Multimodal Approach for Speaker Identification in News Programs

Anthony F. Martone, Cuneyt M. Taskiran, and Edward J. Delp

Video and Image Processing Laboratory (*VIPER*)
School of Electrical and Computer Engineering
Purdue University
West Lafayette, Indiana

ABSTRACT

The process of identifying speakers in a news program is difficult using only text information. We propose a system that will first perform text and video processing separately to identify the start of speech of a speaker. These start of speech locations are aligned and used to identify a change of speaker in the program. An analysis is performed to identify the contribution of the text and video information. It will be be shown that the change of speaker locations identified by our alignment algorithm is more accurate then either mode individually.

Keywords: video processing, feature extraction, histogram features, clustering, multimodal analysis, news videos, text processing, closed captioning, alignment, cost analysis

1. INTRODUCTION

In this paper we address the problem of detecting unique people appearing in television news programs and generate a list of times indicating the time intervals where each person appears. For this work, we have focused on video content obtained from the C-SPAN cable television networks.[1] C-SPAN is a private, non-profit organization in the United States that broadcasts meetings, sessions of the U.S. Senate and House of Representatives, interviews, and other news programming. C-SPAN broadcasts without commercials as a public service. We will demonstrate the use of our approach for the C-SPAN interview program *Booknotes*. However, our approach is applicable to the speaker analysis of any other interview or news program.

Booknotes is a one hour interview program with no commercials. The show has only two speakers: a host and a guest. The structure of the program is as follows: In the introduction, the host introduces the guest and talks about the book. In the main part of the program, the host asks questions to the guest. Finally, the host wraps up the program. The guest speaks an average of 45 minutes with little interruption from the host. The host will also speak at the start of miscellaneous shots, for example wide shots.

Our proposed analysis can provide important results that can be used to study the interactions between speakers. The statistics used to measure these interactions are the length of time spoken by the speakers, number of questions asked by the speakers, and the number of times each person speaks. These statistics can be utilized to find a specific or interesting topics within the program. For example, one could find a controversial question and then examine the length of the response.

Another important application of the speaker interaction analysis is the semi-automatic generation of DVD chapters for interview programs. DVD chapters provide entry points into various locations of a program. The list of DVD chapters also function as a table of contents for the program stored on the disc. Usually, each DVD chapter is identified with a frame extracted from the program from the beginning of the chapter and a title that identifies the contents of the chapter. For interview programs such as *Booknotes* a viewer may want to jump to a particular question asked by the host or the corresponding answer by the guest. Having this functionality in the DVDs for such programs would greatly increase their usability.

This work was supported by a grant from the C-SPAN Archives. Address all correspondence to E. J. Delp, ace@ecn.purdue.edu, telephone: +1 765 494 1740.

Figure 1. Speaker identification system.

Generating the chapters for a program is a task-intensive process for content producers. The time codes of each question and the corresponding answer must be determined, the corresponding frames must be extracted from the programs, and titles for each chapter must be generated. Our analysis method will be instrumental in automating most of these processes.

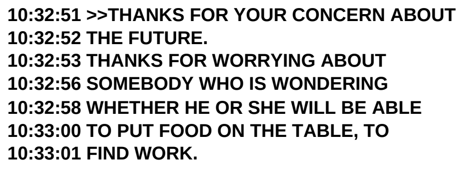
Speaker detection systems have been researched and used in a wide variety of applications. Single model systems have been proposed to identify speakers in broadcast news using only audio information.[2] A multi-model system is proposed by.[3] This system performs a face and name association by first identifying faces of people in a news video. A transcript is then used to find that persons name. An association is then made between the face and name. This system can perform name-to-face or face-to-name retrieval.

In this paper, we describe a system that identifies change of speaker (COS) locations in a news program. A COS location is defined as the time at the start of speech for each interaction of the host and guest. Our system robustly determines the COS locations by processing the video and the corresponding closed-caption text separately and then by combining the results. COS locations in closed-caption text are detected by searching for special tags in text that are placed by captioners whereas in video they are detected by comparing consecutive shots for visual similarity. Neither of these methods is very accurate on their own. It is common for captioners to insert erroneous COS tags or miss some COS locations. Also, the visual content for consecutive shots with different may be similar, leading to missed COS locations if only video processing is used. Each COS location is identified with a timecode. The COS timecodes obtained from closed-caption text processing and video processing are then aligned to obtain the final COS location list for the program. A block diagram of the proposed system is shown in Figure 1.

2. TEXT PROCESSING

Text added to the video signal of a TV program and displayed in the picture are known as *captions*.[4] *Open captions* are captions that are directly added to the frames of the program and become an integral part of the television picture. A common example of open captions are the names and titles of the speakers appearing in a program, which are displayed in the lower part of the TV screen. *Closed captions* provide visual text to describe dialogue and sound effects in television programs.[5] These captions are hidden in the video signal, invisible without a special decoder. The captions are produced by a stenographer, also referred to as a *captioner*, who uses an electronic stenotype keyboard to transcribe the speech content of the program. We refer to the text produced by a captioner as the *caption text*.

The caption text consists of ASCII characters that form several lines of timecodes and words. The timecodes are used to synchronize the audio in the program with the words within the caption text. There exists one timecode and one to ten words in every line of the caption text as shown in Figure 2. As described in Section 4, the timecodes in the caption text will be aligned with the timecodes in the video. It is therefore necessary to find timecodes associated with COS locations. A COS is identified by a double arrow, >>, in the caption text.[3]



```
10:32:51 >>THANKS FOR YOUR CONCERN ABOUT
10:32:52 THE FUTURE.
10:32:53 THANKS FOR WORRYING ABOUT
10:32:56 SOMEBODY WHO IS WONDERING
10:32:58 WHETHER HE OR SHE WILL BE ABLE
10:33:00 TO PUT FOOD ON THE TABLE, TO
10:33:01 FIND WORK.
```

Figure 2. Example of dialog in the caption text.

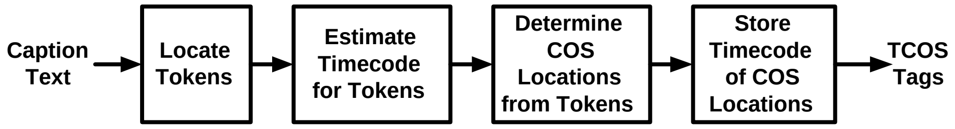

Caption Text
Locate Tokens
Estimate Timecode for Tokens
Determine COS Locations from Tokens
Store Timecode of COS Locations
TCOS Tags

Figure 3. Text processing diagram.

Before the COS tags are located, preprocessing is performed on the caption text.[6] The preprocessor divides the input text into units known as *tokens*, where each token is delimited by white space. Each token is processed to remove all punctuation and non-alphanumeric characters and to also search for special characters. These characters are digits, colons, and right arrows. Digits and colons compose the timecodes and have the format $hh : mm : ss$. A timecode is estimated for each token on the line since several tokens exist for only one timecode. Each token is then searched for the two right arrow marks that indicate a COS. These text COS locations will be referred to as *TCOS tags.* The TCOS tags will be used in the alignment section of our system as shown in Figure 5. The text processing procedure is illustrated in detail by Figure 3.

Since *Booknotes* has only two speakers, it is theoretically possible to identify each speaker by examining the even- and odd-numbered TCOS tags. The even-numbered tags would correspond to the host and the odd-numbered tags would correspond to the guest. A problem with this approach is that errors exist within the caption text. Occasionally the TCOS tag is falsely inserted or is missing. A single erroneous TCOS tag would cause all the following TCOS tags to be false. Therefore it is difficult to identify COS locations based solely on TCOS tags detected from closed-caption text. To make the detection more robust, we perform video analysis in addition to closed caption processing.

3. VIDEO PROCESSING

The closed-caption text provides information of when a person speaks, which is indicated by a timecode. However, even if a single error exists within the closed-caption text, it will be impossible to determine the correct order of speakers from that point onwards. Video analysis will therefore be used to perform shot detection and clustering on the video segment. This will provide information of the time of speech and identify the speaker.

A block diagram of our processing system to derive speaker change timecodes based on video is shown in Figure 4 and is explained in detail by.[7] After extracting luminance histogram and pixel variance features from each frame, shot boundary detection is performed using these features. The middle frame of each shot is used as the representative frame, or *keyframe* for that shot. Another set of features are then extracted from each keyframes, which are used to perform bottom-up clustering to determine shots that are visually similar. We assume that the speakers appearing in the program are visually distinct enough so that such a clustering procedure places shots of different speakers in different clusters. For other details of the shot clustering procedure, refer to.[8]

The clustering algorithm will in general discover more than two clusters in the program. This is due to a number of factors. First, shots the clustering algorithm may erroneously place the shots containing the same speaker in two different clusters. Second, there are shots that are visually impossible to classify as belonging to either speaker, such as wide shots containing more than one speaker or shots not containing any of the speakers. In order to consistently assign cluster labels to such shots, a human operator views the resulting

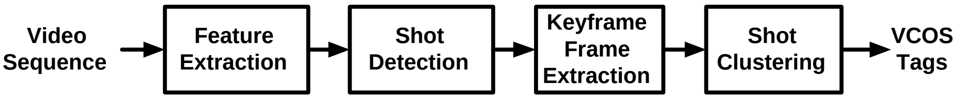


Figure 4. Video processing diagram.

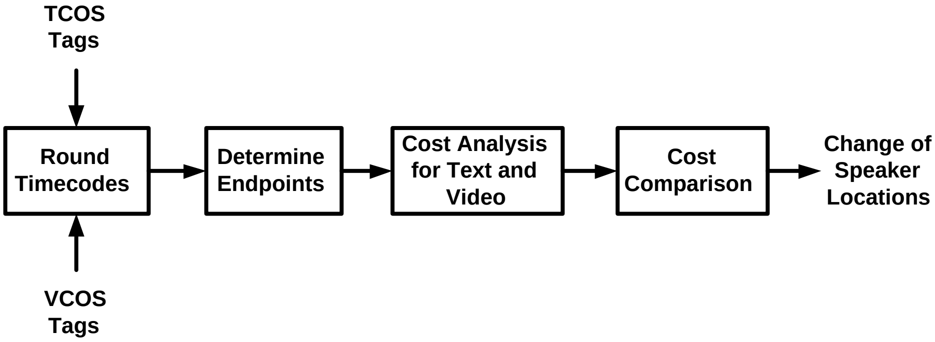


Figure 5. Alignment process of text and video information.

clusters and corrects any errors that are produced by the clustering algorithm. For simplicity, we have label all shots that show neither of the two speakers in *Booknotes* programs as containing the host. The clusters derived above are then used to determine the time locations where each speaker appears. We will refer to this video COS information as *VCOS tags*.

4. ALIGNMENT OF TEXT AND VIDEO INFORMATION

Our alignment algorithm is illustrated in Figure 5. The text information will contribute the start time of a speaker represented by TCOS tags. The video information will contribute the start time of a speaker, represented by the VCOS tags, and identify the speaker. It is therefore necessary to align this information differently.

Since the time resolution of the timecodes from closed-caption text and from video are different, we first round the timecodes obtained from the video analysis. The timecode from the video follows the SMPTE standard and has the following format: $hh : mm : ss : ff$. The timecode associated with the text contains no information for frames as described in Section 2. A video timecode with 15 frames or more will be rounded to the next second.

After rounding, a search is performed to find the timecodes in the text that equal the timecodes in the video. These timecodes signify that both the text and video indicate a COS. This will occur when a TCOS tag and VCOS tag have the same timecode. Unsynchronized TCOS tags and VCOS tags will exist between the synchronized points, or boundary points. These boundary points will be referred to as *endpoints.* We assume that the endpoints indicate a COS location in the program.

The timecodes for the VCOS and TCOS tags may indicate the same COS location, but they could be off by one or two seconds. This is due to the reaction time of the captioner and the shot transition of the video. We will therefore consider an endpoint to be the point when the timecodes of the TCOS and VCOS tags are within two seconds of one another. Two seconds is chosen as a reasonable estimate between the time the shot transitions to the next speaker and the time that the speaker talks. The endpoints will also have cluster numbers associated with them since a VCOS tag is used to create the endpoint. It will therefore be possible to identify the speaker at the endpoints.

The TCOS and VCOS tags that exist between the endpoint are illustrated in Figure 6. These tags are unsynchronized, meaning that they are more then two seconds apart and must be examined. It is possible that some of the unsynchronized TCOS and VCOS tags are in error. It is therefore necessary to search through

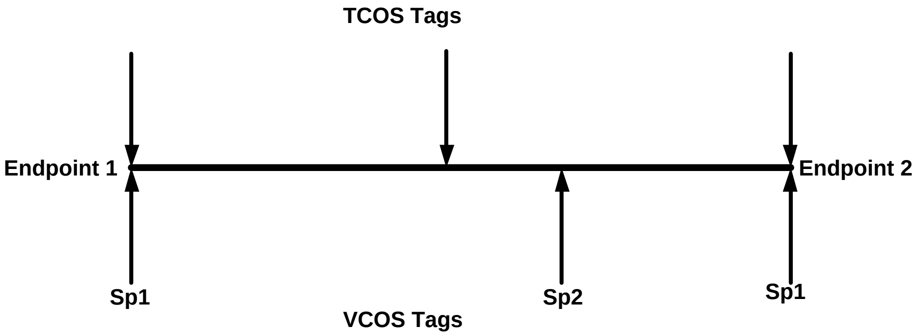


Figure 6. Illustration of endpoints and unsynchronized TCOS and VCOS tags.

the unsynchronized tags between the endpoints to determine the ones that are correct. Costs will be assigned to the unsynchronized tags. These cost will be determined by the likelihood that the tag exists between the endpoints. For example, if both endpoints indicate the same speaker, then a COS transition is highly probably. This would mean that an odd number of tags should exist. It is necessary to examine all combination of unsynchronized tags between each pair of endpoints. Since the TCOS tags do not identify a speaker, it will be necessary to examine the TCOS and VCOS tags separately.

4.1. Text Tag Cost Assignment

A cost is attributed to each TCOS tag in-between the endpoints. The cost will be based on two heuristics:

1. Since two speakers exists, all even TCOS tags should indicate the same speaker and all odd TCOS tags should indicate the other speaker.
2. The total number of true COS locations between the endpoints should be close in value to the total number of TCOS tags.

The first heuristic states that if two TCOS tags are next to one another than they should be different. For example, assume that both endpoints indicate Speaker 1 and one TCOS tag exist between these endpoints as shown in Figure 6. Two scenarios will be examined: the TCOS tag is falsely inserted and therefore ignored or the TCOS tag is correct and is therefore present. Since the endpoints are the same and a COS should occur, a high cost will be assigned to the scenario of the falsely inserted tag. This cost will be referred to as an *alternation cost.* The cost will be dependent upon the number of tags between a pair of endpoints. It also depends on if the speakers indicated by the endpoints are the same as those between the endpoints.

The alternation cost can be estimated for any number of tags between the endpoints. A different scenario will be considered for all combinations of the TCOS tags between the endpoints. Each scenario will be referred to as a *combination set.* The number of combination sets will equal the number of TCOS tags plus 1. For example, if two tags exist between the endpoints then three combination sets will exist. The first combination set will consider zero tags between the endpoints. The second combination set will consider only one tag between the endpoints. The third combination set will consider both tags. An alternation cost will be assigned for each combination set. The alternation cost assignment for the i^{th} combination set is shown by Equation 1. $EP1$ and $EP2$ represent the speaker label at endpoint 1 and endpoint 2 respectively. T_i is the total number of TCOS tags between the endpoints of combination set i. The cost assignment is either 0 or 2.

$$\text{CAlternate}_i = \begin{cases} 2 & \textit{if } EP1 = EP2 \textit{ and } T_i \text{ Even,} \\ 0 & \textit{if } EP1 = EP2 \textit{ and } T_i \text{ Odd,} \\ 2 & \textit{if } EP1 \neq EP2 \textit{ and } T_i \text{ Even,} \\ 0 & \textit{if } EP1 \neq EP2 \textit{ and } T_i \text{ Odd.} \end{cases} \tag{1}$$

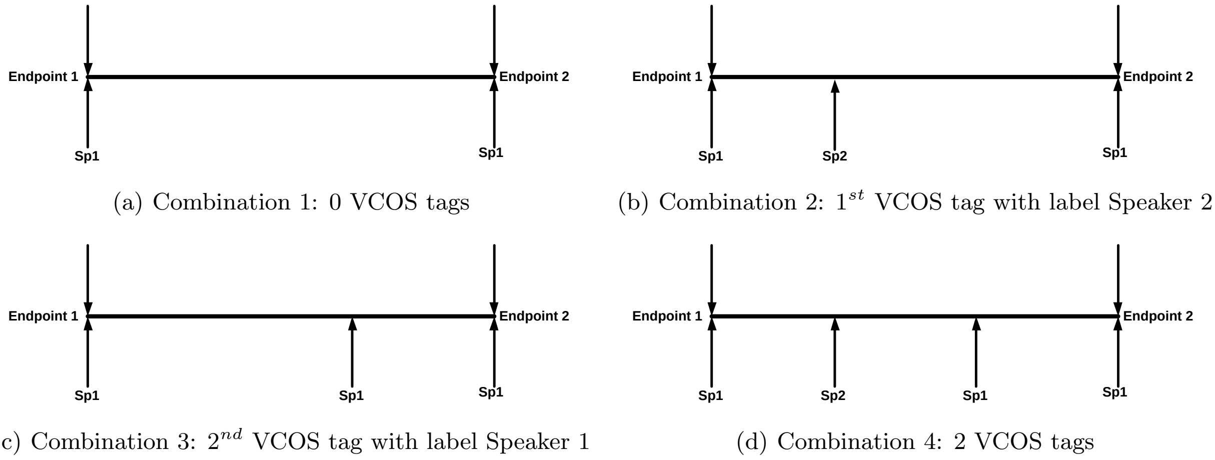
The second heuristic states that the largest number of TCOS tags should be utilized between the endpoints. This is done because the caption accuracy is high. It is therefore unlikely that several consecutive tags are in error. A cost will therefore be assigned for any scenario when a TCOS tag is not considered. For example, in Figure 6, a cost of 1 will be assigned to the scenario when the TCOS tag between the endpoints is not considered. If N tags exists between the endpoints, then a cost of 1 will be added for every tag not considered. This cost will be referred to as a *usage cost* and is shown in Equation 2. T_i is the total number of TCOS tags between the endpoints for the i_{th} combination set. The usage cost and alternation cost can be combined for the same i_{th} combination set as shown in Equation 3.

$$\text{CUsage}_i = N - T_i \tag{2}$$

$$\text{CTotal}_i = \text{CUsage}_i + \text{CAlternate}_i \tag{3}$$

4.2. Video Tag Cost Assignment

The costs for the video will be assigned in the same way as the text. Refer to Section 4.1 for detail. A alternation cost and usage cost will be used. The only difference between the TCOS and VCOS tags is that a cluster number exists for all VCOS tags. The VCOS tags contribute more information than the TCOS tags. Examination of the combination sets will therefore be different from those described in Section 4.1. Consider the case when two VCOS tags are present between the endpoints as shown in Figure 7. Four combination sets exist and a cost will be assigned to each set. The first combination set considers the case when zero VCOS tags are present. Since the endpoints are the same, an alternation cost of 2 will be assigned to this set. Since zero of the two VCOS tags are present, the usage cost will be 2. The total cost for combination set 1 will be 4, as indicated by Equation 3. The second combination set considers the case when the first VCOS tag is present. This tag indicates Speaker 2 and will therefore have an alternation cost of 0. Only one of the two tags are examined, which means that the usage cost is $2 - 1 = 1$. Costs for the remaining two cases are evaluated in the same manner. Combination sets 3 and 4 will incur a total cost of 3 and 2 respectively. Combination set 2 had the lowest cost of the four sets. Combination set 2 should indicate the best alignment of the VCOS tags.



(a) Combination 1: 0 VCOS tags

(b) Combination 2: 1^{st} VCOS tag with label Speaker 2

(c) Combination 3: 2^{nd} VCOS tag with label Speaker 1

(d) Combination 4: 2 VCOS tags

Figure 7. Combinations of VCOS tags between endpoints when N=2.

In general, if N VCOS tags are between the endpoints, then 2^N combination sets must be examined. Equations 1, 2, and 3, are used to define the video cost. The difference is that N refers to the number of VCOS tags.

4.3. Text and Video Cost Comparison

The text combination sets are compared to the video combination sets. One more cost assignment must be added before a comparison is made. This is done if the number of TCOS tags are different from the number of VCOS tags between the endpoints because the costs of the combination sets will not be equal. For example, assume that three TCOS tags and one VCOS tag are between a pair of endpoints. Also assume that a combination set for both the text and video produce a cost of 0. Although they both have the same cost, the text is more probable since it has more tags present. This cost is the same as the usage cost discussed in Section 4.1, which states that the largest number of tags should be utilized. An additional usage cost will therefore be assigned to the text or video combination sets with the smaller number of tags. The amount of the cost will be the difference between the number of VCOS and TCOS tags.

To compare the cost of the text and video combination sets, the union of the text and video sets must be determined. Let the costs for the text combination sets form the set A. Let there be N_t TCOS tags between the endpoints. There will then be $N_t + 1$ elements in set A. Let the costs for the video combination sets form the set B. Let there be N_v VCOS tags between the endpoints. There will then be 2^{N_v} elements in set B. The elements in both sets are the cost values that were discussed in Sections 4.1 and 4.2. The union of the sets is determined and will contain $N_t + 2^{N_v} + 1$ elements. Each element is noted by ζ_i. The probability of each element is determined as shown in Equation 4. The combination set, for either text or video, that corresponds to the most probable element will be used to identify the alignment of the unsynchronized tags between the endpoints. If two probabilities are equal and one is from the text and the other is from the video, then the combination set from the text is favored over the video. This is done since the TCOS tags identify the start of speech, which is the objective of our proposed algorithm.

$$P[i] = \frac{\zeta_i}{\sum_{i=0}^{N_t+2^{N_v}} \zeta_i}, \qquad \text{where } \zeta_i \in \mathrm{A} \cup \mathrm{B} \tag{4}$$

5. RESULTS

For our proposed system we examined two hours of the *Booknotes* program that contained 378 COS locations. During the airing of *Booknotes* the program was digitally encoded to MPEG-1 format. This sequence was then processed by our shot boundary detection and clustering algorithm. The caption text was generated by a closed captioning decoder. The timecodes of the video and caption text must start at the same time. A human operator must trim the video so that the first spoken word aligns with the first word in the text. Further interaction is necessary after shot boundary detection and clustering. The clusters that correspond to the guest must be identified by a human operator. Our system can then distinguish between clusters of the guest and clusters of the host.

Ground truth information is generated and used to identify the true COS locations. This information is used to determine the accuracy of our alignment algorithm. Accuracy is defined in terms of tag error rate (TER), which is shown by Equation 5. The number of speaker tags in error includes *missed*, *false alarm*, and *misclassification* errors. The accuracy of the algorithm is then defined by Equation 6.

$$\text{TER} = \frac{\text{number of COS tags in error}}{\text{number of total tags in ground truth}} \times 100 \tag{5}$$

$$\text{accuracy} = 100 - \text{TER} \tag{6}$$

Before the system was tested, the individual accuracies of the text and video where determined. The captioner identified a COS location with 90% accuracy. The shot boundary detection and clustering accuracy is 80% for a *Booknotes* program. Since the TCOS tags do not identify the speaker, as discussed in Section 2, one TCOS tag error will cause every following tag to be incorrect. The video processing identifies shot boundaries and clusters, but the overall goal of our proposed system is to identify the start of speech of each speaker. Shots will not always indicate a change in speech. The 90% text accuracy and 80% shot boundary detection

	Tags in ground truth	Tags identified	Errors	Accuracy
Proposed System	378	323	102	72.75%
Individual Video	378	210	175	53.70%
Individual Text	378	110	256	32.28%

Table 1. Comparison of speaker identification statistics for text, video, and proposed system.

	Text and Video	Text	Video
Tags	163	104	56
Percentage	50.46%	32.20%	17.34%

Table 2. Percent contribution of COS determined by algorithm. Distribution based on 378 total COS locations.

and clustering accuracy will therefore be a poor indicator for the performance of the system. Table 1 lists the statistics for predicting COS locations by use of our proposed system and by the use of the TCOS tags and VCOS tags individually. As is shown in Table 1, our proposed system is more accurate in predicting COS location than the text or video individually.

The COS locations are identified by the TCOS tags and VCOS tags. It is possible that both tags align and simultaneously predict a COS location. The TCOS and VCOS tags simultaneously occur at an endpoint location. As explained in Section 4, an endpoint signifies alignment between the text and video information. Ideally, all TCOS and VCOS tags should align so that more endpoints exist thereby creating more accurate data. For unaligned data the text and video must individually identify a COS location. For some instances an individual TCOS or VCOS determines a unique location. For example, if a COS occurs and no shot transition exists, the text will still identify the proper location. It is therefore important to note the contribution made by the text and video. The percentage of contribution is shown in Table 2.

Our system produced three types of errors: missed, false alarm, and misclassification. The number for each error is shown in Table 3. As is shown, our system missed several COS locations. This is the largest error that results from our proposed algorithm. This error is produced by both the caption transcript and video. These errors are caused primarily by quick interactions between both speakers. For example, five interactions could occur within a three second time period. At times, the captioner will not be quick enough to identify all the true locations. For video, the shot will stay fixed on only one speaker during these interactions and the COS locations are completely missed.

False alarms are caused mostly by the video, but sometimes by text. As with missing errors, a quick interaction could cause a captioner to accidently insert more tags than needed. Video false alarms are caused by wide shots. Our system assumes that all shot clusters, except that of the guest, refer to the host. Occasionally the guest will speak for wide shots, thereby creating an error. Misinterpretation errors occur if the clustering information is incorrect. If a cluster is falsely identified at an endpoint location, then combinations between the endpoints will also be in error.

6. CONCLUSIONS

In this paper we describe a system that identifies COS locations in news programs. This system uses text and video information to identify these locations. It was shown that individual text processing and video processing is not adequate to determine these locations individually. This is because text information does not identify the speaker and the video determines shot boundary locations (not COS locations). It was shown that individual video analysis determines COS locations with 53.7% accuracy and individual text analysis determines COS

	Missing	False Alarm	Misclassification
Tags	60	16	26
Percentage	58.82%	15.69%	25.49%

Table 3. Errors produced by system.

locations with 32.28% accuracy. Our system aligns the TCOS and VCOS tags by examining combinations between a pair of endpoints. Our proposed system determines COS locations with 72.75% accuracy. It was shown that 50.46% of the identified locations are simultaneously contributed by both the text and video. A higher contribution will produce more endpoints thereby increasing the accuracy of our proposed system. Since 49.54% of the contribution is determined by individual text and video information, it may be necessary to design a trimodel system that uses text, video, and audio.

REFERENCES

1. C-SPAN web site, "http://www.cspan.org/."
2. L. Lu and H.-J. Zhang, "Speaker change detection and tracking in real-time news broadcasting analysis," *Proceedings of the 10th ACM International Conference on Multimedia*, December 1 – 6 2002, Juan-les-Pins, France, pp. 602 – 610.
3. Y. N. S. Satoh and T. Kanade, "Name-it: Naming and detecting faces in news videos," *IEEE Multimedia*, vol. 6, no. 1, pp. 22–35, January 1999.
4. G. Robson, *Inside Captioning.* Castro Valley, CA: CyberDawg Publishing, 1997.
5. FCC web site, "http://www.fcc.gov/cgb/consumerfacts/closedcaption.html."
6. C. Taskiran, A. Martone, and E. J. Delp, "Automated closed-captioning using text alignment," *Proceedings of SPIE Conference on Storage and Retrieval Methods and Applications for Multimedia 2004*, vol. 5307, 2004, San Jose, Ca, pp. 108–116.
7. C. Taskiran, J.-Y. Chen, A. Albiol, L. Torres, C. A. Bouman, and E. J. Delp, "ViBE: A compressed video database structured for active browsing and search," *IEEE Transactions on Multimedia*, vol. 6, no. 1, pp. 103–118, February 2004.
8. C. Taskiran, A. Albiol, L. Torres, and E. J. Delp, "Detection of unique people in news programs using multimodal shot clustering," *Proceedings of the International Conference on Image Processing (ISIP 2004)*, 2004, Singapore.