Augmenting Conversations through Context-Aware Multimedia Retrieval based on Speech Recognition

Kamer Ali Yuksel[1], Aytul Ercil[4]
Computer Vision and Pattern Analysis Laboratory
Faculty of Engineering and Natural Sciences
Sabanci University, Istanbul, Turkey
{kamer, aytulercil}@sabanciuniv.edu

Murat Celik Cansoy[2], Selim Balcisoy[3]
Computer Graphics Laboratory
Faculty of Engineering and Natural Sciences
Sabanci University, Istanbul, Turkey
{muratcansoy, balcisoy}@sabanciuniv.edu

ABSTRACT
Future's environments will be sensitive and responsive to the presence of people to support them carrying out their everyday life activities, tasks and rituals, in an easy and natural way. Such interactive spaces will use the information and communication technologies to bring the computation into the physical world, in order to enhance ordinary activities of their users. This paper describes a speech-based spoken multimedia retrieval system that can be used to present relevant video-podcast (vodcast) footage, in response to spontaneous speech and conversations during daily life activities. The proposed system allows users to search the spoken content of multimedia files rather than their associated meta-information and let them navigate to the right portion where queried words are spoken by facilitating within-medium searches of multimedia content through a bag-of-words approach. Finally, we have studied the proposed system on different scenarios by using vodcasts in English from various categories, as the targeted multimedia, and discussed how it would enhance people's everyday life activities by different scenarios including education, entertainment, marketing, news and workplace.

Categories and Subject Descriptors
H.5 [**Information interfaces and presentation**]: Multimedia Information Systems, User Interfaces, Voice I/O.

General Terms
Algorithms, Performance, Design, Reliability, Experimentation, Human Factors, Theory.

Keywords
Ambient intelligence, multimedia retrieval, speech recognition, spoken document retrieval, video-podcast.

1. INTRODUCTION
The amount of video content available on the Internet has increased dramatically over the past few years. Video streaming is currently the largest bandwidth consumer of the Internet at North America, by accounting for 37% share of all downstream Internet traffic and 41% of all mobile Internet traffic during peak hours, according to the Global Internet Phenomena Report of Fall 2010 [16]. Recently, Cisco estimated that 57% of global consumer Internet traffic would consist of video streams by 2014 [3]. More than half of the information on the Internet is formed by spoken documents that cover excessive amounts of academic, technical and news worthy information, which should be universally accessible.

Among these, vodcasts are one of the most important human-generated information over the Internet. In comparison with other multimedia content, their relatively longer duration and richer instructive content, which is generally conveyed by speech, makes their retrieval more essential. Accordingly, the retrieval of video streams and finding their relevant segments within them has gained importance, as these processes would help users to reduce their time and bandwidth costs.

Researchers investigated the social impact of watching video content and indicated that it promotes the social function of togetherness [10]. Especially within home, it would disallow natural distribution of activities among family members due to attention of viewing [19]. Moreover, the content selection and ownership provided by the Internet video re-formulated the experiences of sharing and viewing videos. However, the literature consistently refers searching for multimedia content as a source of frustration and time loss for users. Using conventional search engines (e.g. Google) to find multimedia content is limited by the textual metadata and suffers the lack of annotation for navigating to the query-related parts within the medium. For that reason, Girgensohn et. al. proposed the use of automatically generated visual summaries, which are composed of thumbnails, to aid user navigation [5].

Using traditional keyword-based search is often insufficient and the indication of relevant sections within each vodcast also becomes crucial due to their significant length. For instance, vodcasts of broadcast news are generally several hours in length and contain several different topics that are rarely represented using keywords. Hence, users seeking the recent news about a specific topic would not be able to find most of relevant vodcasts without going manually through all of them. Besides, video streams (e.g. video-blogs) that are generated and published using mobile devices, often lack detailed metadata due to the difficulty of entering them manually. The traditional information retrieval (IR) techniques can be applied to search the database once a transcription of vodcast is available. However, the manual transcription of such vodcasts is costly and generally there are no means for indexing their content and they are not annotated with links to the relevant points of the vodcast even if their transcription is available.

For that reason, several attempts have been made for creating automatic speech recognition (ASR) based spoken document retrieval (SDR) systems for broadcast news, academic lectures, call-center conversations and multimedia contents [14,20,9]. Most of them relied on the fact that the retrieval accuracy is not significantly affected by the word-error rate (WER) of the ASR for sufficiently long documents. For instance, Thong et. al. reported 69% accuracy for typical user queries despite high WER [20]. Renals et. al. increased the accuracy further by 12% using context-specific language models and dictionaries [14]. To sum up, ASR systems has been proven to be reliable for SDR and can also be used for annotating query-related video segments.

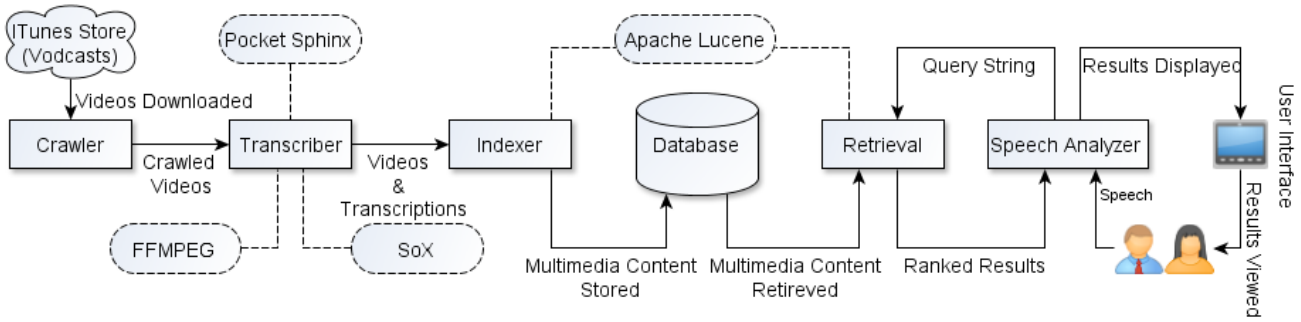

Figure 1. The architecture of the framework.

In this work, we have employed similar SDR techniques to allow users to search the spoken content of multimedia files rather than their associated meta-information and let them navigate to the right portion where queried words are spoken. Moreover, we have utilized the proposed method to implement a framework that augments conversations during daily life activities with related multimedia content. The retrieval results are organized and presented using an intelligent user interface with the ability of playing the segments of interest, which minimizes the time required to find related multimedia content and the amount of information that must be consulted. The primary contribution of this work across other SDR systems is the novel concept of augmenting conversations through indirect or bilateral interaction.

This paper is organized as follows: The following section presents possible application scenarios of the system and a discussion of the provided interaction during different use cases. Then, the architecture and implementation of the proposed method is described in-detail with the high-level explanation of the underlying podcast processing, indexing, and retrieval stages.

2. USAGE SCENARIOS

In this section, possible impacts of the proposed system are discussed through example usage scenarios from everyday life. During Human-human interaction (HHI), signals are used to communicate intention to initiate, availability for communication, or a listener understands what is being said [2]. Focusing on communicative aspects of Human-computer interaction (HCI) rather than cognitive ones, Belloti et. al. [1] argued that humans and systems should manage and repair their communications, and should be able to establish a shared topic, similar to HHI.

The novelty of the proposed system is that it supports indirect and bilateral interaction of the user, in addition to the direct interaction provided by the state-of-art architectures in multimedia retrieval. In other words, it allows humans and computers act like equal partners in dialog by perceiving the topic of the conversation through ASR. Then, it interprets the spoken content using SDR and augments the conversation through relevant vodcast footage. Afterwards, users can monitor the response of the system and navigate through critical frames of the ranked results for playing the segments of their interest at any time of their conversation. Lastly, users are able to manage and repair their communications with the system directly through its voice-control interface or indirectly by refining their conversations.

Firstly, the speech-based interface allows users to directly interact with the framework for the retrieval of spoken documents, similar to a voice-control interface. For example, a context-aware refrigerator, which is able to retrieve vodcasts for most relevant cooking recipes once users closes its door and utter the ingredients that are present within the refrigerator, can be implemented. More importantly, users may indirectly interact with the framework while focusing on other activities. In this case, they may use the system the provide visualization to their audience while focusing on their speech, as a form of assistance for their presentations. Accordingly, the learning efficiency and focus of the audience may also be positively influenced by the retrieved multimedia content due to its effect on their visual memory. For instance, the framework can be utilized through a large display for augmented guiding for cultural or historic virtual tours of cities or museums.

Similarly, users can indirectly interact with the system during their conversations among each other, in order to enhance their productivity and socialization. This type of interaction would be especially useful concerning mobile scenarios. According to researchers, some people have developed a habit of intuitively augmenting their daily conversations using multimedia content via their mobile devices. O'Hara et. al. argued that the socialization of people started to happen in front of mobile devices rather than only through them and discussed the extension of video experience to mobile devices [11]. Repo et. al. indicated that mobile video creates interesting new opportunities for social occasioning and helps users to avoid boring situations by sharing their experiences with other people [15]. Recent mobile phones (e.g. Samsung I7410, LG eXpo) have embedded pico-projectors to remove the limitations of the traditional mobile phone screen during such socializing activities [7].

The most interesting result of augmenting conversations is that the conversation among users would affect the output of the system while it may also be simultaneously affected by the output. Thanks to the facilitated bilateral HCI, the output is a joint accomplishment of users and system rather than users' mental model, as opposed to the previous case. In a bar or café atmosphere, vodcasts about their favorite movie actress or soccer-team can be projected to the tabletop by their mobile devices while users are talking about them, leading to more fruitful and entertaining conversations among people. In similar places, advertisements, which are relevant to the conversation of people, can be retrieved for stronger publicity and marketing of brands. In addition, it may be used for collaborative activities at workplace, such as brainstorming or supporting ideas while in a meeting. For example, brokers or entrepreneurs may utilize it to observe latest broadcast news, discussions or comments about their potential investments.

3. METHODOLOGY

The architecture of the framework (Figure 1) consists of several phases that are described in the following part of this section. The crawling, pre-processing and indexing phases are performed using

a web server while the client device of the user, which may be a mobile or embedded device, performs the retrieval and streaming of vodcasts. The framework is able to process more than eight hours of seventy-two vodcasts per day through a dual-CPU configuration, using the large vocabulary of HUB4 [13] consisting of 64,000 words and a trigram language model. Hence, the final collection, which is indexed in one-week using two dual-core machines, includes nearly thousand vodcasts equally distributed over different categories of iTunes store [8], such as Business, Education, Hobbies, News & Politics, etc. Preliminary experiments have shown promising retrieval accuracies, even in the presence of a high WER up to 25% for each category. The interface of the framework can be utilized through any device equipped with a display and microphone, for several use-cases.

3.1 Crawling and Pre-Processing

Vodcasts are series of digital video files that are episodically distributed using RSS enclosures, generally referred as channels, where a list of all video files currently associated with a given series is maintained centrally on the distributor's server. In this phase, the engine automatically extracts up-to-date content from each channel within well-known podcast directories by downloading vodcasts and their associated meta-information. This phase is performed periodically to process new vodcasts and make them available in the system for people to search, as soon as they are published. Once vodcasts are downloaded, their audio content are automatically extracted and segmented into individual non-silence regions of pre-defined minimum and maximum durations using modules from open-source software packages, FFmpeg [4] and SoX [18] respectively.

Raw audio segments are then converted into texts to obtain the text transcription for each vodcast using PocketSphinx [6], which is a lightweight speech recognition engine developed at Carnegie Mellon University. PocketSphinx takes raw audio segments, splits them into utterances, omits any silent segments and then tries to recognize what's being said in each utterance by taking all possible combinations of words and matching them with the audio. Several vocal features are extracted from the audio file in frames of 10-milliseconds to choose the best matching combination. In order to match spoken words, it employs two models: acoustic and language. The acoustic model contains acoustic properties for each senone, which are context-independent models that contain most probable feature vectors for each phone. The phonetic dictionary contains a mapping from words to phones. The language model is used to restrict word search by defining which word could follow previously recognized words and by stripping words that are not probable.

3.2 Indexing

The indexing and retrieval phases are performed using an open-source high-performance, scalable information retrieval software library in Java, which is called Apache Lucene. After speech-to-text processing, the obtained metadata and text transcripts are then indexed into a special data structure to optimize the speed and performance in retrieving relevant documents for a search query. English-language stop words are filtered from transcripts and inflected words are reduced to their stem, base or root form to reduce the search space and the index size. The index includes the term frequency and inverse document frequency (TF-IDF) weight and the position (relative to the duration of the vodcast) of each recognized word. The weight is a statistical measure to help in ranking the relevance of vodcasts to a query and the position is used to identify word proximity to support searching for phrases. Moreover, position data is later used to calculate the relevance distribution of the timeline according to a query to decrease the amount of each vodcast that should be watched by users to identify their interest to it. The relevance distribution allows users to navigate automatically between the most relevant intervals to minimize the time spent and the bandwidth costs considering that vodcasts are streaming media.

HCIR'11, October 20, 2011, Mountain View, California, USA.

3.3 Retrieval

During each session of conversation established by users, the ambient audio is continuously buffered by a microphone and converted into text using PocketSphinx to form search queries using recognized texts, in real-time. Periodically, new queries are automatically formed and sent to the engine running on the web server. Using the same speech recognition engine along with same models on both sides helps handling out-of-vocabulary words, as they would probably be recognized in the same way. For ranking, Lucene uses the Vector Space Model (VSM) that represents text documents as vectors of weighted terms and determines how relevant a given document is to a query by comparing the deviation of angles between each document vector and the original query vector.

Instead, BM25 scoring function [12] is preferred due to its probabilistic model driven by the uncertainty inherent in search queries, which is more suitable concerning the possibility of a high WER. In fact, the query is a series of weighted set operations on the documents that are listed against each term. The default operation is the union but the intersection may also be employed according to the application scenario. Thus, the scoring function is modified to exploit all words within the hypothesis space, instead of using the set of best hypotheses for each utterance, by imposing an additional parameter for weighting individual scores by their posterior probabilities (Equation 1). The engine examines the index according to specified query and provides a ranked list of best-matching vodcasts, along with their metadata, stream links, relevance distributions and animated thumbnails composed of low-resolution screen captures from the relevant points.

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1) \cdot c_i}{f(q_i,D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{avgdl}\right)}$$

$$IDF(q_i) = log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}$$

Equation 1. Q: set of all hypotheses q_n, c_i: the posterior probability of q_i, $f(q_i,D)$: q_i's term frequency in the document D, |D|: number of words within D, avgdl: average |D| in the collection, N: total number of documents, $n(q_i)$: number of documents containing q_i, k_1,b: free parameters.

3.4 Interface

To be able to interact with speech, a set of words as spoken commands for users are defined to navigate automatically between and through ranked videos without interfering their conversation. The ranked list of results, in which new items added

after each query update, are presented orderly by queuing their thumbnails. Users navigate through the list and change the active item that is presented with a larger size animated thumbnail in the center of the interface and, annotated by the metadata of the vodcast. Users inspect the annotated metadata such as the title or category of the vodcast and captures from possibly relevant scenes to their conversation and then decides if the video is relevant for their talk. If it is thought to be relevant, they may switch to full-screen mode where vodcast starts being streamed by the first critical point.

The timeline of the video-player is enhanced with colored markers to display the computed relevance distribution function and mentions of each queried term. They may still use the same set of commands but now to navigate within critical points and an additional set to control the video player and to switch back to results screen. In full-screen mode the retrieval is restricted to the content of video (which is being watched) and the markers within its timeline are updated accordingly. Similarly, words that are not previously mentioned during the conversation but having high TF-IDF values in the current results are scrolled below the interface in a tag-cloud fashion to help the user refining his query.

4. CONCLUSION

In this work, we have proposed a novel methodology for augmenting conversations during daily life activities with related multimedia content by utilizing automated speech-based multimedia retrieval techniques. Moreover, we have utilized the proposed method to implement a framework that presents relevant video-podcast footage, in response to spontaneous speech and conversations. Finally, we have studied the proposed method on potential scenarios by using video-podcasts in English from various categories as the targeted multimedia and discussed how it may be used to enhance people's everyday life activities by different scenarios. Preliminary experiments have shown satisfactory retrieval accuracy despite high speech recognition error rates, as in state-of-art examples.

In conclusion, we believe that the proposed system may become a tool to introduce and measure the effect of serendipity to ordinary conversations. Serendipitous discoveries are of great importance in science and technology. The uncertainties introduced by the workings of the framework may enable encounters with closely related videos sometimes from unexpected domains. Users may change the direction of their conversation according to those visual stimuli or may ignore them. The presented framework may provide a significant test environment to conduct controlled user studies on collaborative interactive environments.

5. REFERENCES

[1] Bellotti, V.M.E., Back, M.J., Edwards, W.K., Grinter, R.E., Lopes, C.V. and Henderson, A. (2002) Making Sense of Sensing Systems: Five Questions for Designers and Researchers. In Proc. CHI 2002, ACM Press, 415-422.

[2] Button, G., & Casey, N. Generating topic: The use of topic initial elicitors. In J. Atkinson & J. Heritage (eds.) Structures of Social Action. Studies in Conversation Analysis. Cambridge University Press, (1984) 167-189.

[3] Cisco Visual Networking Index: Forecast and Methodology, 2009-2014.

[4] FFmpeg. http://www.ffmpeg.org.

[5] Girgensohn, A., Boreczky J. and Wilcox L. Keyframe-Based User Interfaces for Digital Video. In Proc. IEEE Computer, Vol. 34(9), pp. 61-67, 2001.

[6] Huggins-Daines, D. et. al., PocketSphinx: A free, real-time continuous speech recognition system for hand-held devices. In Proc. ICASSP'06, 185–188.

[7] Hürst, W., Götz G. and M. Welte, Interactive video browsing on mobile devices. In Proc. ACM Multimedia'15, 2007, p. 247–256.

[8] iTunes Store. http://www.apple.com/itunes/whats-on/

[9] Nexidia/Fast-Talk. http://www.nexidia.com.

[10] O'Brien, J., Rodden, T., Rouncefield, M. and Hughes, J. (1999) At home with the technology: an ethnographic study of a set-top-box trial. ACM Transactions on. Computer-Human Interaction, 6(3), 282-308.

[11] O'Hara, K., Mitchell, A. S., Vorbau A. Consuming Video on Mobile Devices. In Proc. CHI 2007, ACM Press, 857-866.

[12] Pérez-Iglesias, J. R. Pérez-Agüera, V. Fresno, and Y. Z. Feinstein, "Integrating the Probabilistic Models BM25/BM25F into Lucene," In Proc. CoRR 2009.

[13] Placeway et. al., The 1996 hub-4 sphinx-3 system. In Proc. DARPA Speech recognition, 1997, 85-86.

[14] Renals, S., Abberley, D., Kirby, D. and Robinson T. Indexing and retrieval of broadcast news. Speech Communication, vol. 32, no. 1-2, p. 5–20, 2000.

[15] Repo, P., Hyvonen, K., Pantzar, M and Timonen P. (2004) Users Inventing Ways To Enjoy New Mobile Services - The Case of Watching Mobile Videos. In Proc. HICSS'04, vol. 4, 40096c.

[16] Sandvine Incorporated. Fall 2010 Global Internet Phenomena Report. http://sandvine.com/news/global_broadband_trends.asp.

[17] Schegloff, E., Jefferson, G., & Sacks, H. The preference for self-correction in the organization of repair in conversation. Language, 53, (1977), 361-382.

[18] SoX - Sound eXchange. http://sox.sourceforge.net.

[19] Taylor, A. and Harper, R. (2003) Switching on to switch off. In Harper, R. (Ed.) Inside the Smart Home. London: Springer-Verlag.

[20] Thong, J. -M. et. al., SpeechBot: an experimental speech based search engine for multimedia content on the web. IEEE Transactions on Multimedia, vol 4.1, 88-96, 2002.