A Survey on Automatic Question-answering process in Speech using Spoken term detection

M.Mamatha[#1], T.Bhaskar Reddy[*2]

[#]*Assistant Professor, CSE Dept, MGIT, Hyderabad, Telangana State, India.*

[*]*Assistant Professor, CSE Dept, SKU, Anatapur, Andhrapradesh State, India.*

Abstract**—** *The advent of WWW has reintroduced the need for user-friendly querying techniques in speech that reduce information overflow, and poses new challenges to the research in automated QA. The goal of current works of the unity of research of Technologies is to improve efficiency of e-learning by introducing intelligence into e-learning environments and automating a set of its features. The system allows learners to post subject related questions / doubts to the subject experts in speech. This usually requires the subject expert to answer the same query with different sentence framing a number of times. This paper discusses the development of an automated frequently asked questions retrieval system techniques in speech. This paper discuss few simple Speech Recognition and retrieval techniques using STD briefly***.**

Keywords— Speech Recognition, Spoken Term Detection(STD,LVCSR, HMM.

I. INTRODUCTION

eLearning is subject expert consultation through this service learners post subject related questions / doubts to the subject expert. The subject expert would then respond to the queries and clarify the doubts posed by the learners in speech. Over a period of time the repository of questions posed by the learners would grow enormously and there is a high probability of posting repeated questions with different sentence framing though they may have similar semantics. The development of automated response generation for frequently asked questions via voice/speech, would stimulate faster response. The goal of Spoken Term Detection (STD) technology is to allow open vocabulary search over large collections of speech content from educational video lectures. In this research, we address cases where search term(s) of interest (queries) are acoustic examples. This is provided either by identifying a region of interest in a speech stream or by speaking the query term. Queries often relate to named-entities and foreign words, which typically have poor coverage in the vocabulary of Large Vocabulary Continuous Speech Recognition (LVCSR) systems. This same concept is also useful in Educational institutes and companies, for customers/users post their questions in terms of spoken terms/voice and get voice response automatically via text to speech synthesizer. This type of query and search using voice saves a lot of time for the customers/users and also more natural and human friendly way.

II. DIFFERENT APPROACHES USED FOR SPOKEN TERM DETECTION ARE AS FOLLOWS:

1. Supervised approaches

(a) Acoustic keyword spotting based

(b) LVCSR (Large Vocabulary Continuous Speech Re-cognition) based

(c) Sub word recognizer based

(d) Query-by-Example (text based STD)

(e) Event based

2. **Unsupervised approaches: QBE (Query-by-example using template matching)**

(a) Frame based template matching

(b) Segment based template matching

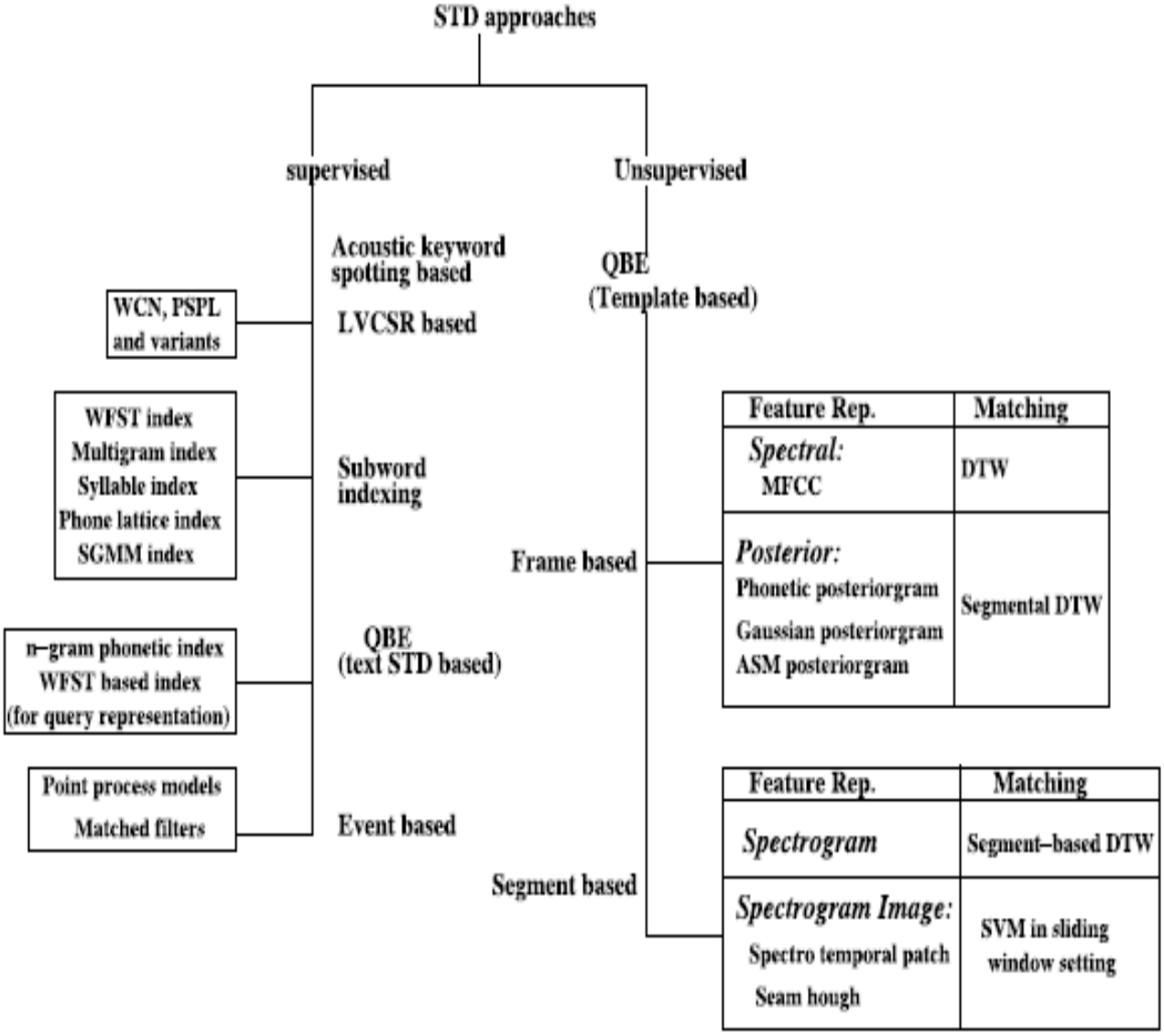

Fig 1:Taxonomy of approaches to STD

Much of the earlier research in this field originated in the framework of acoustic keyword spotting. Many of the present day STD systems use large vocabulary continuous speech recognition (LVCSR) technology that requires supervised training of HMMs (Hidden Markov Model) with huge amount of annotated speech and language resources. However, many of the new languages on which STD tasks are to be performed are under-represented in terms of resources required for building statistical models of HMMs used in LVCSRs.

The second category of these approaches are unsupervised methods based on template matching paradigm where the queried keyword template is matched with the target utterance for detecting a possible presence of the same. These approaches do not require the availability of any kind of labelled resources and hence are most suitable for under-represented languages.

Search for query in speech recording content, retrieve audio instances which are very similar to the query. Query may be speech or an audio clip spoken by a person. Preprocessing is done using an LVCSR system. Generally Word lattices are used for indexing and matching.

For template matching, almost all methods use some or the other variant of Dynamic time warping (DTW),that is essentially a dynamic programming based algorithm to find the degree of similarity between two time series differing in length. The DTW based approaches operate at frame level while comparing a keyword template with its counterpart in the target utterance.

The template based methods have two major steps. The first step provides a template representation of the spoken term. This is followed by matching of the template against a similar representation of the target utterance to determine the possible positions of occurrence of the term in the target utterance

III. QBE approaches using template matching

The principle of template matching has been used for keyword spotting since early days of speech recognition research. Later, the same was extended to detect keywords in a continuous utterance in a sliding window setting. In template matching, several examples of the spoken term to be detected are provided in the form of spoken queries. A template of the spoken term is created from the provided examples by deriving appropriate features. The similarity measure is computed using Dynamic Time Warping (DTW) a technique based on dynamic programming to measure the degree of match between two different sized vector sequences.

In this work, we present an unsupervised framework to address the problem of spotting spoken terms in large speech databases. The segment-based Bag of Acoustic Words (BoAW) framework proposed is inspired from the Bag of Words (BoW) approach widely used in text retrieval systems. Since this model ignores the sequence information in speech samples for efficient indexing of the database, a Dynamic Time Warping (DTW) based temporal matching technique is used to re-rank the results and restore the time sequence information. The speech data is stored efficiently in an inverted index which makes the retrieval very fast, thus making this framework particularly useful for searching large databases. We address the issue of choosing the appropriate size of the segment of speech for reliable indexing.

Template matching techniques are used for matching the query templates against audio segments in a test utterance in order to detect a possible existence of the spoken term.DTW techniques and its variants have been most widely used in this regard. Originally this technique was used for aligning examples of isolated words with reference keyword templates. Later it was extended to detect keywords in a continuous utterance wherein matching is done with segments of speech in a sliding window setting. The following gives a formal description of DTW.

IV. CONCLUSIONS

The paper presents a comprehensive survey of recent developments in the field of spoken term detection. It is seen from the survey that both supervised and unsupervised approaches for STD have own their advantages and disadvantages and the choice is made depending on the context of usage. The best performing supervised approaches clearly surpasses their unsupervised counterparts both in terms of speed and accuracy.

REFERENCES

[1] Baghai-Ravary, L., Kochanski, G., & Coleman, J. (2009*). Data-driven approaches to objective evaluation of phoneme alignment sys-tems.* In Proceedings of the 4th conference on human language technology , Poznan, Poland.

[2]. Anupam Mandal K.R. Prasanna Kumar Pabitra Mitra " *Recent developments in spoken term detection: a survey*" Springer Science+Business Media New York 2013

[3] .Barnwal, S., Sahni, K., Singh, R., & Raj, B. (2012). *Spectrographic seam patterns for discriminative word spotting. InProc. int. conf.acoustics, speech and signal processing* , Kyoto, Japan.

[4]. M.Mamtha, D.Kavitha,T.Swathi '*A Survey on automatic Question-Answering Techniques*' in IJRCM for publication in Volume No.3(2013),Issue No.10(October)

[5]. Bridle, J. (1973). An efficient elastic template method for detecting given key words in running speech. In Proc. of British acoustic society meeting, UK.

[6]. Can, D. (2011). Lattice indexing for spoken term detection. IEEE Transactions on Audio, Speech, and Language Processing

[7] Can, P., Cooper, E., Sethy, A., White, C., Ramabhadran, B., & Saraclar,M. (2009). Effect of pronunciations on oov queries in spoken term detection. In Proc. int. conf. acoustics, speech and signal process-ing , Taipei, Taiwan.

[8]. Chan, C., & Lee, L. (2010). Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping. In Proc. int. conf. speech processing