Audio and Speech Processing for Data Mining

Zheng-Hua Tan
Aalborg University, Denmark

INTRODUCTION

The explosive increase in computing power, network bandwidth and storage capacity has largely facilitated the production, transmission and storage of multimedia data. Compared to alpha-numeric database, non-text media such as audio, image and video are different in that they are unstructured by nature, and although containing rich information, they are not quite as expressive from the viewpoint of a contemporary computer. As a consequence, an overwhelming amount of data is created and then left unstructured and inaccessible, boosting the desire for efficient content management of these data. This has become a driving force of multimedia research and development, and has lead to a new field termed multimedia data mining. While text mining is relatively mature, mining information from non-text media is still in its infancy, but holds much promise for the future.

In general, data mining the process of applying analytical approaches to large data sets to discover implicit, previously unknown, and potentially useful information. This process often involves three steps: data preprocessing, data mining and postprocessing (Tan, Steinbach, & Kumar, 2005). The first step is to transform the raw data into a more suitable format for subsequent data mining. The second step conducts the actual mining while the last one is implemented to validate and interpret the mining results.

Data preprocessing is a broad area and is the part in data mining where essential techniques are highly dependent on data types. Different from textual data, which is typically based on a written language, image, video and some audio are inherently non-linguistic. Speech as a spoken language lies in between and often provides valuable information about the subjects, topics and concepts of multimedia content (Lee & Chen, 2005). The language nature of speech makes information extraction from speech less complicated yet more precise and accurate than from image and video. This fact motivates content based speech analysis for multimedia data mining and retrieval where audio and speech processing is a key, enabling technology (Ohtsuki, Bessho, Matsuo, Matsunaga, & Kayashi, 2006). Progress in this area can impact numerous business and government applications (Gilbert, Moore, & Zweig, 2005). Examples are discovering patterns and generating alarms for intelligence organizations as well as for call centers, analyzing customer preferences, and searching through vast audio warehouses.

BACKGROUND

With the enormous, ever-increasing amount of audio data (including speech), the challenge now and in the future becomes the exploration of new methods for accessing and mining these data. Due to the non-structured nature of audio, audio files must be annotated with structured metadata to facilitate the practice of data mining. Although manually labeled metadata to some extent assist in such activities as categorizing audio files, they are insufficient on their own when it comes to more sophisticated applications like data mining. Manual transcription is also expensive and in many cases outright impossible. Consequently, automatic metadata generation relying on advanced processing technologies is required so that more thorough annotation and transcription can be provided. Technologies for this purpose include audio diarization and automatic speech recognition. Audio diarization aims at annotating audio data through segmentation, classification and clustering while speech recognition is deployed to transcribe speech. In addition to these is event detection, such as, for example, applause detection in sports recordings. After audio is transformed into various symbolic streams, data mining techniques can be applied to the streams to find patterns and associations, and information retrieval techniques can be applied for the purposes of indexing, search and retrieval. The procedure is analogous to video data mining and retrieval (Zhu, Wu, Elmagarmid, Feng, & Wu, 2005; Oh, Lee, & Hwang, 2005).

Diarization is the necessary, first stage in recognizing speech mingled with other audios and is an important field in its own right. The state-of-the-art system has achieved a speaker diarization error of less than 7% for broadcast news shows (Tranter & Reynolds, 2006).

A recent, notable research project on speech transcription is the Effective Affordable Reusable Speech-To-Text (EARS) program (Chen, Kingsbury, Mangu, Povey, Saon, Soltau, & Zweig, 2006). The EARS program focuses on automatically transcribing natural, unconstrained human-human speech from broadcasts and telephone conversations in multiple languages. The primary goal is to generate rich and accurate transcription both to enable computers to better detect, extract, summarize, and translate important information embedded in the speech and to enable humans to understand the speech content by reading transcripts instead of listening to audio signals. To date, accuracies for broadcast news and conversational telephone speech are approximately 90% and 85%, respectively. For reading or dictated speech, recognition accuracy is much higher, and depending on several configurations, it can reach as high as 99% for large vocabulary tasks.

Progress in audio classification and categorization is also appealing. In a task of classifying 198 sounds into 16 classes, (Lin, Chen, Truong, & Chang, 2005) achieved an accuracy of 97% and the performance was 100% when considering Top 2 matches. The 16 sound classes are alto-trombone, animals, bells, cello-bowed, crowds, female, laughter, machines, male, oboe, percussion, telephone, tubular-bells, violin-bowed, violin-pizz and water.

The technologies at this level are highly attractive for many speech data mining applications. The question we ask here is what is speech data mining? The fact is that we have areas close to or even overlapping with it, such as spoken document retrieval for search and retrieval (Hansen, Huang, Zhou, Seadle, Deller, Gurijala, Kurimo, & Angkititrakul, 2005). At this early stage of research, the community does not show a clear intention to segregate them, though. The same has happened with text data mining (Hearst, 1999). In this chapter we define speech data mining as the nontrivial extraction of hidden and useful information from masses of speech data. The same applies to audio data mining. Interesting information includes trends, anomalies and associations with the purpose being primarily for decision making. An example is mining spoken dialog to generate alerts.

MAIN FOCUS

In this section we discuss some key topics within or related to speech data mining. We cover audio diarization, robust speech recognition, speech data mining and spoken document retrieval. Spoken document retrieval is accounted for since the subject is so closely related to speech data mining, and the two draw on each other by sharing many common preprocessing techniques.

Audio Diarization

Audio diarization aims to automatically segment an audio recording into homogeneous regions. Diarization first segments and categorizes audio as speech and non-speech. Non-speech is a general category covering silence, music, background noise, channel conditions and so on. Speech segments are further annotated through speaker diarization which is the current focus in audio diarization. Speaker diarization, also known as "Who Spoke When" or speaker segmentation and clustering, partitions speech stream into uniform segments according to speaker identity.

A typical diarization system comprises such components as speech activity detection, change detection, gender and bandwidth identification, speaker segmentation, speaker clustering, and iterative re-segmentation or boundary refinement (Tranter & Reynolds, 2006). Two notable techniques applied in this area are Gaussian mixture model (GMM) and Bayesian information criterion (BIC), both of which are deployed through the process of diarization. The performance of speaker diarization is often measured by diarization error rate which is the sum of speaker error, missed speaker and false alarm speaker rates.

Diarization is an important step for further processing such as audio classification (Lu, Zhang, & Li, 2003), audio clustering (Sundaram & Narayanan, 2007), and speech recognition.

Robust Speech Recognition

Speech recognition is the process of converting a speech signal to a word sequence. Modern speech recognition systems are firmly based on the principles of statistical pattern recognition, in particular the use of hidden Markov models (HMMs). The objective is to find the most likely sequence of words $\hat{W}$, given the observation data Y which are feature vectors extracted