Let's Hear It for Audio Mining

Neal Leavitt

The Web, databases, and other digitized information storehouses contain a growing volume of audio content. Sources include newscasts, sporting events, telephone conversations, recordings of meetings, Webcasts, documentary archives such as the Visual History Foundation's interviews with Holocaust survivors (http://www.vhf.org), and media files in libraries.

Users want to make the most of this material by searching and indexing the digitized audio content. In the past, companies had to create and manually analyze written transcripts of audio content because using computers to recognize, interpret, and analyze digitized speech was difficult. However, the development of faster microprocessors, larger storage capacities, and better speech-recognition algorithms has made audio mining easier.

Now, the technology is on the verge of becoming a powerful tool that could help many organizations. For example, companies could use audio mining to analyze customer-service and help-desk conversations or even voice mail. Law enforcement and intelligence organizations could use the technology to analyze intercepted phone conversations. Public relations firms could use it to analyze news broadcasts to find coverage of clients.

Broadcast companies like CNN and Radio Free Asia are already using audio mining to quickly retrieve important background information from previous broadcasts when new stories break. And a US prison is using ScanSoft's (http://www.scansoft.com)

audio mining product to analyze recordings of prisoners' phone calls to identify illegal activity.

Several companies such as BBN Technologies (http://www.bbn.com), Fast-Talk Communications (http://www.fast-talk.com), IBM, and ScanSoft have released audio mining software, and industry observers expect the number of products to increase during the next few years.

However, audio mining's accuracy levels are still relatively low, and some products are expensive, with some high-end software packages costing more than $100,000 for a full-scale deployment.

INSIDE AUDIO MINING

Audio mining, also called audio searching, takes a text-based query and locates the search term or phrase in an audio file. This helps users by, for example, letting them quickly get to specific places in a recorded conversation or determine when a company is mentioned in a newscast.

Audio indexing uses speech recognition to analyze an entire file and produce a searchable index of content-bearing words and their locations. This is critical because audio content is in a binary format that is otherwise not readily searchable, explained Robert Weideman, ScanSoft's chief marketing officer.

Indexing audio content thus enables searching, said Jeff Karnes, a group product manager for Virage, an audio mining vendor.

Audio mining approaches

There are two main approaches to audio mining.

Text-based indexing. Text-based indexing, also known as large-vocabulary continuous speech recognition, converts speech to text and then identifies words in a dictionary that can contain several hundred thousand entries. If a word or name is not in the dictionary, the LVCSR system will choose the most similar word it can find.

The system uses language understanding to create a confidence level for its findings. For findings with less than a 100 percent confidence level, the system offers other possible word matches, said Professor Dan Ellis, who leads Columbia University's Laboratory for Recognition and Organization of Speech and Audio (http://labrosa.ee.columbia.edu).

Phoneme-based indexing. Phoneme-based indexing doesn't convert speech to text but instead works only with sounds.

The system first analyzes and identifies sounds in a piece of audio content to create a phonetic-based index. It then uses a dictionary of several dozen phonemes to convert a user's search term to the correct phoneme string. (Phonemes are the smallest unit of speech in a language, such as the long "a" sound, that distinguishes one utterance from another. All words are sets of phonemes.) Finally, the system looks for the search terms in the index.

"A phonetic system requires a more proprietary search tool because it must phoneticize the query term, then try to match it with the existing phonetic-string output," Weideman said. This is considerably more complex than using

one of the many existing text-based search tools.

Phoneme-based searches can result in more false matches than the text-based approach, particularly for short search terms, because many words sound alike or sound like parts of other words.

Thus, Ellis said, it's difficult for a phonetic system to accurately classify a phoneme except by recognizing the entire word that it is part of or by understanding that a language permits only certain phoneme sequences.

However, he added, phonetic indexing can still be useful if the analyzed material contains important words that are likely to be missing from a text system's dictionary, such as foreign terms and names of people and places.

How the technology works

Text- and phoneme-based systems operate in much the same way, except that the former uses a text-based dictionary and the latter uses a phonetic dictionary.

The most important and complex component technology for audio mining is speech recognition. In these systems, explained University of Texas Assistant Professor Latifur R. Khan, "A speech recognizer converts the observed acoustic signal into the corresponding [written] representation of the spoken [words]."

Speech recognition software contains acoustic models of the way in which all phonemes are represented. Also, there is a statistical language model that indicates how likely words are to follow each other in a specific language, said William Meisel, president of TMA Associates, a speech-industry market-research firm. By using these capabilities, as well as complex probability analysis, the technology can take a speech signal of unknown content and convert it to a series of words from the program's dictionary.

Khan noted that this process is more difficult with highly inflected languages, such as Chinese, in which tonality changes the meaning of a word.

Some audio mining dictionaries are domain specific, for use by professionals in different fields, such as law or medicine.

Some products, such as ScanSoft's AudioMining Development System, shown in Figure 1, use XML's ability to tag data so that other XML-capable systems can read it, ScanSoft's Weideman noted. This lets the product export speech index information to other systems, he said.

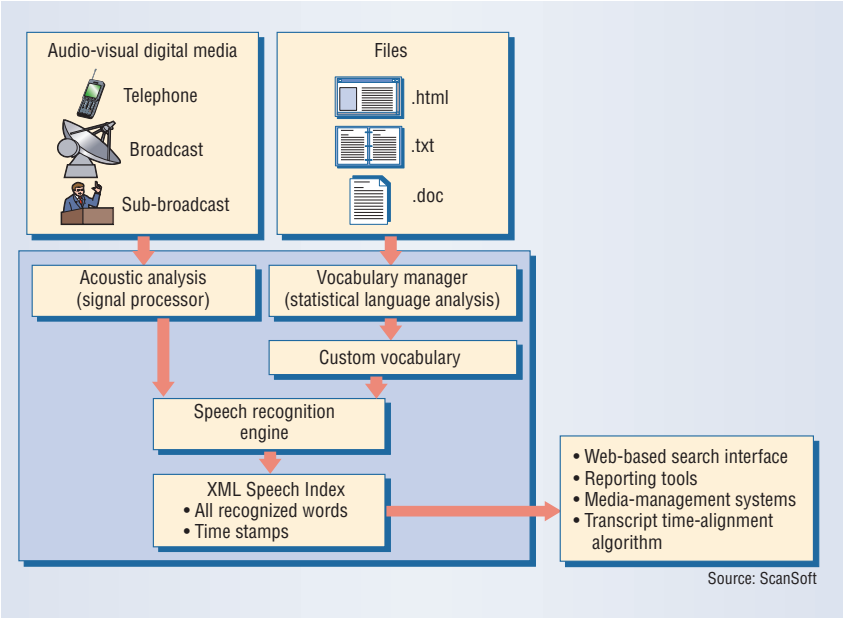

Figure 1. The ScanSoft AudioMining Development System works with audio from various sources. The system analyzes sounds to generate sound strings that are then identified as words by the speech recognition engine, which works with its own dictionary. Material input into the vocabulary manager automatically updates the dictionary. The product's XML Speech Index uses XML's cross-platform capabilities to create files that work with various search engines, servers, and content management systems.

Performance

By working with powerful host-system processors, large memories, and efficient algorithms, most audio mining technology provides high performance levels.

For example, Fast-Talk says its newest technology can index a one-hour audio file in five minutes, and it can process 30 hours of content per second in response to a specific, 10-phoneme search query in a host system running a 2.53-GHz Pentium CPU.

Multiple languages

Demand for multilingual audio mining systems is slowly growing, particularly for those that work with Arabic, Mandarin Chinese, US and UK English, German, Japanese, and Spanish.

BBN Technologies' BBN Audio Indexer plug-in produces indexed, searchable transcriptions of any audio source in Arabic, Chinese, English, or Spanish in real time on a standard PC.

Porting a product to a new language or a significantly different dialect is time consuming and expensive because developers must collect and transcribe acoustic data for the language or dialect and then train and evaluate new acoustic models.

Designers overcome challenges

A major challenge for speech recognition tools has been recognizing the speech of different users in different environments. With this in mind, BBN, IBM, Fast-Talk, and ScanSoft have

designed their audio mining technology to be largely speaker independent. For example, Fast-Talk's acoustic models are trained to recognize numerous speakers via exposure to audio data from males and females representing various ages, dialects, and speaking styles, explained Mark A. Clements, the company's cofounder and a professor at the Georgia Institute of Technology's Center for Advanced Communications.

Some audio mining technology uses acoustic models tuned to understand speech from different environments—such as telephony, TV, or radio.

Meanwhile, Clements said, researchers have had to make other improvements, such as designing filters to reduce the background noise that can interfere with accurate speech recognition, creating efficient data structures for representing content, and developing algorithms that quickly work through the data structures during indexing and searching.

HURDLES TO CLEAR

Although audio mining developers have overcome numerous challenges, several important hurdles remain.

For example, precision is improving but it is still a key issue impeding the technology's widespread adoption, particularly in such accuracy-critical applications as court reporting and medical dictation.

According to Virage's Karnes, audio mining error rates vary widely depending on factors such as background noise and cross talk. He said that Virage's internal testing indicates a 5 to 20 percent error rate for processing news broadcasts and a 30 to 60 percent error rate for processing other content types.

Processing conversational speech can be particularly difficult because of such factors as overlapping words and background noise, noted Professor Howard D. Wactlar, director of Carnegie Mellon University's Informedia Digital Video Library Project.

Breakthroughs in natural language understanding will eventually lead to big improvements, but until then, Clements said, audio mining will get better only incrementally.

Meanwhile, demand for audio mining products is not overwhelming but is slowly growing. "It goes back to the accuracy of the systems," Wactlar said. "Most users have high expectations, and unless these are met, they'll be frustrated with the products and won't use them."

Another factor inhibiting rapid growth and widespread adoption is cost. Prices of audio mining systems can exceed $100,000 because the relatively new systems are still expensive to develop and market, particularly with today's low demand. Volume sales, Karnes said, will eventually bring prices down, but this probably won't be a factor for at least another 12 to 18 months.

Also, Wactlar said, the market doesn't have any killer apps that would rapidly spur organizations to buy today's products.

Meisel speculated that an important market opportunity for the technology lies in call centers that must monitor conversations for employee compliance with company policy.

In addition, said Clements, "As users increasingly use Web conferencing for important meetings and education, they create a large corpus of valuable information that could be [analyzed]. Searchable libraries of Web meetings are thus a likely killer application."

Like many nascent technologies, audio mining shows great promise but won't realize its full potential until accuracy and affordability improve. The technology will most likely find interest in niche markets—such as technical-support centers, help desks, and call centers—during the next three to five years because most companies today don't work extensively with multimedia data, said Jackie Fenn, a vice president and research fellow for market research firm Gartner Inc.

"The technology is just beginning to emerge as a tool that is sufficiently robust," Columbia University's Ellis stated. "I see a period of big changes as these techniques open up a wide range of unanticipated applications."

ScanSoft's Weideman concluded, "Audio mining is an extremely exciting technology that could add tremendous value to knowledge-sharing, intelligence, and productivity applications. But it's not ready for mass adoption yet. It's currently a 'nice to have,' not quite yet a 'need to have.' " ■

Audio Mining: A History in Brief

Serious audio mining research began in the late 1970s. Research has been ongoing at a number of major schools, including Carnegie Mellon University, Columbia University, the Georgia Institute of Technology, and the University of Texas.

However, products have been available for only about four years. And it is only in the past 12 to 18 months that the technology has begun to offer acceptable performance and accuracy levels for commercial use, said Jeff Karnes, a group product manager for Virage, an audio mining vendor.

Audio mining products generally have been integrated into larger systems because, Karnes explained, "the ability to search and replay content is most valuable as part of a [product] that manages large archives of information, such as a media- or content-management system."

Neal Leavitt *is president of Leavitt Communications, an international marketing communications firm with affiliate offices in Paris; Hamburg, Germany; Beijing; and Sao Paulo, Brazil. He frequently writes about technology. Contact him at neal@leavcom.com.*

Editor: Lee Garber, *Computer*, 10662 Los Vaqueros Circle, PO Box 3014, Los Alamitos, CA 90720-1314; l.garber@computer.org