TEST Engineering & Management

January - February 2020
ISSN: 0193-4120 Page No. 8469 – 8472

Speech Retrieval using STD

M. Mamatha, *Research scholor, RU, AP, India.*
T.Bhaskar Reddy, *Professor, SKU, AP, India.*

Article Info
Volume 82
Page Number: 8469 – 8472
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 07 February 2020*

Abstract:
This paper given a brief survey on Speech retrieval techniques. In each and every technique discuss the advantages and disadvantages. The methods which are discussed in this paper are designed for low-useful resource occasions where confined or no in-domain coaching fabric is available and accurate word-centered speech consciousness potential is unavailable. Using these techniques, we demonstrate that it is indeed feasible to perform spoken term detection in low resource scenarios. Moreover, these show that efficient and practical STD systems could be developed which search through several hours of spoken data in real time. Spoken Term Detection (STD) is the task of detecting all occurrences of a query from spoken data. In this thesis, two types of STD techniques are explored. In the first part, a query-by-example STD approach is presented for language independent search. In the second part, a text-based STD technique.

Keywords:*Speech Retrievsl, in-domain, STD systems.*

I. INTRODUCTION

In the digital era, huge amounts of audio data are being produced and consumed every day in a variety of languages. In recent days digital technology is widely using technology. For example broad cast news, classroom lectures, audio books and call centers. Audio mining refers to get the useful information from large amount of database. The increasingly widespread use of digital technology in our everyday lives has made speech data accessible to the masses, thus making it easy to record, preserve and reproduce digital content. The ability to search through this data thus becomes a valuable functionality to improve its access. But, unfortunately the linear and non-deterministic nature of speech signal limit our ability to retrieve information from this repository efficiently. Thus, providing fast and intelligent access to these large speech collections have become a necessity to unlock their true potential as knowledge resources.

There are many languages for which no transcription available on spoken data. For such languages it is not possible to build supervised acoustic or language models. Recently a new technology or algorithms are available to search query on low resource language to perform language independent search. A Large Vocabulary Continuous Speech Recognition (LVCSR) engine converts speech into its corresponding text. In general, it transcribes spoken data into words using apriori-trained acoustic models and language models.

Representing speech as text has many advantages while performing the task of information retrieval. Many well-established techniques from the domain of text data mining could be incorporated into the speech domain. If the LVCSR engine could provide good transcription of spoken data, then the task of spoken information retrieval gets simplified to the problem of searching a text query in a text database. Traditionally, most of the efforts were directed to designing a good speech recognition system. Thus, the problem of audio mining was reduced to the problem of speech recognition

Building a good speech recognizer continues to be a challenging task. First of all, large amount of training data is required for building a good LVCSR engine. Most of the languages of the world do not have such rich training resources. For example, out of about 7000 languages across the world, commercial speech recognition engines are available only for about 80 of them [4, 3]. Secondly, even in languages where sufficient training resources are available, accurate recognition continues to prove difficult, especially in the context of conversational speech, noisy environment etc. For example, word error rates (WER) in English conversational

January - February 2020
ISSN: 0193-4120 Page No. 8469 – 8472

telephone speech remain as high as 30% to 40% in the state-of-the-art-systems. Another issue is the problem of out-of-vocabulary (OOV) words. Since only a finite vocabulary set could be used while training the recognizer, many words would be absent from the dictionary, such as person names, place names and other proper nouns, abbreviations etc. This too brings down the accuracy of the recognizer.

The aim of the thesis is to improve spoken term detection for low resource languages. In particular, the aim is to develop a system with the following characteristics:

• Low resource usage: The system must require only very limited resources of a language.

• Open-vocabulary search: The system must be able to search for any term, without having it to be pre defined in a dictionary, i.e., it must be able to handle OOV queries.

• Accurate spoken term detection: The STD system must be able to retrieve relevant segments from speech collections accurately.

• Fast search capability: The system must be scalable to search through several hours of spoken data in a matter of a few seconds.

II. TECHNIQUES

The goal is to locate the portions from the audio data containing this textual query. The most straight forward approach is to convert the audio database into text using an LVCSR, and then perform a text-based string matching between the query and the database. But, building an LVCSR is not feasible for low resource contexts. We discuss various techniques for Speech retrieval

1. Phonetic based Search: In this search, preprocess the audio into the phonemes and encodes the effect in a lattice of potentialities. Any search term (query) also translated into phonemes and the search appears for the equal sequence in the current lattice.

Pros and Cons of Phonetic process:

advantage is that phrases that are not in a predefined vocabulary can nonetheless be observed ,supplied the phonemes are recognizable.

Drawback is that given that there are various feasible sequences in the lattice, the term may be discovered in lots of locations. For those false positives have to be manually filtered out of the result set. This system is faster at processing i.e., changing into phonemes and far slower in search method, considering phoneme can't be as efficiently listed the way entire phrases can. Whilst phonetic strategies do don't forget what the possible sequences are and their frequency.

2. Large vocabulary Continuous Speech Recognition(LVCSR) speech analysis: It also start with by identifying the phonemes ,then applies a language or dictionary model of 60000-10000 words and phrases to produce a full transcript. In LVCSR every word is identified and nothing is thrown away or omitted. While the initial process of recognizing the full transcript requires more preprocessing power and the resulting transcript makes it much easier and faster for search.

Pros and Cons of LVCSR system:

Capabilities is LVCSR uses statistical approaches to affirm the probability of exceptional phrases, hence the accuracy is far greater than simply the only phrase look up of a phonetic method, the word is discovered it was quite spoken.

Drawback is the words within the search terms want to be in dictionary or vocabulary, the word-phrase of curiosity ordinarily may also be found with the aid of combining words.

Example: “See Alice” for “cialis”

The preliminary processing of the audio takes bit longer on the grounds that of the big vocabulary that has to be analyzed . Search time is far rapid and more accurate. It has greater precision given that it include phrases which might be simply spoken. Curb bear in mind due to individual phrases or consciousness blunders.

Phonemes signify the basic sound unit in a language. It is one in every of many viable sounds within the language. Stated in a defined method. Represented between brackets by using conference. Instance [b], [j] ,[o]..and many others.

Syllables represent the sound units composed of a central nucleus , which is mostly a vowel, with optional initial and final consonants.

3. Spoken Document Retrieval (SDR): SDR is the retrieval of relevant document for the given query. To retrieve the data related to given query. This may be related to may be semantically or acoustic. For example in SDR for given query it retrieves the speech segments. For example to search news data base for given query, the news segments should be apriority, then only news indentified. For this the word error rate(WER) is 20-25%,basing on this the performance is measured. But this work limit only to

January - February 2020
ISSN: 0193-4120 Page No. 8469 – 8472

the easy speech that is the keywords are repeated multiples.

4. Spoken Term Detection(STD): In STD method, it detects all occurrences of a query from spoken data. Compare to SDR the STD give better results, the relevance of document is very clear. The STD system developed such that it works on low resources also that is lack of sufficient linguistic resources. In order to develop the STD results in low resource language, the system has the following characteristics-

1. Low resource language-Limited resources.

2. Open vocabulary search-It has to search any term, not predefined in dictionary that is able to handle OOV queries.

3. Accurate-STD retrieve the most relevant documents.

4. Fast search capability-Able to search several hours of spoken data with in a few seconds.

To achieve this, use sub-word based approaches. The word based method not give the best results for OOV or open-vocabulary search. In this search the main importance is given to reduce the search space, so that search time will be reduced. To improve the accuracy in results, multi-stage search techniques are used.

The STD (Spoken Term Detection) to detect all occurrences of required speech data from Speech data base. The relevance of a document is solely determined by the presence or absence of the query. In addition to retrieving relevant documents, the time stamps within a document where the query occurs may also have to be returned to the user. Sometimes, the term spoken utterance retrieval is used to refer to the task of STD By using multiple Automatic Speech Recognition(ASR) to index the data, the search time is very high. This is not applicable to real time use. To search efficiently for real time, use better indexing representation need. For this one method is STD. STD retrieve data either text or spoken format. In this paper discuss 2-types of STD techniques.

1.Query-by-example STD-This method used for language independent search.

2.Text based STD-In this method if no language model is available ,then use the sub representation.

The "spoken web search" task search an audio query in audio content. It required to build language-independent STD systems. In other cases, if resources are scarce then for that case, trained for basic units(phones) of acoustic models are used. If appropriate language model not exist then use the ASR, which is might be trained to produce a sub word transcription of speech. By applying this method to online audio indexing system, for 64000 words it showed 12% of search terms are OOV. So for 12% of queries it never give correct result. To avoid this use the sub word technique like phone recognition, this method itself is error-phone task

4.1 Query by Example STD:

Query by Example is a database query language for relational databases. It is the first graphical query language using visual tables where the user would enter commands, example elements and conditions. Many graphical front-ends for databases use the ideas from QbE. In information retrieval, QbE has a some what different meaning. The user can submit a document or several documents and ask for "similar" documents to be retrieved from a document database. Similarity search is based on comparing document vectors.

4.2 Text-based STD:

In text-based STD, the query (a single word or a sequence of words) is represented in a text form. The goal is to locate the portions from the audio data containing this text query. Most of the text-based STD tasks assume that enough resources and knowledge of the target language are available, so that transcribed data, phone sets, pronunciation dictionaries, language models etc. could be employed. This becomes a challenge in low resource scenarios with high OOV rate. If the word is OOV, then letter-to-sound rules are used for generating phone transcriptions. A generic architecture for a text-based STD system is given in figure 2.2.

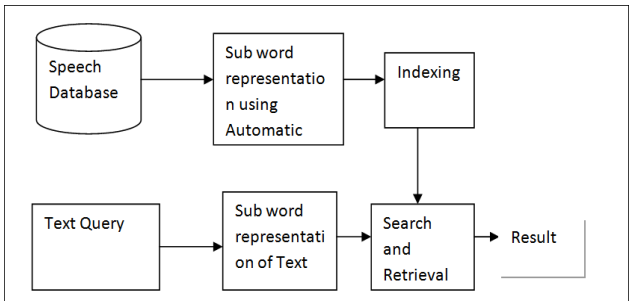

Figure2.2 A generic architecture for a text-based STD system.

January - February 2020
ISSN: 0193-4120 Page No. 8469 – 8472

III. CONCLUTION AND SCOPE OF FUTURE WORK

In this paper, a brief overview of the various aspects of Speech Recognition techniques was given. The limitations of LVCSR systems for transcribing spoken documents in low resource scenarios were mentioned. The need of sub word representations of speech was described. Two broad categories of STD based on the type of query were given. For the applications where the query is given in spoken form, the QbE STD systems are preferred, where as text-based STD can be employed when text queries are given.

More research needs to be done to obtain better representation techniques. Better signal matching algorithms need to be developed. Better variants of Dynamic Time Warping could be explored.

• Improving phone error rates by using improved phone recognition engines could increase text based STD system accuracy.

• Better indexing and search techniques that can include alternate representations of signal could be explored.

REFERENCES

1. G. Aradilla, J. Vepa and H. Bourlard. “Using posterior-based features in template matching for speech recognition,” in Proc. Interspeech, Pittsburgh, Sep. 2006.
2. G. Aradilla, H. Bourlard, and M. Magimai-Doss. “Posterior features applied to speech recognition tasks with user-defined vocabulary,” in Proc. ICASSP, Taipei, May 2009.
3. C. Chelba, T. Hazen and M. Sarac¸lar, “Retrieval and browsing of spoken content,” IEEE Signal Processing Magazine, vol. 24, no. 3, pp. 39–49, May 2008.
4. C. Cieri, D. Miller, and K. Walker, “From Switchboard to Fisher: Telephone collection protocols, their uses and yields,” in Proc. Interspeech, Geneva, Sep. 2003.
5. J. Fiscus, J. Ajot, J. Garofolo, and G. Doddington, “Results of the 2006 Spoken Term Detection Evaluation,” in Proc. 2007 SIGIR Workshop on Searching Spontaneous Conversational Speech, Amsterdam, July 2007.
6. D. Miller, et al, “Rapid and accurate spoken term detection,” in Proc. Interspeech, Antwerp, Belgium, 2007.
7. H. Ney, “The use of a one-stage dynamic programming algorithm for connected word recognition,” IEEE Trans. on Acoustics, Speech and Signal Proc., vol. 32, no. 2, pp. 263-271, April 1984.
8. K. Ng, “Subword-based approaches for spoken document retrieval,” Ph.D. dissertation, Massachusetts Institute of Technology, 2000.
9. L. Rabiner, A. Rosenberg, and S. Levinson, “Considerations in dynamic time warping algorithms for discrete word recognition,” IEEE Trans. on Acoustics, Speech and Signal Proc., vol. 26, no. 6, pp. 575-582, December 1978.
10. M. Sarac¸lar and R. Sproat, “Lattice-based search for spoken utterance retrieval,” in Proc. HLT-NAACL, Boston, 2004.
11. P. Schwarz, P. Matˇejka, and J. ˇCernock´y, “Towards lower error rates in phoneme recognition,” in Proc. Int. Conf. on Text, Speech and Dialogue, Brno, Czech Republic, Sep. 2004.
12. W. Shen, C. White, and T. Hazen, “A comparison of query-by-example methods for spoken term detection,” in Proc. Interspeech, Brighton, England, Sep. 2009.
13. G. Tzanetakis, A. Ermolinsky, and P. Cook, “Pitch histograms in audio and symbolic music information retrieval,” Journal of New Music Research, vol. 32, no. 2, pp. 143-152, June 2003.
14. P. Yu, K. Chen, C. Ma, and F. Seide, “Vocabulary-independent indexing of spontaneous speech,” IEEE Trans. Speech Audio Processing, vol. 13, no. 5, pp. 635–643, Sept. 2005.
15. T. Zhang and C. Kuo, “Hierarchical classification of audio data for archiving and retrieval,” in Proc. ICASSP, Phoenix, March 1999.