SIGNIFICANCE OF ANCHOR SPEAKER SEGMENTS FOR CONSTRUCTING EXTRACTIVE AUDIO SUMMARIES OF BROADCAST NEWS

Sree Harsha Yella[1], Vasudeva Varma[1], Kishore Prahallad[1 2]

[1]International Institute of Information and Technology, Hyderabad, India.
[2]Language Technologies Institute, Carnegie Mellon University, USA.
sreeharshay@research.iiit.ac.in, {vv, kishore}@iiit.ac.in

ABSTRACT

Analysis of human reference summaries of broadcast news showed that humans give preference to anchor speaker segments while constructing a summary. We exploit the role of anchor speaker in a news show by tracking his/her speech to construct indicative/informative extractive audio summaries. Speaker tracking is done by Bayesian information criterion (BIC) technique. The proposed technique does not require Automatic Speech Recognition (ASR) transcripts or human reference summaries for training. The objective evaluation by ROUGE showed that summaries generated by the proposed technique are as good as summaries generated by a baseline text summarization system taking manual transcripts as input and summaries generated by a supervised speech summarization system trained using human summaries. The subjective evaluation of audio summaries by humans showed that they prefer summaries generated by proposed technique to summaries generated by supervised speech summarization system.

Index Terms— Summarization, Broadcast news summarization, Speaker tracking.

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 in less time. Broadcast news (BN) is one of the most common media through which people obtain news. Hence, summarization of BN is of great importance.

BN 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 without any alteration. The early speech summarization 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 [1] [2]. In [3] importance of sentences obtained from ASR output was determined using significance score, linguistic score, confidence score of recognition. [4] explores the relation between style of a BN story and different summarization techniques. In [5] multi layer perceptrons were employed to eliminate ASR errors and utterances were picked based on term frequency (TFIDF) scores and named entity frequency. [6] attempts to summarize BN using structural features. Methods have been proposed to combine lexical features derived from ASR transcripts with acoustic/prosodic features derived from speech signal, in a supervised frame work, where gold standard human summaries are used for training a classifer using these features to classify an utterance as belonging to summary or not. [7] scores the sentences based on prosodic features and lexical features. [8] combines lexical, acoustic/prosodic, structural and discourse features to train a supervised system to classify an utterance as belonging to summary or not. [9] attempts to summarize speech without lexical features, using only acoustic features in a HMM frame work. [10] 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.

All the above mentioned methods depend on the availability of human/ASR transcribed speech, or gold standard human reference summaries for training. However, ASR systems may not be available for all languages, and it involves considerable amount of resources and effort in building an ASR system for a new language. Also, constructing gold standard human reference summaries is a tedious job and they are not easily available for all speech files.

In this paper, we propose an approach to summarize BN using anchor speaker tracking. We analyze human reference summaries of BN and find that anchor speaker segments are important as they are picked in most of human reference summaries. The idea lies in exploiting the characteristics of BN, where a specific structure is followed to deliver the news content. We make use of the fact that in BN, there is a pattern of anchor-speaker and on-field reporter taking turns to cover

each story. We propose a method to perform anchor speaker tracking based on Bayesian information criterion (BIC) technique [11]. Once the segments of anchor-speaker's speech are extracted, a summary is obtained for desired compression ratio by using positional features of these segments. The summaries are provided in audio format as it prevents errors due to automatic speech recognition (ASR) and preserves characteristics of natural speech. Our aim is to find the segments in the news show, that when concatenated together form a meaningful and coherent audio summary that is acceptable and useful for humans. The summaries generated by current techniques will be indicative or informative extractive summaries.

2. DATA SET

2.1. BBC news corpus

All the news shows used in the experiments belong to globalnews podcast of BBC podcasts[1] available on-line. The show provides a daily update of global news and features different anchor speakers. We have used a total of 20 news shows each around 30 min of duration. Each show was sampled at 16 kHz and contains a single anchor speaker and multiple other speakers. There are a total of eight anchor speakers in 20 shows, of which three are male and five are female speakers.

2.2. Human reference summaries

The text transcripts of the speech files along with their corresponding audio are presented to 4 human annotators for constructing a summary. All the annotators are graduate students with a good background of English. The annotators were instructed to generate a summary of five minutes in length. They were instructed to pick meaningful phrases or sentences present in original story without altering them. Their aim was to generate a generic extractive speech summary that is coherent and meaningful. The number of human reference summaries used in this work was fixed following document understanding conference (DUC)[1] framework.

3. ANALYSIS OF HUMAN REFERENCE SUMMARIES

The way in which human abstractors perform summarization may help us a great deal in building automatic summarization systems [12]. 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.

In order to study how humans perform summarization of BN, we have asked four graduate students with good English knowledge to summarize each news show in the data set. These audio summaries are transcribed into text manually for analysis purpose. Given these multiple human reference summaries for a news show, it would be interesting to observe the measure of overlap between them and also type of segments present in the overlap. This would help us to identify the features in the input that humans use and agree on, to pick segments in summary. If such features can be identified, it would help in design of automatic summarization systems.

As anchor speaker performs an important task of delivering news and running the show, we investigate his/her contribution to human reference summaries. Tab. 1 shows the % of anchor speaker sentences (An) in human summaries, % of sentences picked in all human summaries which indicates overlap (Ov) among human summaries, % of anchor speaker sentences in the overlap (An_Ov) and % of initial sentences (first two) in each news story (In) that are picked in human summaries.

Table 1. *Statistics of human summaries averaged over 20 news shows.*

type	An	Ov	An_Ov	In
%	74%	63%	92%	89%

Tab. 1 shows that human annotators give importance to anchor speaker utterances while summarizing and they also have a good agreement on this (92 % of the segments in the overlap belong to anchor speaker segments). The bias of human annotators towards anchor speaker segments may be due to their preciseness and salience which are essential for an audio summary. Also the picking of 89% of initial sentences in a story (In) shows the importance of anchor speaker utterances in the starting of story.

4. ANCHOR SPEAKER TRACKING

4.1. Feature extraction 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 [13]. In our study, 13 MFCC features were extracted for each speech frame, with a frame length of 10 ms and frame shift of 5 ms.

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

[1] http://www-nlpir.nist.gov/projects/duc/guidelines.html

4.2. Anchor speaker tracking using BIC

Speaker tracking using BIC method is performed in two stages. In the first stage, the BN show is divided into homogeneous regions containing speech from a single speaker, by detecting speaker change points. In the second stage, agglomerative clustering of these segments is performed using BIC as distance measure. As, anchor speaker has more speech instances spread across the show, the cluster containing more speaker turns is hypothesized as the cluster belonging to anchor speaker.

4.2.1. Speaker change detection

The speaker change detection is performed by the dissimilarity measurement between two adjacent windows based on the comparison of their parametric models. The comparison is performed using Bayesian Information Criterion (BIC) [11]. Bayesian Information Criterion (BIC) is a maximum likelihood criterion penalized by the model complexity (number of model parameters). If X is a sequence of data and M is a parametric model with m parameters, and likelihood $L(X, M)$ is maximized, the BIC for model M is defined as

$$BIC(M) = logL(X, M) - \lambda \frac{m}{2} logN_x \qquad (1)$$

where N_x is the number of points in the data sequence.

The first term represents the extent of match between model and the data. The second term denotes the model complexity. The value of λ is data dependent (theoretical value of λ is 1). The BIC allows us to select a model that best fits the data with less complexity. For speaker change detection, two hypothesis are tested. Consider two windows X and Y adjacent to each other. The first hypothesis (H_1) is that there is no speaker change between X and Y and the second hypothesis H_2 states that a speaker change occurs between the two windows. In H_1 a single multi-dimensional Gaussian distribution is assumed to model the data in the two windows better. In H_2 two multi-dimensional Gaussian distributions one for each window are assumed to model the data better. Let N_x, N_y be the number of data points in X and Y windows respectively and Z be the combined sequence of X and Y windows (N_z = N_x + N_y).

The ΔBIC value between the two hypothesis H_1 and H_2 is given by

$$\Delta BIC(H_1, H_2) = \frac{N_z}{2} log|\Sigma_z| - \frac{N_x}{2} log|\Sigma_x| - \frac{N_y}{2} log|\Sigma_y| + \frac{\lambda}{2}(p + \frac{p(p+1)}{2}) logN_z$$

where λ is a tuning factor which is data dependent and p denotes dimensionality of feature vector (in present case 13). A positive ΔBIC value indicates that a speaker change occurs between two windows. The windows are slid along time axis to detect speaker changes. A speaker change point is hypothesized at time instant i such that

$$max_i \Delta BIC(i) > 0. \qquad (2)$$

The performance of the above technique on the current data set is reported in terms of false alarm rate (FAR) and missed detection rate (MDR) in Tab. 2.

Table 2. *Performance of ΔBIC on current data set*

error type	FAR	MDR
%	9.8%	11%

The BIC technique works better for long speaker turns as there is sufficient data to compute the dissimilarity measure reliably. The window size used in our experiments for computation of BIC was five seconds as speaker turns in news data are typically long. The graph of ΔBIC values with actual speaker change points marked is shown in Fig. 1.

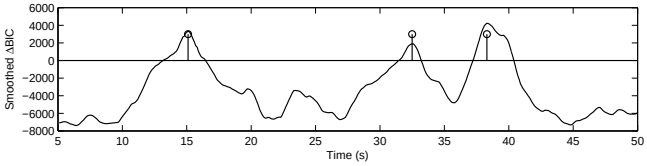

Fig. 1. *ΔBIC based smoothed distance graph with actual speaker change points marked.*

It can be observed from Fig. 1 that the speaker change points coincide with the peaks in smoothed ΔBIC graph. These peaks are considered as speaker change points.

4.2.2. Clustering anchor speaker segments

Homogeneous segments containing speech from a single speaker are obtained by taking segments between two speaker change points. To find the segments of anchor speaker, the segments are clustered by using the ΔBIC values as the distance measure. Initially each individual segment is treated as a cluster and the ΔBIC is calculated for each segment with all other segments. The segments that have ΔBIC values less than or equal to zero are assigned to the cluster of corresponding segment. Ideally, the cluster containing highest number of segments will be anchor speaker's, as anchor speaker will be in all the news stories. But it was observed that there are a few missed anchor speaker segments in this cluster. To reduce these, a global similarity matrix is constructed, by the intuition that segments of same speaker will have similar clusters. If A and B are two clusters then the distance d_{AB} is given by

$$d_{AB} = n(A \Delta B), \qquad (3)$$

where Δ denotes symmetric difference between two sets A and B. d_{AB} gives number of segments that are not present

in both the clusters A and B. Hence, smaller the d_{AB}, the more similar are the two clusters A and B. But considering only d_{AB} as similarity measure might assign a small cluster of other speaker to anchor speaker. To prevent this, an intersection score i_{AB} is introduced which is given by

$$i_{AB} = n(A \cap B). \tag{4}$$

Finally the similarity score s_{AB} is given by

$$s_{AB} = i_{AB} - d_{AB}. \tag{5}$$

All the clusters that have similarity score greater than a threshold (empirically decided as 1) with the cluster containing highest number of segments are treated as clusters of anchor speaker. All these clusters are merged into one cluster and this cluster represents the segments of anchor speaker. The technique can be easily extended to multiple anchor speakers by taking top n clusters which are most dissimilar to each other according to the global similarity matrix. n is equal to the number of anchor speakers. The assumption here is that anchor speakers have more turns in the show than other speakers. The performance of anchor speaker tracking is reported in Tab. 3.

Table 3. *Performance of anchor speaker tracking*

error type	FAR	MDR
%	14%	3%

5. SUMMARY CONSTRUCTION

Each anchor speaker segment can be treated as start of a news story in the show. But there are also instances where anchor speaker interacts with the other speakers within a story. Such segments are typically small and filtered out by removing anchor speaker segments less than 5 seconds in duration.

5.0.3. Concatenation with compression

After removing short segments, we obtain final anchor speaker regions that need to be concatenated to form a summary. 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

$$Sl = cr \times Tl, \tag{6}$$

where Tl is the total length of the show in seconds. The number of stories is approximately equal to the number of final anchor speaker regions (N). Duration (D) of each news story in a summary is obtained as

$$D = Sl/N. \tag{7}$$

Initial D seconds 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 seconds 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.

6. EVALUATION

The evaluation is done on 20 news shows of globalnews podcast of BBC news, details of which are presented in Sec. 2.1. Two types of evaluations are carried out, one using traditional text summary evaluation system ROUGE and the other using human evaluation for audio summaries. ROUGE based evaluation provides an objective measure of quality of the summaries where as human evaluation was done to evaluate the usefulness of the audio summaries for humans. The summaries generated by proposed techniques are compared with summaries generated by a text summarization system , and a supervised speech summarization system similar to the systems proposed in the literature.

6.1. Text summarization system

The manual transcripts of speech files corresponding to each BN show are given as input to the text summarizer to generate a summary. The text summarizer is built using MEAD [14] which uses positional features and tf.idf scores for ranking sentences in a document. The top ranking sentences are picked into the summary until desired summary length is reached. The summaries are generated for a compression ratio of 30 %.

6.2. Supervised speech summarization system

An artificial neural network classifier is trained on gold standard human labelled summaries which contains segments from all four human summaries. The classifier is trained with class labels -1 for class 'non summary' and 1 for class 'summary'. The features on which the classifier is trained consist of minimum, maximum, mean, standard deviation of RMS energy (I), ΔI, F_0, ΔF_0 over each segment and duration of the segment. The F_0 and I contours are normalized using z-score normalization. The corpus is divided randomly into two non overlapping halfs. Classifier was trained on one half and tested on the other. While testing, the classifier outputs a score between -1 and 1 for a given speech segment. This score is used for ranking the speech segments to generate audio summaries for desired length. Summaries are generated

for a compression ration of 30 %.

6.3. ROUGE based evaluation

Recall oriented understudy for gisting evaluation (ROUGE) [15] which is commonly used for evaluating text summaries, measures overlap units between automatic and manual summaries. 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 four. The ROUGE scores of the current system are compared against a baseline text summarization system built using MEAD and supervised speech summarization system trained on gold standard human reference summaries. Audio summaries generated by the system are transcribed manually into text for evaluation purpose. In order to evaluate the summarization capability of the proposed techniques for different summary lengths, summaries are generated for different compression ratios (5, 10, 15, 20, 25 and 30). The size of human reference summaries was not altered for evaluating automatic summaries of different compression ratios. The ROUGE scores of audio summaries for different compression ratios (5, 10, 15, 20, 25 and 30) are presented in Fig. 2.

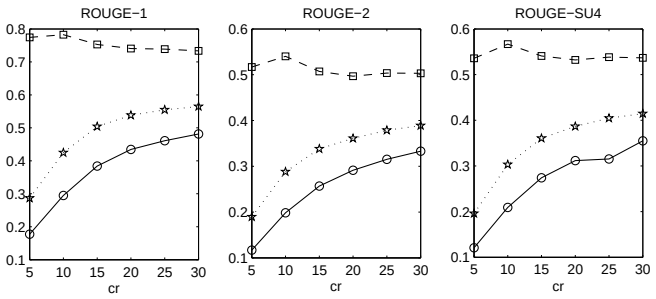

Fig. 2. *Plots showing recall (solid line), precision (dashed line) and F-measure (dotted line) values for various compression ratios (cr) of audio summary generated using BIC based speaker tracking.*

It can be observed from Fig. 2 that recall values of the summaries increase with increase in compression ratio as expected. The precision values are fairly constant for all compression ratios which shows that the new segments that are being added to the summary due to increase in desired summary length are relevant to summary. Precision values are important for an extractive summary, because if the number of extracts is increased, the recall values might increase but the percentage of segments relevant to summary might drop.

The ROUGE scores for summaries generated using proposed speaker tracking techniques, text summarizer built using MEAD and supervised speech summarizer trained on gold standard human summaries for 30 % compression ratio are presented in Tab. 4

Table 4. *F-measure values and 95% confidence intervals for ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-SU4 (R-SU4) metrics for speaker tracking based summaries, summaries generated by supervised system and MEAD summarizer.*

system	R-1	R-2	R-SU4
BIC	0.564[0.54 0.58]	0.388[0.36 0.40]	0.414[0.39 0.43]
supervised	0.553[0.52 0.57]	0.382[0.36 0.40]	0.402[0.38 0.42]
MEAD	0.572[0.55 0.59]	0.394[0.37 0.41]	0.421[0.40 0.44]

It can be observed from Tab. 4 that the proposed speaker tracking techniques produce summaries as good as MEAD based text summarizer and supervised system.

6.4. Human evaluation

6.4.1. Question & Answer based evaluation

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. All the subjects were asked to answer the questionare before listening to summaries to factor out their prior knowledge on the news stories. The subjects were not restricted from listening to a summary multiple times. The percentage of the questions answered correctly after factoring out their prior knowledge for each compression ratio is presented in Tab. 5.

Table 5. *Percentage of questions answered correctly for different compression ratios (cr)*

cr	5	10	15	20	25
BIC	42.4 %	55.6 %	62.0 %	65.5 %	71.0 %
Supervised	36.2 %	41.6 %	47.3 %	53.4 %	60.2 %

The results of Q&A based evaluation in Tab. 5 show that humans are able to understand the audio summaries produced by anchor speaker tracking easily and were able to get more information from them than summaries generated by a supervised system.

6.4.2. Coherance evaluation

In order to evaluate coherence of the audio summaries, subjective evaluation by is performed by 10 subjects. The subjects are asked to evaluate the summaries based on coherence, ease of understanding and appropriateness as a summary. They are provided with text transcript of the news show

before they listen to the summaries, so that they get an idea of the contents of the show. They are asked to rate the summaries at five levels: 1-very bad, 2-bad, 3-normal, 4-good, 5-very good. The mean opinion scores (MOS) of these ratings for summaries of 20 news shows are presented in Tab. 6

Table 6. *MOS of summaries generated by various methods.*

method	BIC	Supervised
MOS	4.05	3.2

From Tab. 6 it can be inferred that human beings prefer summaries generated by the proposed techniques than summaries generated by standard speech summarization systems based on a supervised classifier.

7. CONCLUSIONS AND FUTURE WORK

We have demonstrated an automatic speech-to-speech summarization system for BN shows. The proposed approach does not require any transcripts or reference summaries, and summaries are generated in speech such that the naturalness in the original signal is preserved. The proposed system generates summaries for different compression ratios without degrading the quality of the summaries. 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 for summarizing monologue broadcast news shows and speeches by identifying useful acoustic cues for summarization.

8. REFERENCES

[1] K. Zechner, "Automatic generation of concise summaries of spoken dialogues in unrestricted domains," in *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2001, pp. 199–207, ACM.

[2] Gabriel Murray, Steve Renals, and Jean Carletta, "Extractive summarization of meeting recordings," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, 2005, pp. 593–596.

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

[4] H. Christensen, B. Kolluru, Y. Gotoh, and S. Renals, "From text summarisation to style-specific summarisation for broadcast news," in *ECIR*, Sunderland, UK, 2004.

[5] B. Kolluru, H. Christensen, and Y. Gotoh, "Multi-stage compaction approach to broadcast news summarisation," in *Eurospeech*, Lisbon, Portugal, 2005, pp. 69–72.

[6] S. Maskey and J. Hirschberg, "Automatic speech summarization of broadcast news using structural features," in *EUROSPEECH*, Geneva, Switzerland, 2003.

[7] A. Inoue, T. Mikami, and Y. Yamashita, "Improvement of speech summarization using prosodic information," in *Speech Prosody*, Japan, 2004.

[8] S. Maskey and J. Hirschberg, "Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization," in *ICSLP*, Lisbon, Portugal, 2005, pp. 621–624.

[9] S. Maskey and J. Hirschberg, "Summarizing speech without text using hidden markov models," in *NAACL '06: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers on XX*, Morristown, NJ, USA, 2006, pp. 89–92, Association for Computational Linguistics.

[10] S. Maskey and J. Hirschberg, "Intonational phrases for speech summarization," in *Interspeech*, 2008.

[11] Scott Shaobing Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *DARPA Speech Recognition Workshop*, 1998, pp. 127–132.

[12] Inderjeet Mani, *Automatic Summarization*, John Benjamins, 2001.

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

[14] Dragomir Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Çelebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam Winkel, and Zhu Zhang, "MEAD - a platform for multidocument multilingual text summarization," in *LREC*, Lisbon, Portugal, May 2004.

[15] C. Y. Lin, "Rouge: a package for automatic evaluation of summaries," in *ACL Text summarization Workshop*, 2004.