Proc. of the Fifth Intl. Conf. Advances in Computing, Communication and Information Technology- CCIT 2017

ISBN: 978-1-63248-131-3 doi: 10.15224/ 978-1-63248-131-3-54

Usage of acoustic cues in spoken term detection / keyword spotting for zero/low resource languages

Patha Sreedhar

Suryakanth V. Gangashetty

***Abstract*:**

The proposed work exploits acoustic cues at various levels and incorporates them in the present (Spoken Term Detection) STD frame work. Recently proposed new syllabification method [1] for speech signal is being used for STD. In STD, a query and reference speech signals are provided, these speech signals are syllabified using the new syllabification method and features like Mel-frequency cepstral coefficients (MFCC), posterior grams are extracted. These features are then matched using template based match techniques like dynamic time warping (DTW) at syllable level instead of regular frame level. This essentially reduces the unwanted matching done at frame level.

Keywords- Syllabifiaction, Spoken term detection, Dynamic Time warping.

I. INTRODUCTION

Spoken term detection (STD) / keyword spotting (KWS) refers to system that retrieves the utterances containing the query term or exact time spans of query term. Conventionally KWS refers to task with predefined keyword set but for STD, query can be any term including out of vocabulary (OOV). In this task, both query term and retrieval content are audio based.

If spoken query is used to retrieve text based content then it is called as voice search. In some cases of KWS, keywords can be in text format.

Although this kind of task is very successful with usage of automatic speech recognition (ASR) in cascade with text retrieval techniques, it fails drastically if ASR performance is poor. ASR performance can be improved using large hand labelled database which is very difficult to obtain in real world services or it can be improved by using very complex models like deep neural networks. But in spite of these efforts handling OOV terms in ASR is much challenging task.

Patha Sreedhar
PhD student, International Institute of Information Technology, Hyderabad
Reseach Scientist ,Pralhad P. Chhabria research Center, Pune

Suryakanth V. Gangashetty
Associate Professor, International Institute of Information Technology, Hyderabad

Hence developing a STD or KWS system which depends less on ASR accuracy or bypasses ASR completely or partially is required.

This essentially reduces the errors caused by ASR and complexity involved in developing full-fledged ASR.

This work focuses on developing the STD or KWS system which not only bypasses ASR but also exploits significant acoustic cues and matches at acoustic level, which significantly reduces the computing power. These acoustic cues will also help in developing the low resource automatic speech recognition or phonetic engine.

II. OVERVIEW

Many systems were developed by various researchers by using ASR or by bypassing ASR. These systems can be broadly classified into following categories [2]. Cascaded ASR with text information retrieval: The spoken content is converted into word or sub word sequences or lattices using ASR and then text retrieval techniques are applied. Modified ASR for retrieval purpose: Here the ASR and retrieval performances are jointly optimized instead of doing separately.

Exploiting the information not present in ASR output: Some potentially useful information like temporal structure is lost in ASR, hence this kind of information is added in complementary to ASR output to improve retrieval performance. Direct matching the acoustic level without ASR: The acoustic features are matched directly without going to ASR. Two stages are involved in accomplishing this task namely representation and matching.

Representation of speech signal plays an important role as it required to differentiate various underlying classes. This can be achieved in two ways: discrete and continuous representation of speech. In discrete representation, the speech signal is represented as discrete symbols by using large vocabulary continuous speech recognizers (LVCSR) or sub word unit recognizers. Where as in continuous representation, speech models are built using hidden Markov models (HMM), multilayer perceptron (MLP), Gaussian mixture models (GMM) etc. Representation of speech either

Proc. of the Fifth Intl. Conf. Advances in Computing, Communication and Information Technology- CCIT 2017

ISBN: 978-1-63248-131-3 doi: 10.15224/ 978-1-63248-131-3-54

discrete or continuous in turn can be achieved either from supervised or unsupervised way.

In matching stage, if speech is represented in discrete fashion, then lattice matching methods are followed and for continuous way of representation of speech template based matching techniques are used. In evaluating the system, two approaches can be employed: ranked and unranked retrieval results. In evaluating unranked retrieved results, the performance is evaluated based on correctness of retrieved objects while the order is not considered. Precision, Recall, F-measure metrics are used. In evaluating ranked results, where the order of retrieved results is important, Precision@N metric is used.

III. PROPOSED APPROACH

The proposed work exploits acoustic cues at various levels and incorporates them in the present STD frame work. As a first step, syllable level information is incorporated, the syllabification method of speech signal proposed in [1] is being studied for STD.

In STD, a query and reference speech signals are provided, these speech signals are syllabified using the new proposed method and features like Mel-frequency cepstral coefficients (MFCC), posterior grams are extracted. These features are then matched using template based match techniques like dynamic time warping (DTW) at syllable level instead of regular frame level. This essentially reduces the unwanted matchings done at frame level.

Syllabification method:

Speech is divided into syllables by locating syllable nuclei, several features like energy, fundamental frequency, duration helps in extracting the required information.

Many phoneticians, linguists and other writers did not converged to a particular definition of a syllable [3], hence working towards marking boundary of a syllable in speech signal is pretty difficult task.

The theory that explains syllable at its best is based on speech production mechanism. It is explained in terms of pulmonic air stream mechanism. It says that 'speaking is modified breathing'. This syllable producing movement of the respiratory muscles has been called a chest pulse or breath pulse or syllable pulse [3].

It can also be defined as a minimal pulse of initiatory activity bounded by a momentary retardation of the initiator, either self-imposed, or, more usually, imposed by a consonantal type of articulatory stricture [4]. Every syllable consists of a nuclei and it is mostly vowel, it optionally contains onset and/or coda. Onset and coda are generally consonants. Onset is a sound which occurs before the nuclei and coda is a sound which occurs after the nuclei. Syllable nuclei is detected using the concept of sonority. Sonority has a long history in literature, it is dated as back as 1876 when Sievers first attributed it as sound fullness. He explained it as relative loudness of speech sounds [5]. Ladefoged defined it as loudness relative to other sounds having same length, stress and pitch [6].

Clements defined it as relative resonance of speech sounds. A sonorant sound tend to have low degree of resistance or acoustic loss, leading to slow decay of formant oscillations, this leads to formant bandwidth reduction [7]. Sonority sequencing principle states that the sonority increases from onset to nuclei and then decreases to coda [5].

Sonority hierarchy [6] is as follows:

low vowels > mid vowels > high vowels > nasals > fricatives

Procedure:

ZFF (Zero Frequency Filtering) [8] analysis is done on speech signal to get fundamental frequency and voice – non voice decision. Speech signal is band passed in the range of 500 – 1700 Hz, which is considered as sonorant band [9] and ZTL (Zero Time Liftering) [10] analysis is performed over it to obtain strength of DRF (Dominant Resonant Frequency).

Spoken term detection:

For every query and reference speech file, the syllable level segmentation is done and the posterior grams are extracted. DTW is employed [12] for this query and reference file at syllable level. From the figure 1, it can be observed that there is match from syllable number 14 to 16 of reference speech file to that of query file.

This first approach employed used the posterior grams, which require rich resources for a particular language. For low/zero resourced languages, we need a set of acoustic cues which can match the syllables more effectively. Towards developing acoustic cues, features like center of gravity, dominant resonance frequency, strength of excitation, fundamental frequency and formants are being analyzed.

Proc. of the Fifth Intl. Conf. Advances in Computing, Communication and Information Technology- CCIT 2017

ISBN: 978-1-63248-131-3 doi: 10.15224/ 978-1-63248-131-3-54

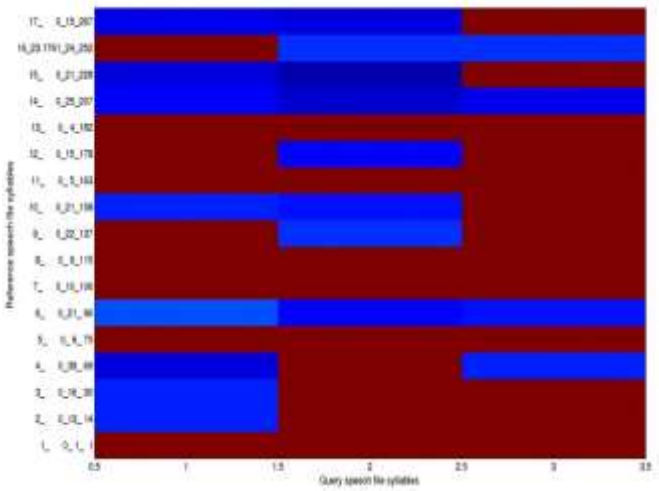

Figure 1: DTW employed at syllable level

The following Figures 2 and 3 shows how the proposed approach works when manually syllable boundaries are available and DTW is employed at syllable level. The results shows promising content which has to be exploited further.

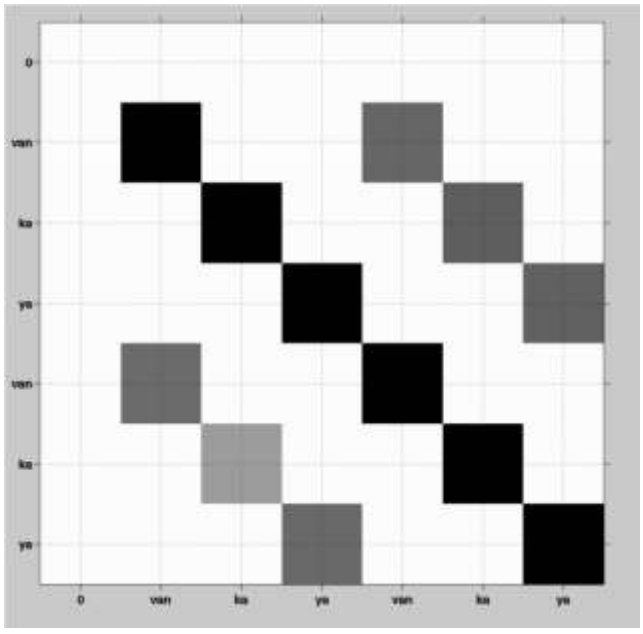

Figure 2: Proposed approach with manual syllable boundaries of a Telugu utterance

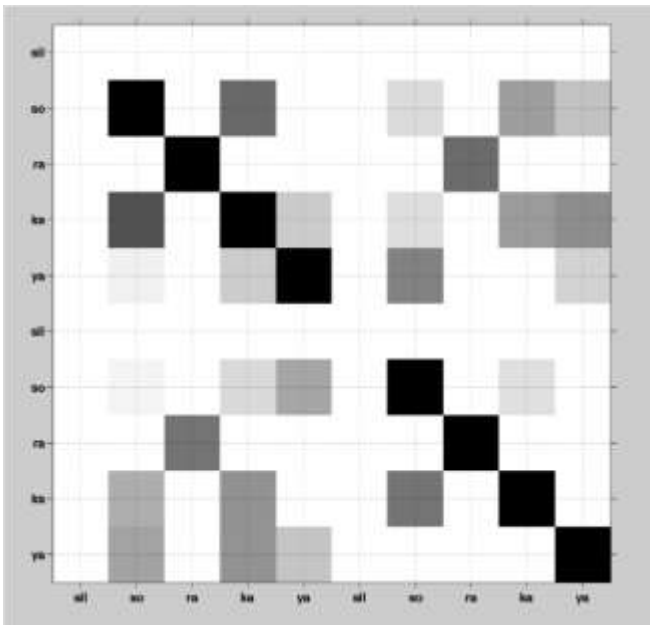

Figure 3: Proposed approach with manual syllable boundaries of a Telugu utterance.

ACKNOWLEDGMENT

The authors would like to thank Prof. Bayya Yegnanarayana for his valuable suggestions and guidance.

REFERENCES

[1] Sreedhar Patha and Yegnanarayana Bayya and Suryakanth V Gangashetty, “Syllable nucleus and boundary detection in noisy conditions,” in Speech Prosody 2016, Boston USA.

[2] Lin-shan Lee, James Glass, Hung-yi Lee, and Chun-an Chan, “Spoken content retrieval-Beyond cascading speech recognition with text retrieval,” in IEEE/ACM transactions on audio, speech, and language processing, Vol 23, No. 9, September 2015.

[3] D. Abercrombie, Elements of general phonetics, 1 ed., Edinburgh: Edinburgh University Press, 1982, pp. 34-38.

[4] J. C. Catford, A practical introduction to phonetics, 2 ed., New York: Oxford university press, 2010, pp. 168-170.

[5] S. Parker, "Sound level protrusions as physical correlates of sonority," Journal of Phonetics, no. 36, pp. 55-90, 2008.

[6] P. Ladefoged and K. Johnson, A course in phonetics, 6 ed., Delhi: Cengage Learning, 2012.

[7] G. N. Clements, "Does sonority have a phonetic basis?," in Contemporary views on architecture and representations in phonological theory, MIT press.

[8] K. S. R. Murthy and B. Yegnanarayana, "Epoch extraction from speech signals," IEEE Trans. Audio, Speech, Lang. Process., vol. 16, no. 8, pp. 1602-1613, 2008.

[9] Y. Nakajima, K. Ueda, S. Fujimaru, Y. Ohsaka and Y. Ohsaka, "Sonority in British English," Proceedings of meetings on acoustics, vol. 19, June 2013.

[10] N. Dhanunjaya, "Signal processing for excitation-based analysis of acoustic events in speech," Ph.D. Thesis, IIT-Madras, 2011.

ISBN: 978-1-63248-131-3 doi: 10.15224/ 978-1-63248-131-3-54

[11] S. Errede, "Lecture notes: Illinois University," [Online]. Available: https://courses.physics.illinois.edu/phys406/Lecture_Notes/P406P OM_Lecture_Notes/P406POM_Lect7.pdf. [Accessed 24 July 2014].

[12] G. Mantena, S. Achanta, and K. Prahallad, "Query-byexample spoken term detection using frequency domain linear prediction and non-segmental dynamic time warping," submitted to IEEE Trans. Audio, Speech and Lang. Processing, 2013.