploring the style-technique interaction in extractive summarization of broadcast news. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*. (pp. 495--500). St. Thomas.

Kong, S.-Y. & Lee, L.-S. (2006). Improved spoken document summarization using probabilistic latent semantic analysis (PLSA). In *Proc. ICASSP 2006*. (pp. I-941--944). Toulouse.

Koumpis, K. & Renals, S. (2000). Transcription and summarization of voicemail speech. In *Proc. ICSLP 2000*. (pp. 688--691). Beijing.

Kupiec, J., Pederson, J. & Chen, F. (1995). A trainable document summarizer. In *Proc. ACM SIGIR 1995*. (pp. 68--73).

Lin, C.-Y. (2004). Looking for a few good metrics: ROUGE and its evaluation. In *Proc. Working Notes of NTCIR-4*. Vol. Supl. 2. (pp. 1--8).

Mani, I. & Maybury, M. T. (Ed.). (1999). Advances in automatic text summarization, *MIT Press*. Cambridge. MA.

Mrozinski, J., Whittaker, E., Chatain, P. & Furui, S. (2006). Automatic sentence segmentation of speech for automatic summarization. In *Proc. ICASSP 2006*. (pp. I-981--984). Toulouse.

Murray, G., Renals, S., Carletta, J. & Moore, J. (2005a). Evaluating automatic summaries of meeting recordings. In *Proc. ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization (MTSE)*. (pp. 33--40). Ann Arbor.

Murray, G., Renals, S. & Carletta, J. (2005b). Extractive summarization of meeting recordings. In *Proc. Interspeech 2005*, (pp. 593--596). Lisbon.

Sameer, M., & Hirschberg, J. (2005). Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *Proc. Interspeech 2005*, (pp. 621--624). Lisbon.

Shinozaki, T. & Furui, S. (2004). Spontaneous speech recognition using a massively parallel decoder. In *Proc. ICSLP 2004*. Vol. 3. (pp. 1705--1708). Jeju Island.

Shriberg, E., Stolcke, A., Hakkani-Tur, D & Tur, G. (2000). Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*. Vol. 32. No. 1-2. (pp. 127--154).

Steinberger, J & Jezek, K. (2004). Text summarization and singular value decomposition. T. Yakhno (Ed.) *Lecture Notes in Computer Science 3261. Third International Conference on Advances in Information Systems (ADVIS 2004)*. (pp. 245--254). Springer-Verlag. Berlin and Heidelberg.

Valenza, R., Robinson, T., Hickey, M. & Tucker, R. (1999). Summarization of spoken audio through information extraction. In *Proc. ESCA Workshop on Accessing Information in Spoken Audio*. (pp. 111--116). Cambridge.

Zechner, K. (2002). Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*. 28. 4. (pp. 447--485).

Zechner, K. (2003), Spoken language condensation in the 21st century. In *Proc. Eurospeech 2003*. (pp. 1989--1992). Geneva.

Zechner, K. & Waibel, A. (2000). Minimizing word error rate in textual summaries of spoken language. In *Proc. NAACL 2000*. Seattle.