International Archive of Applied Sciences and Technology
Int. Arch. App. Sci. Technol; Vol 9 [2] June 2018 : 41-44

[ISO9001: 2008 Certified Organization]
www.soeagra.com/iaast.html

IAAST
ONLINE ISSN 2277- 1565
PRINT ISSN 0976 - 4828

CODEN: IAASCA ORIGINAL ARTICLE

DOI: .10.15515/iaast.0976-4828.9.2.4144

Development of Algorithm to Detect End of Utterance in Continuous Speech for Gujarati Language

Jinal Tailor and Dipti Shah
[1]Ph.D Scholar, S.P.University, Vallabh Vidhyanagar, Gujarat, India
[2]Professor, G.H.Patel Post Graduate Dept. of Computer Science, S.P. University, Vallabh Vidhyanagar, Gujarat, India
Email: jinal.tailorssa@gmail.com, dbshah66@yahoo.com

ABSTRACT
Speech Recognition System (SRS) is the emerging field of Natural Language Processing (NLP). Speech Recognition System involves methodology and techniques to convert acoustic signals in form of speech into text. SRS involves acoustic model and language model to convert audio signal into speech. Language model matches with word units and similar sounds for conversion. This emerging area involves many challenges such as real time response, system accuracy and performance. One of the major challenges is finding End of Utterance in isolated word recognition and in continuous speech. The system is required to find when the speaker stop speaking and waiting for system response. Delay in response can decrease system performance. It is difficult job to find start and end of sentences in continuous speech as speaker is delivering content in a natural way and it is not well defined. Sometimes it is difficult to identify pauses between words and end of utterance as for some of the speakers all are having different rhythm of speaking. We have presented proposed algorithm to find end of utterance in continuous speech for Gujarati language. This proposed work distinguish end of utterance of speech by analyzing time instant between frequencies and defining words to terminate system. This proposed approach also includes minimization of false alarm generated by speaker and improve accuracy and response time. The average performance ratio achieved for EOU identification is 94.8%.
Key words: *End of Utterance, Gujarati Language, Speech Recognition System, Threshold*

Received 12/01/2018 Revised 26/03/2018 Accepted 19/05/2018

Citation of this article
Jinal Tailor and Dipti Shah. Development of Algorithm to Detect End of Utterance in Continuous Speech for Gujarati Language. Int. Arch. App. Sci. Technol; Vol 9 [2] June 2018. 41-44.

INTRODUCTION

Speech Recognition System involves major features to provide ease to speaker by converting speech into text format. Basic idea of SRS is to convert acoustic speech signal into meaningful abstract form of text. The most common applications of SRS are dictation in the domain of business, education, medical or legal notes, hands-free environment, and real-time embedded system as well as for the people with physical disabilities. It is more useful when this task gets done in real time. Low processing speed and less accuracy in SRS can affect user satisfaction. To operate the system in real time some of the major challenges have to be resolved such as real time feedback to speaker and detection of End of utterance (EOU) [1-3]. It is important to detect EOU in continuous speech that identify when speaker stops speaking and waiting for the system action. Delay in detecting EOU can waste speech recognition resources as well as lose user interest towards operating the system. EOU detection algorithm mainly focuses upon differentiation between long pauses between words and EOU. Each speaker has his own style of speaking with short as well as long pauses. Sometimes when speaker is delivering spontaneous speech it might be the chance to get

maximum number of pauses occur during speech. The proposed algorithm focuses upon specifying threshold value to check with pauses and EOU [4-6]. Time distinct that exceeds threshold value considered as EOU. Another major challenge occur in identification of EOU is, in continuous speech when an actual EOU occurs speaker has to wait till the threshold value time decided by the system. This will increase extra waiting time for the speaker to get response from the system.

MATERIALS AND METHODS

Pre – processing of speech

Speech is the analog signal that is converted into text format. Speech recognition involves many challenges related to real time response to the speaker. It is difficult to identify EOU in continuous speech as some speakers take more time while speaking words that are conjunct words, sometimes unfamiliar literature that makes speaker conscious while speaking and he takes more time between words. These long pauses are difficult to differentiate with EOU through the system. Long pauses generate false alarms which system needs to ignore to reduce confusion. The detection algorithm must have to detect these false alarms to easily identify EOU occur in continuous speech.

End point detection algorithm

Speech recognition system requires finding ending point in speech signal waveform. End point recognition algorithm focuses upon voiced region and unvoiced region in the speech. To define voiced and unvoiced region we need to divide speech signal into various frames. Different frames considered as samples of different frequencies.

The endpoint detection algorithm contains various steps.

In first step we need to analyse EOU with comparing unvoiced region. If we have total N samples then squared value of samples are added and divided by N for each speech signal to calculate average power of each frame. Average power is compared with standard threshold value defined in algorithm. If average power is greater than threshold value then it is considered as voiced region and if the value is less than threshold value then it is considered as unvoiced region.

Another feature of speech signal is pitch or fundamental frequency F0 can be analysed for EOU. F0 is found constant during continuous speech but analysed that value of fundamental frequency drops down while end of utterance is about to occur. But dropped occurrence of F0 other than EOU can also be occurring when speaker is trying to speak conjunct word or unfamiliar content.

Zero-crossing rate in Hz:

$$Zcross = n\_cross \cdot \frac{f_s}{N} \qquad (1)$$

Zero crossing rates are rate at which signal changes from positive to negative and vice versa. ZCR is used in voice activity detection technique to check that presence of speech in signal equation (1). In ZCR it measures the amplitude of the speech signals falls to zero for number of times in a given time interval or frame.

Second step contains identification of interjections used in speech. It is a normal tendency of a speaker that whenever he speaks any interjection words, he takes longer pause than normal pause between other utterances. This longer pause can be greater than threshold value considered to be as silence. This will generate false alarm to the system. To overcome this false alarm we can check the previous utterance spoken by the user. If it is interjection word then it will not be considered as EOU and consider as false alarm.

Major problem when dealing with EOU is that when the speaker is already done but still has to wait until threshold value to generate EOU breakpoint. This mechanism delays the runtime response time from the system. We have proposed approach to minimize this response time. According the approach, we can assign particular unusual word to consider defining EOU explicitly by speaker after ending his speech. As in Gujarati Language we can use word “પૂર્ણ” or “સંપૂર્ણ” to explicitly terminate the system by defining EOU by the speaker. Each speaker has to utter this predefined words to work as EOU and no need to wait till the threshold value. This will resolve the problem related to waiting response time. It is possible

that the speech of speaker contains "પૂર્ણ" or "સંપૂર્ણ" in meaningful context. At that point system will not recognize these words as meaningful context and consider as EOU and get terminated. This will generate problem in case where speaker has to add these words in his speech. For this solution we can generate other sequence of characters or word that does not have any meaning. For example "અખ:", "ખ:" and "ખે:" etc. according to Gujarati language these words have no such meaning and has zero probability to present in any speech. We can select and define any sequence from this and consider as EOU that can be used by the speaker to terminate the system after completing speech.

We can use language model to calculate probability distribution p (w) for occurrences of word W in speech. P (w) can be decomposed as equation (2).

$$P(w) = P(w1,w2,\ldots,wn)$$
$$= P(w1)P(w2|w1)P(w3|w1,w2)\ldots P(wn|w1,w2,\ldots,wn-1) \quad (2)$$
$$= \pi P(wi|w1,w2,\ldots wi-1)$$

Where p (wi|w1, w2, wi-1) is the probability that wi will follow given that the word sequence w1, w2, wi-1 in above equation.

Where we can find that P ("છે") =0.01 as in Gujarati language every hundred statement we find "છે" word at the end. While P ("અખ :") =0 since it is meaningless word to add in Gujarati language speech. Same happen for P ("ખ :") and P ("ખે :")

RESULT AND DISCUSSION

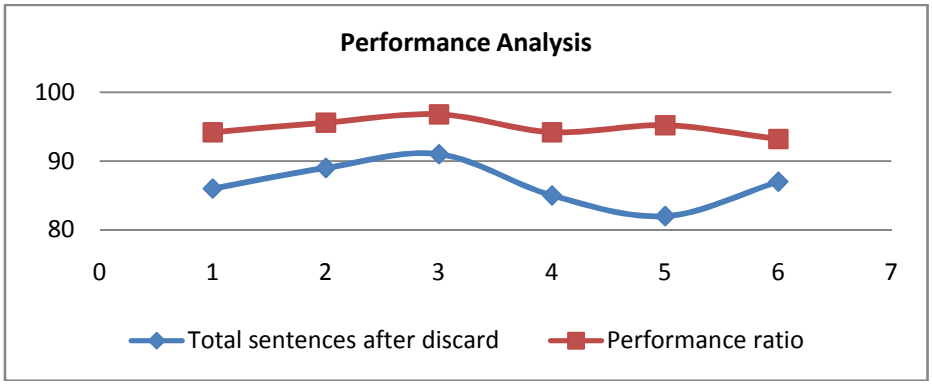
For experimental setup we have used audacity software for frequency analysis. Speech data recorded at 16KHz. Total 6 speakers are selected gender wise to speak continuous routine Gujarati speech. We have prepared total list of interjections that contains 8 words. We have discussed use of interjections in their speech to delay the pause duration after that. Total 600 sentences were recorded by total 6 speakers (100 sentences to each). From which 80 sentences were discarded due to low quality. Total 520 sentences were processed to detect EOU. From total speakers 3 speakers were instructed to terminate the system using "પૂર્ણ" or "સંપૂર્ણ". Other 3 speakers were instructed to add these words in their speech and terminate the system by speaking "અખ:", "ખ:" or "ખે:" We have analyzed output as well as response time to the speaker.

The experimental results are:

Table 1. Algorithm Performance Measures

Speakers	Total sentences after discard	Total No. of Utterances	Identified utterances	Identification of EOU	Error rate for EOU	No of utterances that generated False alarm	Performance ratio
S1	86	1032	1016	81	5.8%	16	94.2%
S2	89	979	969	85	4.4%	10	95.6%
S3	91	910	903	88	3.2%	7	96.8%
S4	85	1020	1009	80	5.8%	11	94.2%
S5	82	902	883	78	4.8%	9	95.2%
S6	87	870	855	81	6.8%	15	93.2 %

Performance criteria of three techniques used to detect EOU:



The average performance ratio for identification of EOU is 94.8%. Average error rate to detect EOU is 5.1 %. Total identification ratio of utterances is 98.63%. We can improve system performance by using highly configured devices as well as improved noise removal techniques.

CONCLUSION

We have proposed approach to detect EOU by minimizing waiting response time till threshold value. There are multiple approaches to find EOU detection such as dividing each speech signals into frames and calculate average power of each frame and compare with standard threshold value. If the average power of frame is less than threshold value then it is consider as EOU. But in some situation it can generate false alarm when low pitch frequency found due to environmental variability. It is found that in general scenario when speaker use interjections to express feelings he takes longer pause then normal pause. When we found longer pause greater than threshold value we first need to check previous word. If previous word is any interjections then we need to consider it as false alarm not as EOU. Through this approach we can minimize generation of false alarm. Another major issue found is when speaker is already completed speech he has to wait till threshold value. Both the issues get resolve by proposed approach for ending speech to speak pre-defined words “પૂર્ણ” or “સંપૂર્ણ”. This approach provides proper detection of EOU when speech does not contain words “પૂર્ણ” or “સંપૂર્ણ”. For all the possible sentences we have defined some sequence of characters in Gujarati language that has no meaning and zero probability to occur in any speech such as “અખ:”, “ખ:”, and “ખે:”. We have achieved system performance for detection of EOU is 94.8%. Average Error rate is found 5.1% which can be minimize using improved noise removal techniques.

REFERENCES

1. Tailor J.H., Shah D.B. (2018) HMM-Based Lightweight Speech Recognition System for Gujarati Language. In: Mishra D., Nayak M., Joshi A. (eds) Information and Communication Technology for Sustainable Development. Lecture Notes in Networks and Systems, vol 10. Springer, Singapore
2. Tailor, J. H., & Shah, D. B. (2016). Speech Recognition System Architecture for Gujarati Language. International Journal of Computer Applications, 138(12).
3. Akila, A., & Chandra, E. (2014). Comparative study of endpoint detection algorithms suitable for isolated word recognition. BVICAM's Int. J. Inf. Technol., 6(2), 764-766.
4. Arsikere, H., Shriberg, E., & Ozertem, U. (2015, March). Enhanced end-of-turn detection for speech to a personal assistant. In AAAI Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction.
5. Atterer, M., Baumann, T., & Schlangen, D. (2008). Towards incremental end-of-utterance detection in dialogue systems. In Proceedings of the 22nd International Conference on Computational Linguistics.
6. Bachu, R. G., Kopparthi, S., Adapa, B., & Barkana, B. D. (2008, June). Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal. In American Society for Engineering Education (ASEE) Zone Conference Proceedings (pp. 1-7).