4. Automatic Spoken Document Processing for Retrieval and Browsing

Ciprian Chelba, Google, and T. J. Hazen, MIT

Ever increasing computing power and connectivity bandwidth together with falling storage costs is resulting in overwhelming amounts of multimedia data being produced, exchanged, and stored. One key application area in this realm is the search and retrieval of spoken audio documents. As storage becomes cheaper, the availability and usefulness of large collections of spoken documents is limited strictly by the lack of adequate technology to exploit them. Manually transcribing speech is expensive and sometimes outright impossible due to privacy concerns. This leads us to exploring an automatic approach to searching and navigating spoken document collections. This tutorial will present an overview of speech transcription, indexing, and search technologies for spoken documents, with an emphasis on a corpus containing recorded academic lectures. The tutorial will point out general problems in this area and suggest possible solutions. Included in the tutorial will be a discussion of scenarios and previous projects in the area of spoken document retrieval, issues of automatic transcription of long audio files, and techniques for the indexing and retrieval of spoken audio files.

4.1 Tutorial Outline

1. Introduction: Scenarios/Previous Work/Corpora
   - Scenarios:
     * Economic considerations for viability of such technology
     * Scenarios where technology is not expected to be useful
     * Scenarios where technology is expected to be useful:
   - Broadcast News
     * Characteristics
     * Meta-data annotation
     * Past work (HP SpeechBot, BBN, TREC, PodZinger, etc.)
   - Academic & Scientific Lectures
     * Examples (OCW, CSJ, MICASE)
     * Characteristics
     * Challenges and opportunities
2. Automatic Speech Transcription
   - Overview of speech recognition models and processing
   - Vocabulary Issues
     * Examination of vocabulary statistics and coverage
     * Vocabulary expansion from supplemental materials
   - Language Modeling Issues
     * Spontaneous conversational speech vs. read speech
     * Appropriateness of written materials
     * Language model adaptation
   - Acoustic Modeling Issues
     * Speaker independent modeling
     * Speaker dependent modeling
     * Supervised and unsupervised adaptation
   - Out-Of-Vocabulary (OOV) modeling
     * Methods for recognizing OOV words
     * Phonetic transcription of OOV words

Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume, pages 305–306,
New York City, June 2006.

3. Audio Retrieval
   - Overview of text retrieval algorithms:
     - TF-IDF/vector space methods
     - Probabilistic methods
     - Large scale web search (Google)
     - Inverted indexing; query processing/language.
   - Speech recognition lattices:
     - Word/phone/OOV-models for generation
     - Lattice accuracy vs. 1-best accuracy
   - Query processing (OOV problem)/language:
     - ”Soft”-indexing with pruning to control size
     - Combine sub-word and word-level indexing/recognition results
   - Relevance scoring:
     - Proximity
     - Incorporating multiple data streams: speech, text, title, author, abstract, etc.
     - Tuning precision/recall at query run-time
   - Evaluation:
     - Basic Metrics: Precision/Recall
     - Ordered list metrics: Kendall-Tau, Spearman
     - TREC measures and package (Mean Average Precision, R-precision)
     - Issues with evaluating speech data
   - User interface:
     - Issues in consuming speech (as opposed to text, images)
     - Pro’s/con’s for displaying transcription
     - Navigation in long documents with errorful transcriptions
     - Segmentation: topic boundaries, keywords, summaries

4.2 Target Audience

This tutorial is designed for people interested in learning about the technologies used to transcribe, process, search, and retreive spoken audio materials. Detailed prior knowledge of speech recognition and/or search technologies is not required.

Ciprian Chelba is a Research Scientist with Google. Previously he worked as a Researcher in the Speech Technology Group at Microsoft Research. His core research interests are in statistical modeling of natural language and speech. Recent projects include speech content indexing for search in spoken documents, discriminative language modeling for large vocabulary speech recognition, as well as speech and text classification.

Timothy J. Hazen is a Research Scientist at the MIT Computer Science and Artificial Intelligence Laboratory where he works in the areas of automatic speech recognition, automatic person identification, multi-modal speech processing, and conversational speech systems. For the last two years he has been a key contributor to the MIT Spoken Lecture Processing Project.