Access to Large Spoken Archives: Uses and Technology

Moderator Dagobert Soergel

College of Information Studies, University of Maryland, College Park, MD 20742 dsoergel@umd.edu.

Samuel Gustman

Survivors of the Shoah Visual History Foundation, P.O. Box 3168, Los Angeles, CA 90078-3168, sam@vhf.org**.**

Mark Kornbluh

Department of History, Michigan State University and The National Gallery of the Spoken Word (NGSW). mark@mail.matrix.msu.edu

Bhuvana Ramabhadran

Human Language Technologies Group, IBM T.J. Watson Research Laboratory, bhuvana@us.ibm.com.

Discussant: Jerry Goldman

Political Science Department, Northwestern University,

Co-chair DELOS EU-NSF Working group on Spoken Archives j-goldman@northwestern.edu

With recent advances in information technology, digital archiving is emerging as an important and practical method for capturing the human experience. Large amounts of spoken materials and audiovisual materials in which speech is an important component are becoming available. This panel will discuss the uses of these materials for education, information retrieval and dissemination, and research, the requirements that arise from these uses, and speech recognition and retrieval technologies being developed to meet these requirements. These materials have tremendous potential for enriching the presentation of information in education, newscasts and documentaries, but retrieval from and access to these large repositories pose significant challenges. The panel will provide an overview of these issues.

ASIST 2002 Proc. of the 65. ASIST Annual Meeting, Philadelphia, PA, November 18-21, 2002). Medford, NJ: Information Today. p. 469-470

Samuel Gustman

The collection of the Survivors of the Shoah Visual History Foundation and its uses.

The Shoah foundation has assembled 116,000 hours of digitized interviews in 32 languages from 52,000 survivors, liberators, rescuers and witnesses of the Nazi Holocaust. Based in part of a study of requests, this presentation will discuss present and future uses of this collection with emphasis on the promotion of tolerance. It will then analyze the retrieval and representation requirements that arise from these uses and discuss these requirements in light of the special nature of the materials.

Mark Kornbluh

The collection of The National Gallery of the Spoken Word and its uses

NGSW is creating a significant, fully searchable online database of spoken word collections spanning the 20th century to be used for development of a rich set of exhibits and educational curricula that fully incorporate sound files. This presentation will describe the collection and discuss novel uses of its content.

Bhuvana Ramabhadran

Issues in applying automated speech recognition to difficult speech

Automated speech recognition (ASR) is a fundamental technology for accessing large corpora of speech that defy manual transcription or even detailed manual indexing. But many of these collections, such as the Shoah Foundation collection, contain speech that is often difficult even for a human listener: spontaneous and emotional speech; whispered speech; speech with background noise and frequent interruptions; speech from elders; speech that switches between languages; heavily accented speech; speech with words such as names, obscure locations, unknown events, etc. that are outside the recognizer lexicon; disfluent speech. The presentation will discuss techniques being developed in the MALACH project to deal with these difficulties

Dagobert Soergel

Issues in retrieval from large spoken archives

Access to large spoken archives poses significant problems particularly in light of requirements such as finding particularly vivid testimonies or material suitable for certain age groups in addition to topical access. This requires techniques of automatic classification developed from a training corpus and methods whereby users can share indexing and evaluation of material. Creation and presentation of suitable surrogates also poses problems well beyond creating abstracts of written text. This presentation will provide a framework for investigating these issues in both monolingual and multilingual contexts.

References

Shoah Foundation www.vhf.org

National Gallery of the Spoken Word www.ngsw.org

Malach project
http://www.clsp.jhu.edu/research/malach/malach.html