Summarization of Broadcast News Using Speaker Tracking

Sree Harsha Yella
LTRC
IIIT, Hyderabad
sreeharshay@research.iiit.net

Kishore Prahallad, Vasudeva Varma
LTRC
IIIT, Hyderabad
{kishore, vv}@iiit.ac.in

Abstract

In this paper we demonstrate an automatic summarization system for broadcast news shows. The proposed technique does not require ASR transcripts or human reference summaries. The system exploits the role of anchor speaker in a news show by tracking his/her speech to construct indicative extractive summaries. Speaker tracking is done by autoassociative neural network model. Summaries are generated for desired compression ratio . The output summary is presented in the form of speech. The experiments were carried out on BBC news podcasts available online. The evaluation results show that summarization of structured speech documents like broadcast news shows can be performed with good accuracies comparable to text summarization.

1 Introduction

In recent years the amount of multimedia data available has increased rapidly, especially due to increase of broadcasting channels and availability of cheap and efficient mass storage means. In this era of information explosion, there is a great need for systems that can distill this huge amount of data automatically with less complexity and time. News broadcasts are of specific interest since they are filled with information that is relevant for many classes of people. Hence, summarization of broadcast news is of great importance.

The problem of broadcast news summarization has been attacked from different directions in previous research work. Broadcast news corpora were used for experiments in many speech summarization systems. Speech summaries are mainly extractive in nature. Extractive summaries are those formed by concatenation of important parts in the original signal. The early approaches applied the text summarization approaches such as maximum marginal relevance (MMR),latent semantic analysis (LSA), on Automatic Speech Recognition (ASR) system's output of spontaneous speech (Zechner, 2001) (Murray et al., 2005). In (Furui et al., 2004) importance of sentences obtained from ASR output was determined using significance score, linguistic score, confidence score of recognition. (Christensen et al., 2004) explores the relation between style of a broadcast news story and different summarization techniques. In (Kolluru et al., 2005) multi layer perceptrons were employed to eliminate ASR errors and utterances were picked based on term frequency (TFIDF) scores and named entity frequency. (Maskey and Hirschberg, 2003) attempts to summarize broadcast news using structural features. (Inoue et al., 2004) scores the sentences based on prosodic features and lexical features. (Maskey and Hirschberg, 2005) combines lexical and acoustic features to train a supervised system to classify an utterance as belonging to summary or not. (Maskey and J.Hirschberg, 2006) attempts to summarize speech without lexical features, using only acoustic features in a HMM frame work. Several approaches based on sentence generative probability using word topical mixture model (WTMM) (Chiu and Chen, 2007) were attempted. (Maskey and Hirschberg, 2008) attempts to determine the choice of unit of extraction for speech summarization and it was proposed that the intonational phrases are better choice of extraction than sentences and pause based boundaries.

Proceedings of ICON-2009: 7th International Conference on Natural Language Processing, Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2009

All the above approaches depend either on output of ASR system or manual transcripts in some way or the other, or they need human reference summaries which are not easy to obtain for all types of corpora. Moreover construction of human reference summaries is time consuming and tedious job. In this paper we propose an approach to summarize broadcast news using speaker tracking technology. The idea lies in exploiting the characteristics of broadcast news, where a specific structure is followed to deliver the news content. We make use of the fact that in broadcast news, there is a pattern of anchor-speaker and on-field reporter taking turns to cover each story. The anchor-speaker provides introduction to each news-topic and often summarizes briefly before the on-field reporter provides details of the story. Our approach aims at performing the task of summarization by tracking anchor-speaker's speech. Once the segments of anchor-speaker's speech are obtained, a summary is obtained for desired compression ratio by using positional features of these segments. Please note that our approach does not use ASR or manual transcripts and also does not require human reference summaries for training. Moreover, the summaries are provided in audio format as it prevents errors due to ASR, preserves the emotions and characteristics of natural speech and also users can listen to the summaries without having to concentrate on the task of reading.

The proposed approach performs speaker tracking using an Auto Associative Neural Network (AANN) model in broadcast news to construct summaries. The method achieves good recall and precision scores which indicate that speech summarization on structured data like news broadcast can perform as good as text summarization. Section 2 describes the scope of the problem and the motivation for our approach. Section 3 describes the features and the approach in detail, Section 4 describes our method of evaluation and shows the results. Finally Section 5 presents discussion and Section 6 our conclusions and future work.

2 Scope of the Problem and Motivation for our Approach

2.1 Scope of the Problem

Broadcast news show follows a certain structure depending on the genre of the show. There are regular news bulletins featuring stories across all domains. There are news shows dedicated to specific domain like financial reports, sports news, weather forecast, entertainment news. There are daily news shows, weekly and monthly reviews, etc. Most of the broadcast news have an anchor speaker who starts the show by reading the headlines and then presents each story where reporters and others speakers may be involved. Our approach assumes this structure of a broadcast news show. Our aim is to find the segments in the news show, that when concatenated together form a meaningful summary. The summaries generated by our technique will be indicative extractive summaries. Indicative summaries are those that announce the contents of a document without describing them in much detail. The summaries are provided in audio format.

2.2 Motivation for our Approach

The field of summarization dates as early as library science, where human abstractors write abstracts for books and articles for their easy access to readers. The way professional abstractors perform summarization may help us a great deal in building automatic summarization systems (Mani, 2001). Professional abstractors do not focus on understanding a document for summarizing it, instead they make use of the properties of structure of the document such as title, position of a sentence in the paragraph (beginning and ending) and also cue phrases to find important parts in the document. Once they have found the parts of the document that describe the content of the document, they construct simple sentences on the contents of these segments to present it as an abstract. Hence, to summarize any document it is important to first find informative sections in the document. This may not be trivial for all types of documents.

In the field of automatic text summarization it has been shown that structural and positional features are very important (Kupiec et al., 1995) (Lin and Hovy, 1997). It has been shown that initial sentences of a news story provide relevant information regarding the news story and can be included in the summary. In the case of a broadcast news show, if we consider the whole show as a document, it is organized in the form of news stories one after the other where each news story is started by the anchor speaker who provides background information for the story before specifying the actual news. Some times he performs the task of summarizing the views of other speakers also. The anchor speaker's speech contains at least the information about what news stories are presented in the show if not details about these stories. Also, anchor speaker's speech is more planned and precise. Hence, while doing extractive summarization, its more meaningful to extract the anchor speaker turns to obtain good indicative summaries. Our approach aims at performing this task of tracking anchor speaker's speech at the beginning of each news story in a news show, without using any lexical information. Meaningful summaries are constructed by concatenating these extracts according to a given compression ratio.

3 Proposed Approach

The proposed approach performs speaker tracking in a news show to get anchor speaker information, and the turns that are in the beginning of each news story are hypothesized to contain relevant information about the news story. These extracts of anchor speaker snippets are concatenated according to a given compression ratio to form a meaningful summary. The block diagram of the proposed approach is presented in Figure 1.

3.1 Dataset Used

All the news shows used in the experiments belong to globalnews podcast of BBC podcasts[1] available online. The show provides a daily update of global news and features different anchor speakers. Each show was sampled at $16kHz$. We evaluate our method on 10 shows each around 30 min of duration. Each show contains a single anchor speaker. While there are a total of 5 anchor speakers in 10 shows, 3 male and 2 female.

3.2 Extraction of Features from Speech Signal

To perform speaker tracking, speaker-specific features are extracted from the speech signal. Typically these features represent the short-time spectral information such as mel-frequency cepstral coefficients (MFCCs) which describe the vocal tract properties of an individual broadly (Rabiner and Juang, 1993). In our study, 13 MFCC features were extracted from the anchor speaker's speech for each frame, with a frame length of 10 ms and frame shift of 5 ms. These features are given as input to an Auto-Associative Neural Network (AANN) to capture the variability in an individual (Yegnanarayana et al., 2001).

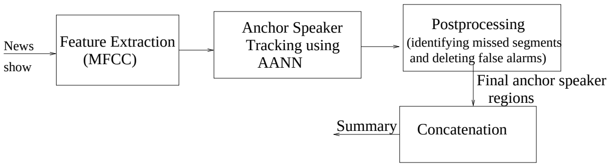
News show
Feature Extraction (MFCC)
Anchor Speaker Tracking using AANN
Postprocessing (identifying missed segments and deleting false alarms)
Final anchor speaker regions
Summary
Concatenation

Figure 1: *Block diagram of the summarization system.*

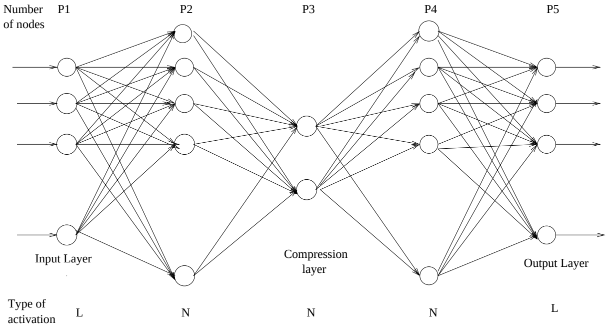

Figure 2: *Five layer Auto Associative Neural Network.*

3.3 Speaker Tracking using AANN model

Artificial Neural Network (ANN) models consist of interconnected processing nodes, where each node represents the model of an artificial neuron, and the interconnection between two nodes has a weight associated with it. ANN models with different topologies perform different pattern recognition tasks (Yegnanarayana, 2004). For example, a feedforward neural network can be designed to perform the task of pattern mapping, whereas a feedback network could be designed for the task of pattern association. A special case of feedforward network is autoassociative neural network (AANN) models which perform identical mapping of input space. It has been shown such networks effectively captures speaker characteristics and could be used for speaker recognition and tracking (Yegnanarayana et al., 2001).

The structure of AANN model is similar to the one followed in (Yegnanarayana et al., 2001). The network structure that was used in our experiments consists of 5 layers: $13\ L\ 39\ N\ 4\ N\ 39\ N\ 13\ L$, where the numbers indicate the number of nodes in the corresponding layer. L represents linear output function and N represents tangential output function. The AANN network layout is shown in Figure 2.

[1] http://www.bbc.co.uk/podcasts/series/globalnews/

The above structure was estimated over a few trials with different number of units in each layer. 13 MFCC features extracted for each frame are given as input to the network with the same feature vector as desired output. The weights in the network are modified by standard backpropagation learning law (Yegnanarayana, 2004). The weights of the network are adjusted for 200 cycles of presentation of data, where each cycle involves presentation of all training data once.

The proposed speaker tracking method follows an iterative technique to identify the segments of speech belonging to anchor speaker and the speaker model is refined in each iteration. An AANN model is trained with initial 30 s of speech of the show which contains anchor speaker's speech mostly. This is a reasonable assumption to make as in most cases, anchor speaker starts the show by greeting the audience and reading the headlines. 13 MFCC features of each frame (generated by a frame length 10 ms and shift of 5 ms) of the show are given as input to the model. The mean squared error ($e[n]$) between the actual output and desired output is calculated. When MFCC features are given as input to AANN model, error as a function of time is not uniform in time. So, we used a confidence measure similar to the one proposed in (Yegnanarayana et al., 2001) defined as,

$c[n] = exp(-e[n])$, where

$e[n]$ is the mean squared error for the nth frame.

$c[n]$ is the confidence score for the nth frame.

The confidence score will be high for the regions belonging to the speaker on whom the model is trained. These confidence scores are smoothed by a moving average window of length 2 s. The valleys in the smoothed confidence contour belong to speech of speakers other than anchor speaker. The smoothed confidence contour is shown in Figure 3.

This smoothed confidence contour is divided into non overlapping segments of 5 s each and mean confidence score is calculated for each segment. Length of the segment is chosen as 5 seconds as average length of speaker turn in a news show is around 5 seconds. Mean confidence score of a segment is compared against a threshold to classify it as belonging to anchor speaker or not. The threshold is calculated automatically as mean value of the smoothed confidence contour in the region belonging to initial 30 s (training) speech.

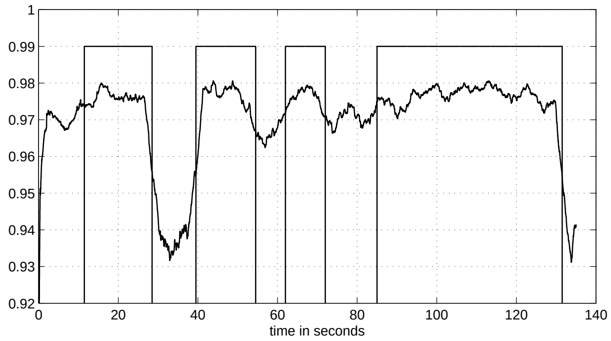

Figure 3: *Smoothed confidence contour with a moving average window of 2 s with anchor speaker regions marked.*

All the segments that have mean confidence score greater than or equal to the threshold are identified as anchor speaker's speech. The MFCC features of these identified segments are used as training data for the next iteration. The above process is repeated untill the model converges. The threshold ensures that only the segments that have a high likelihood of belonging to the modeled speaker are identified. The speaker tracking efficiencies for each iteration are calculated in terms of precision and recall. A segment is considered as belonging to the anchor speaker if it contains more than half of his speech. The speaker tracking performance for each iteration is shown in Table 1.

Table 1: *Speaker tracking performance.*

iter-no	Recall	Precision
1	0.220	0.968
2	0.458	0.962
3	0.541	0.967
4	0.595	0.970

The anchor speaker turns are distributed roughly as shown below in a typical broadcast news show.

Anchor(Headlines): ....
Anchor (Story 1): Its not often when an US
president quotes lines..
Reporters and other speakers: ....
Anchor (Story 2): Its several days now
since opposition leader ..
Reporters and other speakers: .....

We are interested in tracking those segments that indicate topic shift or in other words utterances of anchor speaker at the beginning of each news story. The anchor speaker has a significant duration (> 10 s) of speech at the beginning of each news story where he provides background for the story and specifies the story. Anchor speaker turns inside a topic (if it involves some interviews or interaction with other speakers) are short typically less than 10 s.

We can observe from Table 1 that the recall increases for each iteration while the precision is almost constant. The process was stopped after 4 iterations as it was observed that segments identified by 4th iteration cover anchor speaker's speech in all the topics. Even though the recall is not high, it was observed that the segments identified cover the starting portions of all the news stories which contain significant amount of anchor speech. There may be some missing segments in between identified segments in each news story and some false alarms which are taken care in the summary construction step.

3.4 Construction of Summary

The segments authored by the anchor speaker are obtained from the speaker tracking module. It is evident from Table 1 that recall is not high, which indicates that there are a fair number of segments that need to be identified further. High precision values indicate that there are less false alarms but there is a need to eliminate regions of speech that indicate the interactions of anchor speaker with other speakers which are not relevant for summary. The summary construction involves 3 steps.

- Missed segment identification
- False alarm detection
- Concatenation with Compression

3.4.1 Missed segment identification

Segments attributed to anchor speaker are distributed among different news stories in a news show. Given a news story, there could be segments that belong to anchor speaker that are not identified by speaker tracking model. We treat them as missed detections. It was observed that most of the missed detections are between two identified segments. Therefore the segments that lie within a window of $10s$ between two identified segments are assumed to be of anchor speaker's.

3.4.2 False alarm detection

The output of speaker tracking contains some isolated segments attributed to anchor speaker. These segments are mostly false alarms or instances of interaction of anchor speaker with other speakers. These isolated segments are ignored while constructing summary. A segment is treated as isolated if there is no anchor speaker segment within 10 s of it's boundaries towards left and right.

3.4.3 Concatenation with Compression

After identifying the missed detections and deleting the false alarms, we obtain final anchor speaker regions that need to be concatenated to form a summary. Each anchor speaker region is seperated by speech of reporters or other speakers. The beginning of an anchor speaker region is assumed as beginning of news story because anchor speaker is the person who starts a news story and turns of anchor speaker in between two news stories are deleted during false alarm detection.

The compression ratio (cr) is defined as the ratio of desired summary length to the total length of a document. The required summary length (Sl) is obtained from the given compression ratio (cr) as

$Sl = cr * (Tl)$, where

Tl is the total length of the show in seconds.

The number of stories is approximately equal to the number of anchor speaker regions (N).

Duration(D) of each news story in a summary is obtained as

$D = Sl/N$.

Initial D s of speech from each anchor speaker region are taken as candidates for concatenation. This type of selection makes sure that all news stories are covered in the summary. If anchor speaker's speech in a particular news story is less than D s then the boundary is adjusted accordingly to the end point of his speech. The boundaries of these candidate regions are not meaningful, either acoustically or linguistically, and they may be abrupt. To make them smooth the boundaries of these regions are extended to the nearest 250 ms pause in the signal. The final candidates are concatenated to form a meaningful audio summary.

4 Evaluation and Results

The evaluation was carried out on globalnews podcast of BBC news, details of which are presented in Section 3.1. Two types of evaluations are carried out, one based on traditional text summary evaluation system ROUGE and the other, human evaluation for audio summaries.

Recall Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004) which is commonly used for evaluating text summaries, measures overlap units between automatic and manual summaries. Manual summaries are obtained by asking human annotators to mark the important segments that can be extracted for concatenation from the original show. These segments are extracted and concatenated to form a human reference summary. Human annotators were not given any restriction on summary length. Headlines are removed from human reference summaries and automatic summaries to eliminate their effect on final scores. The size of human reference summaries was not altered for evaluating automatic summaries of different compression ratios. The audio summaries are transcribed manually into text for the evaluation purpose. ROUGE-N computes the n-gram overlap between the summaries where N indicates the size of n-grams. We report ROUGE-1, ROUGE-2 and ROUGE-SU4 scores. ROUGE-SU4 indicates the skip bi-gram score within a window length of 4. The ROUGE scores for different compression ratios are presented in Tables 2-6 below.

Table 2: *ROUGE based scores for compression ratio* 5

type	Recall	Precision	F-measure
ROUGE-1	0.48859	0.90347	0.63220
ROUGE-2	0.44632	0.82937	0.57843
ROUGE-SU4	0.43692	0.81358	0.56667

In human evaluation, 5 human subjects were asked to listen to a summary of a given compression rate and answer a questionnaire given to them. All the subjects are in the age group of 20-23 and are graduate students who can understand and speak english. As the aim of our summarizer is to generate indicative summaries, which announce the contents of a document, the questionnaire consisted of simple questions based on facts of a news story. The questions are of type what, when, who, where etc. The subjects were given strict instructions not to use their prior knowledge on the news stories in answering the questions. They answered the questions based on the information present in the summary. The subjects were not restricted from listening to a summary multiple times. All the answers are one word answers like name of place or person etc present in news story. The percentage of the questions answered correctly for each compression ratio is presented in Table 7.

Table 3: *ROUGE based scores for compression ratio* 10

type	Recall	Precision	F-measure
ROUGE-1	0.68696	0.89393	0.77462
ROUGE-2	0.63130	0.82390	0.71269
ROUGE-SU4	0.62455	0.81564	0.70529

Table 4: *ROUGE based scores for compression ratio* 15

type	Recall	Precision	F-measure
ROUGE-1	0.80461	0.87675	0.83514
ROUGE-2	0.74845	0.81595	0.77693
ROUGE-SU4	0.74302	0.81030	0.77141

Table 5: *ROUGE based scores for compression ratio* 20

type	Recall	Precision	F-measure
ROUGE-1	0.85261	0.85928	0.85218
ROUGE-2	0.79442	0.79818	0.79280
ROUGE-SU4	0.79053	0.79458	0.78905

5 Discussions

The ROUGE scores show that there is a good extent of overlap between human reference summaries and automatic summary, showing that humans also believe that anchor speaker's speech can to some extent describe the contents of the show. The precision and recall values are high as the reference summaries are also extracts from the news show and not abstractive. The ROUGE scores for different compression ratios suggest that recall values increase with increasing compression ratio without much decrease in precison. The system achieves good precision scores which indicates that most of the extracted segments belong to the summary. It is necessary for an extractive summary to have a good precision because if the number of extracts is increased the recall scores are high but still the quality of summary is low.

Table 6: *ROUGE based scores for compression ratio* 25

type	Recall	Precision	F-measure
ROUGE-1	0.88570	0.83869	0.85651
ROUGE-2	0.83345	0.78606	0.80432
ROUGE-SU4	0.82746	0.78058	0.79860

Table 7: *Percentage of questions answered correctly for different compression ratios(CR)*

CR	5	10	15	20	25
correct(%)	0.43	0.54	0.61	0.66	0.70

The results support the hypothesis that initial sentences in a news story are informative and are good candidates for inclusion into summary.

The audio summary evaluation done by human subjects also agrees with the above observations. The percentage of the questions answered increased with the compression ratio as expected. If we treat the number of questions answered as an indication of information gained by the subject, it can be seen that more than 50% of the information in the show is gained from just 10% compression ratio summary. Even higher compression ratios show good increase in the information acquired which shows the summarization capability of the system. The fact that 5% compression ratio summary has nearly 44% of information conveys the importance of the initial sentences in a news story.

6 Conclusions and Future work

We have demonstrated an automatic summarization system for broadcast news shows with a single anchor speaker. The novelty of our approach lies in the fact that it does not require any transcripts or reference summaries, and in that the summaries are generated in the form of speech such that the naturalness in the original signal is preserved. The proposed system generates summaries for different compression ratios without compromising much on the information coverage. Good recall and precision scores indicate that it is possible to build extractive speech summarization systems with performance comparable to text summarization systems provided they have some inherent structure that can be identified.

In future we plan to extend this work to broadcast news shows with multiple anchor speakers and also try out different approaches for speaker tracking.

References

H. S. Chiu and B. Chen. 2007. Word topical mixture models for dynamic language model adaption. In *Proc. ICASSP*.

H. Christensen, B. Kolluru, Y. Gotoh, and S. Renals. 2004. From text summarisation to style-specific summarisation for broadcast news. In *ECIR*.

S. Furui, T. Kikuchi, Y. Shinnaka, and C. Hori. 2004. Speech-to-text and speech-to-speech summarization of spontaneous speech. *Speech and Audio Processing, IEEE Transactions on*, 12(4):401–408, July.

A. Inoue, T. Mikami, and Y. Yamashita. 2004. Improvement of speech summarization using prosodic information. In *Proc. Speech Prosody*, Japan.

Balakrishna Kolluru, Heidi Christensen, and Yoshihiko Gotoh. 2005. Multi-stage compaction approach to broadcast news summarisation. In *Proceedings of Eurospeech*, pages 69–72.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 68–73, New York, NY, USA. ACM.

Chin-Yew Lin and Eduard Hovy. 1997. Identifying topics by position. In *Proceedings of the fifth conference on Applied natural language processing*, pages 283–290, Morristown, NJ, USA. Association for Computational Linguistics.

C. Y. Lin. 2004. Rouge: a package for automatic evaluation of summaries. In *ACL Text summarization Workshop*.

Inderjeet Mani. 2001. *Automatic Summarization*. John Benjamins.

S. Maskey and J. Hirschberg. 2003. Automatic speech summarization of broadcast news using structural features. In *EUROSPEECH*.

S. Maskey and J. Hirschberg. 2005. Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *ICSLP*.

S. Maskey and J. Hirschberg. 2008. Intonational phrases for speech summarization. In *Interspeech*.

S. Maskey and J.Hirschberg. 2006. Summarizing speech without text using hidden markov models. In *NAACL-HLT*.

Gabriel Murray, Steve Renals, and Jean Carletta. 2005. Extractive summarization of meeting recordings. In *Proceedings of the 9th European Conference on Speech Communication and Technology*, pages 593–596.

Lawrence Rabiner and Biing-Hwang Juang. 1993. *Fundamentals of speech recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

B. Yegnanarayana, K. Sharat Reddy, and S. P. Kishore. 2001. Source and system features for speaker recognition using aann models. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 409–412.

B. Yegnanarayana. 2004. *Artificial Neural Networks*. Prentice-Hall of India Pvt.Ltd.

K. Zechner. 2001. Automatic generation of concise summaries of spoken dialogues in unrestricted domains. *R and D in IR*, pages 199–207.