International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-12, October 2019

Voice Based Retrieval using Convolution Neural Network in Deep Learning

M.Mamatha, T.Bhaskar Reddy

Abstract: ***In this paper, we propose a Content Based Voice Retrieval(CBVR) is used to search a specific audio files from a large data base. Using Deep learning features are learned automatically in the training phase. Convolution Neural Network is used in this research for CBVR. On this paper, we recommend a novel technique to key word search(KWS) in low-resource languages, which presents an replacement method for retrieving the phrases of curiosity, in particular for the out of vocabulary (OOV) ones. Our procedure contains the approaches of question-by using-illustration retrieval tasks into KWS and conducts the hunt by the use of the subsequence dynamic time warping (sDTW) algorithm. For this, text queries are modeled as sequences of function vectors and used as templates within the search. A Convolution neural network-headquartered model is informed to gain knowledge of a frame-degree distance metric to be used in sDTW and the right question model frame representations for this realized distance. This new procedure can be used as a substitute to traditional LVCSR-situated KWS programs, or in combination with them, to attain the intention of filling the gap between OOV and in-vocabulary (IV) retrieval performances.***

Keywords **: keyword search, low resource languages, out of vocabulary (OOV) terms, query modeling, distance metric learning, subsequence dynamic time warping.**

I. INTRODUCTION

Within the digital generation, giant amounts of audio information are being produced and consumed every day in a type of languages. We use it day-to-day in a type of contexts to communicate with this world. This may be in the form of TV news, class room lectures ,audio books, call center archives and even personal audio recordings. 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. This democratization of media, enabled by easy access to technology, has made digital speech a medium of vast and heterogeneous information and a valuable knowledge resource. Naturally, the ability to search through this data becomes a valuable functionality to improve its access.

Revised Manuscript Received on October 30, 2019.
***** Correspondence Author
M.Mamatha*, ResearchArea-Speech Processing
T.Bhaskar Reddy,Membership-ICITE,GATE Qualified with good score Professor in SKU,Anatantapur,AP. Research Area- Computer Science

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.

CBVR conjointly referred to as question By Voice Content(QBIC) is that the application of voice retrieval looking out and/or browsing a information of digital audio supported the content of the audio. “Content based” implies that the search can analyze the particular contents of the audio. The term content in this context might refer to words spoken, music etc , or any other information derived from the audio itself. A query Audio provides rich content information Finding the relevant Audio by sequentially comparing the low level Audio features of the query Audio with those of key features of Audio in database takes long response time for query computation.

Finding similarity measure requires key features matching and hence computing distance parameter across the entire feature vector array of the whole Audio Feature Database. These huge computations cause long response time to the users and thus, the problem of high computation cost in computing Audio features persistent. Raw audio recordings are analyzed and segmented based on abrupt changes of features. Audio segments are classified and indexed. They are stored in corresponding archives. The audio archives are organized in a hierarchical way for the ease of the storage and retrieval of audio clips. The query audio samples in the archives, he may put a set of features or a query sound into the computer. A brief flow chart furnished in Fig 1.

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

is prepared as per journal the template. 3. Contents of the paper are fine and satisfactory. Author (s) can make rectification in the final paper but after the final submission to the journal, rectification is not possible.

II. METHODOLOGY

The search engine will then find the best matched sounds and present them to the user.

Uncooked audio recordings are analyzed and segmented situated on abrupt changes of points.

Audio segments are labeled and listed. They are stored in corresponding archives.

Step 1:Voice preprocessing - approach the input audio utilizing MFCC(Mel Frequency Cepstral Coefficients) encoding to extract the features.

Mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, headquartered on a linear cosine transform of a log vigor spectrum on a nonlinear mel scale of frequency

Step2 : Voice Feature Extraction - Extract the regions of the Voice parts of the input Audio which has the features that can be used for finding Audio Similarity

Step 3: Relevance Scoring Algorithm - Compare the feature present in the extracted Object to the other Audio in the repository and calculate relevance score. Use the Relevance score to Sort the results in ascending order of relevance score to get the Most Relevant Audio at the top.

Convolution Neural Network

Convolutional Neural Networks(CNN) are a special type of Deep Neural Network focusing on Image Processing. Convolutions and Pooling layers are 2 special type of layers helpful for Image recognition in CNNs

CNNs have following 3 parts in their Hidden Layers

1.Convolution Layers

2.Pooling Layers

3.Fully Connected Layers

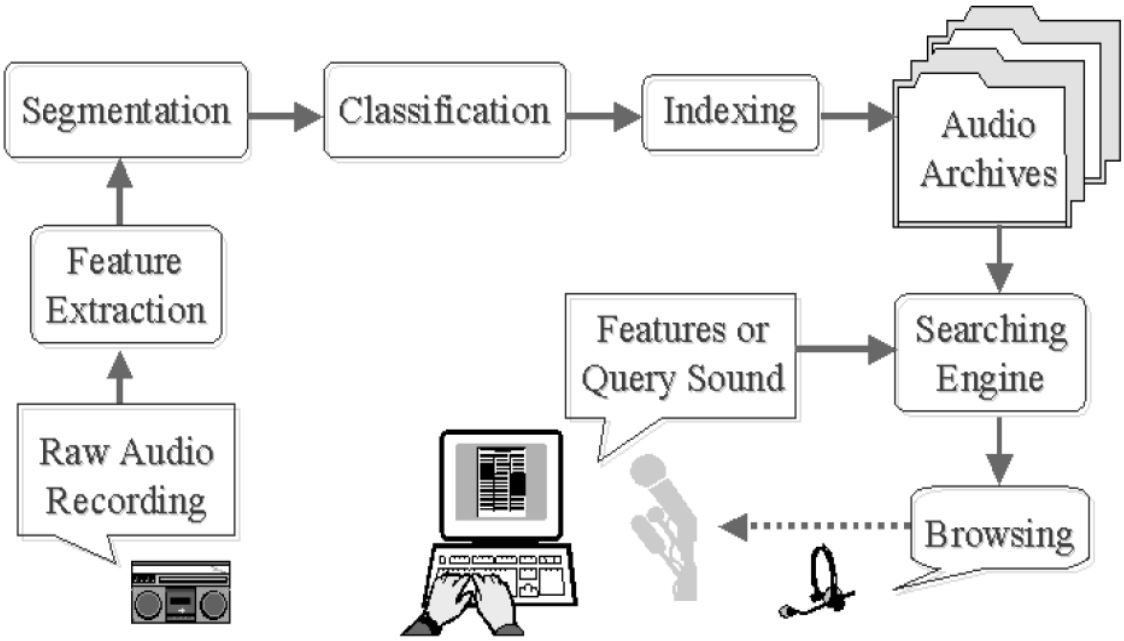

Fig 1: A generic architecture for search engine on CBVR.

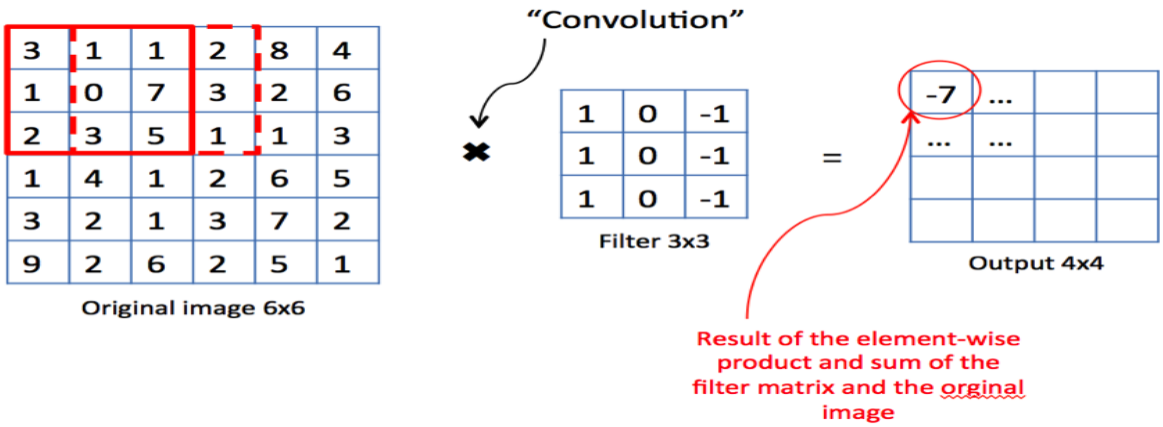

Fig 2: Convolution layer for Matrix operations

Why CNN for CBVR

CNNs are special type of Neural networks which work very well with data that is spatially connected. MFCC output is like a Image of 2D Vector. Hence normal neural networks will not be able to accurately understand the 2D Vector features correctly CNNs can Automatically learn features from Spatial and Temporal Relationships

CNN(Convolution Layers)

Convolution Layer in CNN is the first Hidden layer .Convolution layer performs

Convolutions which is a Matrix multiplication operation with specific types of Filters.

1. Filters are usually a 3x3 matrices which convolves the MFCC data into processed forms like word Detection, music pattern Detection, etc
2. Each 3x3 sub section of original matrix are multiplied with the filter and result is stored as output and we take fixed stride to start next matrix multiplication operation.

This also reduces the input matrix dimension Pooling layer helps in extracting the important features and removing unessential parts of an image. A brief flow char furnished in Fig 2.

CNN Pooling Layers

- Pooling also helps in Dimension reduction and it also it gives to some extent rotational invariance
- Common pooling options are Min, Max, Sum, Average
- Given below is an example of max pooler and average pooler

III. CONCLUTION AND FUTURE WORK

CBVR is used to search a specific audio files from a large database. Using Deep learning features are learned automatically in the training phase. Convolutional Neural Network is used in this research for CBVR.Feature based relevance scoring for Content Based Voice Retrieval using Relevance scoring algorithm like cosine similarity is used for ranking. More research needs to be done to obtain better representation techniques. Although Gaussian posteriorgrams are easier to model when only very less resources exist for a language, better representation techniques need to be explored to improve search accuracy. Better signal matching algorithms need to be developed. Better variants of Dynamic Time Warping could be explored

REFERENCES

1. S. Khudanpur, G. Chen, D. Povey, , D. Yarowsky, O. Yilmaz and J. Trma, "Quantifying the value of pronunciation lexicons for keyword search in lowresource languages," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 8560–8564.
2. D. Wang, J. Frankel, J. Tejedor, and S. King, "A comparison of phone and grapheme-based spoken term detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 4969–4972C
3. B. Gündoğdu, M. Saraçlar and B. Yusuf "Joint learning of distance metric and query model for posteriorgram based keyword searching" *IEEE* vol. 11, no. 8, pp. 1318–1328, Dec. 2017.
4. D.Allauzen, M. Mohri, and M. Saraclar, "General indexation of weighted automata: Application to spoken utterance retrieval," in *Proc. Workshop Interdiscip. Approaches Speech Indexing Retrieval HLT-NAACL*, 2004, pp. 33–40.
5. D. Karakos and R. Schwartz, "Subword and phonetic search for detecting oov keywords," in *Proc. Interspeech*, 2014, pp. 2469–2473.
6. X. Anguera, L. J. Rodriguez-Fuentes, I. Szoke, A. Buzo, and F. Metze, "QbE STD evaluation on low-resource languages," in *Proc. Int. Workshop Spoken Lang. Technol. Underresourced Lang.*, 2014, vol. 24, pp. 24–31.
7. A. Jansen, C. Liu, G. Chen, J. Trmal, , K. Kintzley and S. Khudanpur, "Low- resource open vocabulary KW search using point process models," in *Proc. Interspeech*, 2014, pp. 2789–2793..
8. A. Garcia and H. Gish, "Keyword spotting of arbitrary words using minimal speech resources," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2006, vol. 1, pp. 949–952.
9. S.w. Lee, K. Tanaka, and Y. Itoh, "Combining multiple subword representations for open-vocabulary in SDR" in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2005, vol. 1, pp. 505–508.

AUTHORS PROFILE

M.Mamatha, M.Tech, Researchscholor, Ru,Kurnool. Researcharea-Speech Processing ,Membership-Icite,Gate Qualified With Good Score.

T.Bhaskar Reddy, Ph.d,, Professor in SKU,Anatantapur,AP. Research Area- Computer Science