Transcript-Free Search of Audio Archives for the National Gallery of the Spoken Word

[Extended Abstract]

John H.L. Hansen
University of Colorado
Cntr. for Spoken Lang. Res.
Boulder, CO 80309 USA
jhlh@cslr.colorado.edu

J.R. Deller, Jr.
Michigan State University
Dept. Elec. & Comp. Engr.
E. Lansing, MI 48824 USA
deller@msu.edu

Michael S. Seadle
Michigan State University
E308 Main Library
E. Lansing, MI 48824 USA
seadle@msu.edu

ABSTRACT

The National Gallery of the Spoken Word (NGSW) project is creating a carefully organized on-line repository of spoken-word collections spanning the 20th century. Unprecedented technical challenges are inherent in the development of an archive of such extensive scale and diversity. This paper describes research on the development of text-free search-engine technology used to locate requested content in the audio records. A companion paper in these proceedings addresses watermarking technologies for copyright protection.

Categories and Subject Descriptors

H.3.7 [**Info. Storage & Retrieval**]: Digital libraries—*Sys. issues*; E.5 [**Data**]: Files—*Sorting & searching*

1. THE NGSW

The National Gallery of the Spoken Word (NGSW) project is sponsored by the Digital Libraries II Initiative. The goal of the NGSW project is to create a carefully organized on-line repository of spoken word collections, based largely upon the renowned Vincent Voice Library at Michigan State U. (MSU). The collaborative project among specialists in the library sciences, humanities, engineering, and education, will provide the first large-scale repository of its kind. MSU is creating the NGSW in partnership with several universities and agencies, the U. Colorado–Boulder (CU) representing the key collaborator in the engineering developments described here. Further information is found on the NGSW web site at URL www.ngsw.org.

This paper is one of two papers appearing in these proceedings whose purpose is to describe the key technical research issues that are being investigated by academic engineers and allied colleagues in connection with the NGSW. The use of the term "academic engineers" is deliberate in an effort to distinguish the work described here from the significant amount of engineering and technical development (principally creation and maintenance of hardware and software infrastructure) essential to the existence of such a vast resource. In particular, this paper describes research focused on the construction of integrated search mechanisms for locating audio resources in the NGSW collection. This work is being led by researchers at the Center for Spoken Language Research (CSLR) at CU in collaboration with the MSU's Speech Processing Laboratory and MSU Libraries. The companion paper describes research centered at MSU on the development of "watermarking" technologies for secure, efficient, delivery of aural material [6].

JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.

2. A CHALLENGING SEARCH PROBLEM

A long-term goal of the NGSW project is to be able to automatically "mine" audio resources for material that is responsive to gallery-users' queries. This feature will be of particular utility to researchers at all levels - from primary school students through professional scholars, journalists, and authors. This aspect of the engineering research is fraught with daunting challenges and uncertainties, as the NGSW represents an audio database whose large scale and content diversity have never been remotely approached.

Although the problem of audio stream search is relatively new, it is related to a number of previous research problems. Many systems developed for audio search, however, assume the existence of associated text or a clean audio stream [5]. Direct information retrieval via audio mining generally focuses on relatively noise-free, single-speaker recordings.[1] Alternative methods have included ways to time-compress or modify speech to allow listeners the ability to skim more quickly through recorded audio data [1]. While a keyword spotting system can generally be used for topic, gisting, or phrase search applications, the system must be able to recover from errors in both a user's text query and in rank-ordered phrase hypotheses in the stream. Phrase search focuses more on locating a single requested occurrence, whereas keyword/topic spotting systems assume a number of possible outcomes. Great strides have also been made in large-vocabulary, continuous speech recognition, suggesting the use of forced transcripts of the NGSW material. While this

[1] speechbot.research.compaq.com and www.dragonsys.com.

may be a manageable task for even the larger databases used for speech research (e.g., the Broadcast News Database of 100 hours [9]), the initial offering for NGSW will be 5000 hours (with a potential of 40,000+ total hours based on the existing collection alone), and it is not feasible to achieve accurate forced transcription, even if the text data were available. Further discussion of the NGSW database complexity is found in [7].

3. SYSTEM CONCEPT

A diagram of the current search-engine concept is shown in Fig. 1. The system employs a multifaceted approach to solve this complex search problem. Component modules include[2]

Natural language parsing (NLP): In response to a user query, an N-best parser will be used to rank order audio streams for search, and to correct ill-formed requests. Related research involves the development of adaptive assessment of search success likelihood based on query content.

Environment characterization: An environment processor will be used in conjunction with meta-tag information for the input stream under test. This processor will identify distortion type: acoustic background or recording media noise, restricted channel, reverberation, multiple speakers, etc.

Three adaptation modules: •*Noise adaptation* measures will be based on the results of the environment characterization. The current focus is on rapid methods for parallel model combination [8]. • For repeated searching of material from the same speaker, *hidden-Markov model* (HMM) *adaptation* [4] is performed. Limited data for adaptation are available. Methods being considered include selective training [3] and decision-bias correction [2].

Restricted channel adaptation: The bandwidth of audio streams from Edison cylinder disks, for example, is very small ($\sim$ 2 kHz). We are investigating methods to normalize feature sets for models trained with 8 and 16kHz speech.

Speech enhancement: A set of speech enhancement algorithms will be available for user for quality improvement, and audio feature enhancement prior to stream search.

HMM recognition search (e.g., [4])**:** Following adaptation, an HMM phrase search will be performed. If the environment classifier determines that front-end enhancement could be effective, the input parameter set will be appropriately modified. Detected phrase sets will be rank-ordered using confidence measures and the NLP processor. Essential higher-level knowledge to be incorporated into the recognition search will include extensive metadata records that

[2]For readers from other disciplines, we offer a few general remarks to clarify the need for each of the system modules. Roughly speaking, "NLP" refers to the use of preprogrammed rules of syntax and grammar of a language to expedite speech search and recognition. The "environment" of an utterance refers to any quantifiable characterization of the utterance (e.g., background noise, accents) that would not be evident (e.g., words), or inherent (e.g., grammar), in a written transcript of the message. "Restricted channel" refers to the different frequency ranges that characterize different recording media – a factor that must be compensated for in comparing them to template "features" derived from full audio band data. Speech enhancement algorithms are used to improve the quality of digitized speech which is degraded by various processing measures. The "HMM" is the principal tool used in modern speech processing to capture the statistical structures among sounds and word orderings in a language. For further information, see, for example, [4].

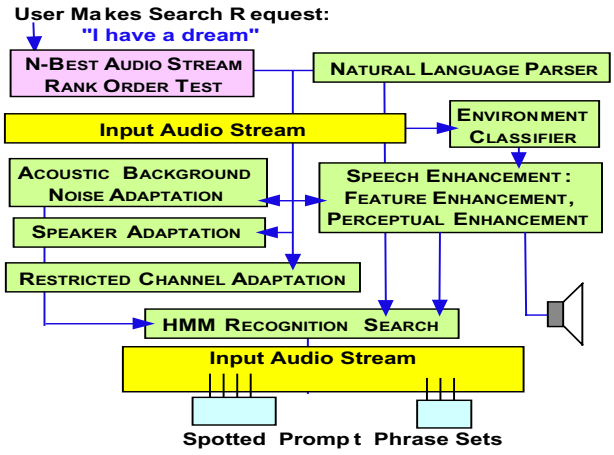

Figure 1: Flow diagram of the audio-stream search engine under development for the NGSW.

are being developed by the MSU Libraries using Encoded Archival Description (EAD). This blending of information sources – the stochastic engineering framework with the deterministic archival information – poses one of the interesting problem areas arising out of this digital library research.

Some preliminary results on the application of these search modules to NGSW speech data are found in [7].

Acknowledgments

This material is based upon work supported by the National Science Foundation under Cooperative Agreement No. IBIS-9817485. Opinions, findings, or recommendations expressed are those of the authors and do not necessarily reflect the views of the NSF.

4. REFERENCES

[1] B. Arons. A system for interactively skimming recorded speech. *ACM Trans. Computer-Human Interaction*, 4:3–38, 1997.

[2] L. Arslan and J. Hansen. Likelihood decision boundary estimation between HMM pairs. *IEEE Trans. Speech & Audio Process.*, 6:410–414, 1998.

[3] L. Arslan and J. Hansen. Selective training in HMM recognition. *IEEE Trans. Speech & Audio Process.*, 7:46–54, 1999.

[4] J. Deller, Jr., J. Hansen, and J. Proakis. *Discrete-Time Processing of Speech Signals*. IEEE Press, 2d edition, 2000.

[5] J. Foote *et al.* Talker-independent keyword spotting for information retrieval. *Proc. Eurospeech*, volume 3, pp. 2145–2149, 1995.

[6] J. Deller, Jr., A. Gurijala, and M. Seadle. Audio watermarking techniques for the National Gallery of the Spoken Word. Elsewhere in these proceedings.

[7] J. Hansen *et al.* Stream phrase recognition for the NGSW: One small step. *Proc. Int. Conf. Spoken Lang. Process.*, vol. 3, pp. 1089–1092, Beijing, 2000.

[8] R. Sarikaya and J. Hansen. PCA-PMC: A novel use of *a priori* knowledge for fast parallel model combination. *Proc. IEEE ICASSP*, 2000.

[9] P. Woodland *et al.* Experiments in broadcast news transcription. *Proc. IEEE ICASSP*, pp. 909–912, 1998.