On the Use of Automatic Speech Recognition for Spoken Information Retrieval from Video Databases

Luis R. Salgado-Garza and Juan A. Nolazco-Flores

Computer Science Department, ITESM, Campus Monterrey
Av. Eugenio Garza Sada 2501 Sur, Col. Tecnológico
Monterrey, N.L., México, C.P. 64849
{lsalgado,jnolazco}@itesm.mx

Abstract. This document describes the realization of a spoken information retrieval system and its application to words search in an indexed video database. The system uses an automatic speech recognition (ASR) software to convert the audio signal of a video file into a transcript file and then a document indexing tool to index this transcripted file. Then, a spoken query, uttered by any user, is presented to the ASR to decode the audio signal and propose a hypothesis that is later used to formulate a query to the indexed database. The final outcome of the system is a list of video frame tags containing the audio correspondent to the spoken query. The speech recognition system achieved less than 15% Word Error Rate (WER) and its combined operation with the document indexing system showed outstanding performance with spoken queries.

1 Introduction

The most natural way to transmit information among humans is by voice, unfortunately a permanent recording of such class of information has been less common than text archiving mainly because two reasons: the large amount of storage required for acoustics and its dificulties for indexing. Nowadays, massive storage technology is more affordable than ever, new algorithms for data compression permit to record several hours of video and audio in just only tenths or thousands of megabytes. The boost of broadcast communication via television, radio and Internet has promoted the use of digital audio and video (multimedia) recording as one of the fastest and most accessible resources for information transmission and storage [1]. Trends are that in the near future, due to the amount of information contained in multimedia documents, multimedia will surpass text based documents as the preferred archiving method for information storage [2].

In recent years, new techniques for text indexing and automatic information retrieval have been developed [3] and succesfully used over the Internet, focusing on digital documents containing text information. However, the future of these search engines requires its use on multimedia databases, if not, its application

A. Sanfeliu et al. (Eds.): CIARP 2004, LNCS 3287, pp. 381–385, 2004.

in mobile devices and hand-busy environments will be very limited. At present, several research groups have conducted efforts to develop search engines for multimedia repositories of information [4], commonly known as multimedia digital libraries. However, there is still much work to be done, particularly regarding speech information and spoken queries [5, 6].

This paper presents the architecture of a spoken information retrieval system and its application to video databases, using acoustic information for indexing. The system is structured using two main components: an automatic speech recogniton (ASR) system and a document indexing software. The input to the system is a speech signal, containing the words included in the video frames of interest to the user, the output is a sequence of indexes on the video for such frames. The system was trained in spanish and tested in the same language but using spontaneous speech and different speakers.

The organization of the document is the following, section 2 presents the general architecture and explanation of the modules of the system. Subsection 2.1 explains the acoustic and language models training procedure, subsection 2.2 presents the methodology used for the video database indexing while subsection 2.3 focuses in the spoken query procedure. Section 3 discuses the results and conclusions are presented in section 4.

2 System Architecture Description

Our system architecture is structured in three modules (Fig. 1), acoustic and language models training, video's acoustic information transcription and indexing and spoken query processing. A detailed description of each of these modules is presented in this section.

2.1 Acoustic and Language Models Training

In order to perform speech recognition we train triphone acoustic models using the CMU SPHINX-III system, that is a HMM-based automatic speech recognition environment for continuous speech and large vocabulary tasks. A speech database, containing several hours of audio, produced by different speakers, is used for training. First the analog signal, from each audio file in the database, is sampled and converted to MFCC coefficients, then the MFCC's first and second derivatives are concatenated [7], the number of MFCC is 13 then the total dimension of the feature vector accounts for 39.

The acoustic models are finally obtained using the SPHINX-III tools. This tools use a Baum-and-Welch algorithm to train this acoustic models [8]. The Baum-and-Welch algorithm needs the name of the word units to train as well at the label and feature vectors. The SPHINX-III system allows us to model either discrete, semicontinuous or continuos acoustic models. In SPHINX-III system tools allow to select as acoustic model either a phone set, a triphone set or a word set.

The language models (LM) are obtained using the CMU-Cambridge statistical language model toolkit version 2.0 [9]. The aim of the LM is to reduce

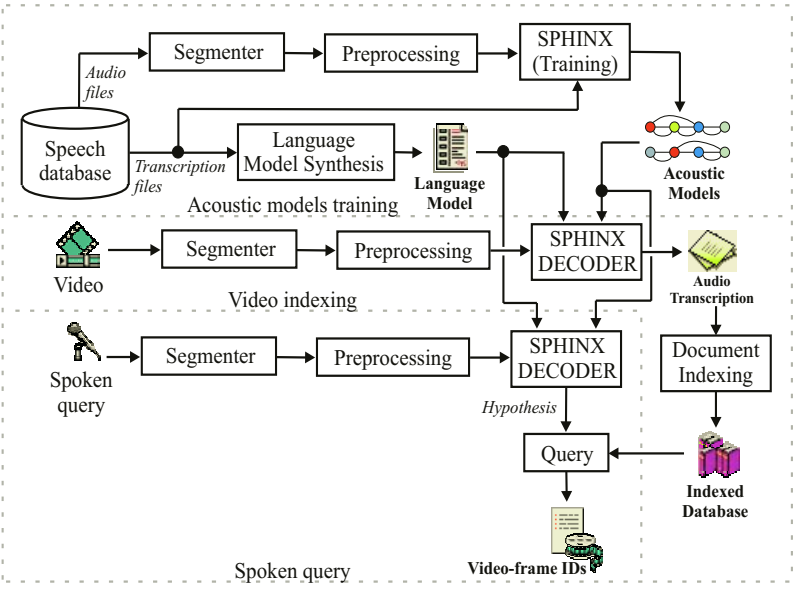

Fig. 1. Architecture of the spoken information retrieval system applied to video inexing

the perplexity of the task, by predicting the following word based in the words' history. Several techniques can be used to synthetize a LM [10], among them N-grams is the easiest technique with very good results. If all the n-grams are not contained in the language corpus, smoothing techniques need to be applied. In the CMU-Cambridge language model toolkit, unigram, bigrams or trigrams can be configured for this tool, as well as four types of discount model: Good Turing, Absolute, Linear and Witten-Bell.

2.2 Video Indexing

For our information retrieval system, the querable repository will be the audio into a video file. In order to build our indexed database we first need to get the transcript from the audio signal. The initial stage of video indexing deals with segment boundaries definition in the acoustic signals, for this we use the CMUseg [11] tool. Then, audio frames are preprocessed in order to calculate 13 MFCC coefficients, computing also first and second derivatives. The 39 sized feature vector for each frame is used as the input for the decoder. Each hypothesis from the decoder represents its best transcriptions for the audio signal, the whole set is stored as the audio transcription archive.

The transcription file is indexed using the MG [2] tools suite, that is a large scale inverted index based text retrieval system. The indexed database spans over the entire audio signal extracted from the video file, including every word

detected by the speech recognizer. This database indexes the video in a frame-time bases and will be used for the queries of the next stage.

2.3 Spoken Information Retrieval

The spoken query module uses a close talk microphone to capture a query, composed by isolated words, uttered by a speaker. The audio signal is stored as a `wav` file. The acoustic stream is preprocessed to calculate MFCC coefficients and feeded into the speech recognition decoder. The best hypothesis is taken as the transcription for the query using the MG tools. The output from the system is a sequence of file tags, pointing to specific time of the video where each queried word appears.

3 Experiments

The configuration of the SPHINX-III system for our experiments used 13 mel-frequency cepstral coefficients (mfcc) and also their first and second deritatives, therefore the feature vector accounted for 39 elements. The speech lower and higher frequencies were stablished at 300 Hz and 7,000 Hz, respectively. The frame rate was set to 50 frames per second using a 30ms Hamming window. A 512 samples FFT length was used and the number of filterbans was set to 40. Five states continuous HMM was used as acoustic modeling technique with mixtures of 16 gaussians per state. The training data base for the speech recognition system had a vocabulary of 22,398 different words (409,927 words in 28,537 sentences). Triphones were used as the word unit and language modeling is based in word trigrams.

The baseline configuration of our speech recognition system was as reported in [10], for our latest experiments we used a new version of the SPHINX-III system leading us to the scores reported in table 1.

Table 1. Word Error Rates for the speech recognition system

Experiment	Language Weight	WER
Baseline	10.0	26.44%
Experiment 1	9.5	14.98%
Experiment 2	10.0	14.52%

With so emboldening results from the speech decoder, we were motivated to test the combined efficency with the MG indexing and retrieval system. For our experiments were performed spoken queries using single and multiple words, uttered by different speakers, independent form the ones in the training set for the ASR system. We found that the spoken query always performed correctly, reporting every video frame segment where the queried words were actually present.

4 Conclusions

As seen from the results, the overall performance of the system showed that the integration of a speech recognizer, a text based document indexing and retrieval tools comprise an effective architecture for a spoken query system on multimedia databases. However, more work is needed in order to increase the robustness of the system to OOV words and acoustic noise conditions. Worth to try is to use noise compensation methods, syntactically inspired language model techniques [10, 12] and n-Best list hypothesis from the decoder and evaluate the effect of these in the information retrieval cogency. We also propose the use of phonetic features into the document indexing algorithms (multifeatured information indexing), because the use of this kind of acoustic information could lead to more comprehensive classification system for spoken information retrieval.

The encoraging results shows the reliability of the system architecture and settles a testbed for future applications as Internet based user interface for spoken search engines and information storage and retrieval in mobile applications.

References

1. Chen, B., H.M. Wang, and L.S. Lee. "Retrieval of Broadcast News Speech in Mandarin Chinese Collected in Taiwan using Syllable-Level Statistical Characteristics", Proceedings of ICASSP-2000.
2. Witten, I.H., Moffat, A., and Bell, T.C. Managing gigabytes: compressing and indexing documents and images. Van Nostrand Reinhold, New York (1994).
3. D.R.H. Miller, T. Leek and R.M. Schwartz. "A hidden Markov model information retrieval system", In Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), pages 214-221, 1999.
4. Witten, I.H., Don, K.J., Dewsnip, M., and Tablan, V., "Text Mining in a digital library", Journal of Digital Libraries, 2003 (In Press).
5. Wolf, P.P.; Raj, B., "The MERL SpokenQuery Information Retrieval System: A System for Retrieving Pertinent Documents from a Spoken Query", IEEE International Conference on Multimedia and Expo (ICME), Vol. 2, 317-320, August 2002.
6. K. Spärck Jones, G. J. F. Jones, J. T. Foote, and S. J. Young. "Experiments in spoken document retrieval", Inf. Processing and Management, 32(4):399-417, 1996.
7. Deller, J.R., Proakis, J.G., Hansen, J.H.L., Discrete-Time Processing of Speech Signals, Prentice Hall, Sec. 6.2, 1993.
8. Dempster, A.P., Laird, N.M., Rubin, D.B., "Maximum likelehood for incomplete data via the EM algorithm", J. Roy. Stat. Soc., Vol. 39, No. 1, 1-38, 1977.
9. Clarkson, P., Rosenfeld, R., "Statistical Language Modelling using the CMU-Cambridge Toolkit", Proceedings of Eurospeech, Rodhes, Greece, 1997, 2707-2710.
10. Luis R. Salgado-Garza, Richard M. Stern, Juan Arturo Nolazco-Flores. "N-Best List Rescoring Using Syntactic Trigrams", Proceedings of MICAI 2004, Advances in Artificial Intelligence, Springer-Verlag (LNAI 2972:79-88)
11. M. Seigler, U. Jain, B. Raj, R. Stern. "Automatic segmentation, classification, and clustering of Broadcast news audio", Proc. Of the DARPA speech recognition workshop, February 1997.
12. Hsin-Min, W., Berlin, C., "Content-based Language Models for Spoken Document Retrieval", International Journal of Computer Processing of Oriental Languages (IJCPOL), Vol. 14, No.2, 2001.