Speech Audio Retrieval Using Voice Query

Chotirat Ann Ratanamahatana and Phubes Tohlong

Dept. of Computer Engineering, Chulalongkorn University, Bangkok 10330 Thailand
ann@cp.eng.chula.ac.th, Tohlong@hotmail.com

Abstract. Multimedia data has increasingly become a prevalent resource in Digital Library system; this includes audio, video, and image archives. However, each type of these data may need specific tools to help facilitate effective and efficient retrieval tasks. In this paper, we focus on retrieval of speech audio collection, which includes audio books, speech recordings, interviews, and lectures. Currently, most of the audio retrieval systems are based on keyword/title/author search typed into the system by users. The system then searches for particular keywords and gives a list of entire audio files that potentially are relevant to the query. Nonetheless, browsing audio content for particular section of the audios without knowing the actual content is yet a very difficult task. Moreover, since audio transcription or keyword annotation is very labor intensive and becomes infeasible for large data, we introduce here a preliminary framework that locates subsections of the audio that correspond to the voice query made by a user. We demonstrate a utility of our approach on query retrieval tasks in various types of audio recordings. We also show that this simple framework can potentially help retrieve and locate the voice query within the audio accurately and efficiently.

Keywords: Audio retrieval, time series, query by example, voice search.

1 Introduction

Speech processing has established itself in research communities since 1950s. However, it still has many unsolved problems and remains a very challenging and active area of research nowadays due to its exceptionally complex nature of the problem itself. Many speech processing techniques have been proposed for speech audio retrieval [1][6][7]. In spite of this, none of them have really solved a problem of searching the actual content within large retrieved audio files. Instead, the search processes usually are text-based or voice query, searching for entire audio files according to provided titles, authors, and keywords [1][3]. Some system may need manual transcription of the speech/audio into text before searching can be performed. It would be very helpful and more convenient if we can search *any* part of the speech audio using our voice as a query without having to do the transcription. This paper proposes a preliminary alternative to textual annotations, which is based on time series features extracted from the raw speech data.

S. Sugimoto et al. (Eds.): ICADL 2006, LNCS 4312, pp. 494 – 497, 2006.

1.1 Motivation

Our motivation started from an attempt to search the recorded lectures archived in the digital library collection. At this point, we can generally search for audio files as a whole, based on keywords, titles, and authors. However, if the retrieved audio file is very long, it is still extremely hard to browse or locate the exact content within the audio, when we are interested only in some parts of the recording; it is more likely that a user would like to only hear subsections within the audio file instead of having to listen to the whole audio from the beginning.

This work is based on a query-by-example (QBE) technique, where users provide voice/speech examples of the word or phrase they seek. Some may argue that this query-by-example approach has a major limitation when the users really want to search for semantic concepts rather than the "exact" word or phrase; rather, query by keyword approaches may be more appropriate. Since these types of research have been a research of interest within speech communities for years and still have not been considered a completely solved problem, we are taking this opportunity to explore an alternative in approaching the problem without using the full speech processing techniques. We would like to be able to search *inside* each audio file to locate the exact content that we want based on a given voice query.

2 Time Series Representation for Voice Searching

More complex analysis of the speech audio cannot be achieved by looking at the raw audio plots alone. To get some information about the frequency distribution, harmonics, and others, some signal processing such as Fourier analysis is needed. In this work, we propose a simple approach to approximately represent audio features using time series representation, an approach recently used in query by humming system [4][5]. Note that by looking at the raw audio plot (.WAV file), we can extract several features, such as volume from the amplitude and the "timbre" of the voice. However, these characteristics are irrelevant to the task of differentiating one word from another. Instead, we propose to simply use the frequencies information as approximations of words in the audio. Note that the effectiveness and accuracy of this approach essentially depend on the nature of the spoken languages themselves as well. In this work, we test our method in Thai language, a tonal language with 5 tones.

We start off with acquiring the Thai digital audio recording in WAV file format. In our experiment, all recordings are originally recorded at the sampling rate of 22,050 Hz, but we decide to downsample the data to only 2,000 Hz (16 bits mono) to significantly speed up the search process and make sure that we do not lose too much of important features during the reduction and calculation. In speech community, such sampling rate is considered unacceptably low; however, our proposed work has one big difference in the algorithm in that we process the speech in *word level*, instead of *phoneme level* as typically being done in speech processing. This in turn allows us to easily process a 1-hour audio which almost seems unfeasible if we were to employ a traditional automatic speech recognition process.

We then preprocess the data by transforming a raw audio into a frequency domain using Fast Fourier Transformation (FFT), which gives the frequency distribution information about the spoken word or subsequence of the recording. This is a time series to be later used in similarity search in our framework. In addition, to further

remove noise and outlier, we also apply some smoothing and z-score normalization to all datasets in our work before utilizing a Dynamic Time Warping distance measure (DTW) to locate the *K*-nearest neighbor query word within the given recording.

The algorithm is simply a subsequence matching using a sliding window of the size of the query window. Starting from the beginning of the recording until the end, it looks for the one with best match using a similarity measure. To simplify the implementation, a Euclidean distance metric could be used. However, we believe that a more sophisticated similarity measures, such as Dynamic Time warping [2][8], could significantly improve the accuracy of the result since it could gracefully resolve the problem of discrepancies or minor time variation in the time series, where we could intuitively map the time series query to the appropriate section of the recording.

3 Experimental Evaluation

We have put together a collection of various audio recordings for our experiment; some are audio books, and some are real lectures with both male and female speakers. Each one is approximately 45 to 60 minutes in length, with word content ranging from 6,000 to 9,000 words. We have chosen some words from each recording and exclusively removed those occurrences from the recording to avoid getting an exact match during the search. To evaluate the retrieval's effectiveness, we calculate the Precision, Recall, as well as the F-Measure to compare results among various parameter settings and approaches.

3.1 Experiment Results and Discussions

At this preliminary stage of our work, the evaluation process must be done manually. After the query words are selected, we have to actually listen to the whole recording and mark all the actual occurrences of each word within the recording, since there is no transcription available. The main contribution of our work is an ability to perform a voice search within a large audio file, where speech processing community may still have difficulties with. We demonstrate our utility by querying a word in an hour-long audio then measure the retrieval effectiveness both by looking at the precision/recall as well as the running time. Up to this point, we have demonstrated that Dynamic Time warping distance measure always outperforms the classic Euclidean distance metric in terms of the accuracy but with the price of higher time complexity.

In addition, we also consider another approach using Mel Frequency Cepstral Coefficients or MFCC that is regularly employed in speech processing to see if its superiority still holds for voice search in the word level. We first compare its time complexity with the Euclidean and Dynamic Time Warping distance measures. With exactly the same parameters and settings, MFCC measure is running 30 times slower

Table 1. Comparison of results between FFT with DTW and MFCC measures, showing that DTW gives more accurate results

Approach	Precision	Recall	F-Measure
FFT with DTW	80%	75%	77.42%
MFCC	61.11%	68.75%	64.71%

than Euclidean distance and about 5 times slower than the Dynamic Time warping. The retrieval's effectiveness between the two approaches is shown in Table 1.

Since the MFCC's running time is larger and its F-Measure is much lower, FFT with DTW distance measure is then employed in our experiments. With speaker-dependent experiment, as expected, we get much worse results; there are many more query words that were left undetected, as well as a lot more false alarms. Ideally, we would like to minimize the number of False Negatives as much as possible, with an acceptable number of False Positives. Looking closely, we found that the results are affected across genders as well. We look at the Fourier analysis of the same word spoken by different speakers and discover that they approximately have the similar shape but relatively shifted along the frequency axis. That means the structure of the word spoken are quite similar across the speakers, but the overall speaking frequency for each person differs and can be thought of as a frequency offset.

4 Conclusions and Future Work

In this preliminary work, we have proposed a simple approach to approximately represent speech audio features using time series representation, then to locate a voice query within the audio recordings. We have demonstrated the utility of our approach on query retrieval tasks for audio recordings in Thai language, i.e., to locate a voice query within the lecture recordings. From the experiment results, we have demonstrated that this simple framework can potentially help retrieve and locate the audio according to voice query inputs, especially in the speaker-dependent situation. Since the pitch discrepancies among speakers pose a limitation in our current framework, we need to look more closely into these features and see if any normalization among various speakers could be attained. Together with a Dynamic Time Warping distance measure as well as some lowerbounding and dimensionality reduction techniques, this could potentially resolve the problem and to help speed up the overall search process.

Referenes

[1] Franz, A. & Milch, B. (2002). Searching the Web by Voice. In Proceedings of COLING.
[2] Kruskall, J. B. & Liberman, M. (1983). The symmetric time warping algorithm: From continuous to discrete. In Time Warps, String Edits and Macromolecules.
[3] Klabbhankao, B. (2000). Online Information Retrieval Using Genetic Algorithms. NECTEC Technical Journal Vol 2, No.7. March-June.
[4] Zhu, Y., Shasha, D., & Zhao, X. (2003). Query by Humming – in Action with its Technology Revealed. ACM SIGMOD, June 9-12.
[5] Zhu, Y. & Shasha, D. (2003). Warping Indexes with Envelope Transforms for Query by Humming. ACM SIGMOD, June 9-12.
[6] Hazen, T.J., Saenko, K., La, C.-H., & Glass, J.R. (2004). A Segment-Based Audio-Visual Speech Recognizer: Data Collection, Development, and Initial Experiments. Proc. ICMI.
[7] Gutkin, A. & King, S. (2004). Structural Representation of Speech for Phonetic Classification. In Proc. 17th International Conference on Pattern Recognition (ICPR), volume 3, pages 438-441, Cambridge, UK, August 2004. IEEE Computer Society Press.
[8] Ratanamahatana, C.A. & Keogh, E. (2005). Three Myths about Dynamic Time Warping Data Mining. SIAM International Conference on Data Mining (SDM).