Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 8, August : 2024

A CTC Framework Approach to Speech Recognition in Collaborative Programming IDEs

Sasmita Lenka, Aliva Pattanaik,Pratyasha Pradhan

Dept. of Computer Science and Engineering, GIFT Autonomous, Bhubaneshwar, 752054, India
Email : sasmitalenka@gift.edu.in

***Abstract*— Speech Recognition is the process of automatically Recognizing a certain word spoken by a particular speaker based on individual information included in speech waves. This technique makes it possible to use the speakers voice to verify his/her identity and provide controlled access to services like voice based biometrics, database access services, voice based dialing , voice mail and remote access to computers. Voice can be a powerful tool for use in human computer interaction because it is the fundamental means of human communication. With the rapid growth of wireless communications, the need for voice recognition techniques has increased greatly. Portability and wearability , which are necessary items for being computationally powerful computer devices, will be reinforced by attaching voice applications, since voice can support invisible communication with a computer device as a natural way of communicating. Signal processing front end for extracting the feature set is an important stage in any speech recognition system. The optimum feature set is still not yet decided though the vast efforts of researchers. There are many types of features, which are derived differently and have good impact on the recognition rate. This project presents one of the techniques to extract the feature set from a speech signal, which can be used in speech recognition systems.**

Keywords: **Speech-to-Text, Text-to-Speech,ISA,TPA,CTC.**

I. INTRODUCTION

With advances in new technologies, computer devices have grown in popularity to become one of the most common consumer devices. Even as these devices are shrinking in size, however, their capability and content are changing into more complex and diverse functionalities to meet user requests. Now, it is common for many computer devices to include a phone, a personal directory, a memo capability ,an alarm clock, a scheduler, a camera, games and several applications which were working in Personal Digital Assistants(PDAs) before, so there is no boundary between computer devices and PDAs.

However computer devices in which designers have worked more and more to decrease their size are likely to have small keypads and screens, whereas they should have more complex and diverse functions for users. Their functionality and ease of use are greatly limited, and thus many researchers look to find alternative communication channels when interacting with these devices. In recent years, many researchers in the area of human computer interaction (HCI) have attempted to enhance the effectiveness and efficiency with which work and other activities are performed using voice based interfaces. Even if voice technology has been explored for use in desktop computer and telephone information system, the role of voice in interfaces has received little attention because of its difficulty of use and tiresomeness of recognition.

Actually, in the past, the accuracy of voice recognition was unacceptably low, and it's a role in a system was questionable because of ambiguity and error. However, voice technology has reached the point of commercial viability and reliability now, and also many computer devices adapt voice applications for providing better services to users. Using voice allows the interface size to be scaled down because voice interaction requires only audio I/O devices such as a microphone and speaker, which are already quite small and inexpensive. Currently, in a computer device voice interfaces need only small space and power consumption, but are able to provide every user with a friendly interfaces are sufficient to replace graphical user interfaces for accessing all information and content without using keywords, buttons and touch screens, since voice is the fundamental means of human communications.

II. OBJECTIVE

1. To build voice operated system for physically disabled persons.
2. To design hardware for voice recognition and corresponding action.
3. To recognise the voice of the person by analysing the speech signal.
4. To increase the speed of the task execution on processor.
5. To calculate time complexity of normal access and voice based access.

III. LITERATURE SURVEY

Connectionist Temporal Classification Lattices, efficient and effective modular speech recognition approaches, second pass rescoring for large vocabulary continuous speech recognition and phone based keyword spotting are also proposed [1]. The proposed voice casting system explores Gaussian mixture model based acoustic models and multilabel recognition of pre-received paralinguistic contain for the voice casting of professionally acted voice [2]. Recent advances in the field of speaker recognition have resulted in highly efficient speaker comparison algorithm [3]. The idea of using the GMM super vector machine classier. We proposed two new SVM kernels based on distance matrices between GMM models [4]. For modern human system interaction, traditional methods based on keyboard and mouse use do not prove sufficient. Especially when thinking about the poor and ill-qualified citizens of the Information Society, we must try to find the easiest and the most comfortable method for human system interaction [5]. To handle the variety of spontaneous effects in human-to-human dialogs, special noise models are introduced representing both human and non-human noise, as well as fragments. It is shown that both the acoustic and the language modelling of the noise increase the recognition performance significantly. In the experiments, a clustering of the noise classes is performed and the resulting cluster variants are compared, thus allowing one to

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 8, August : 2024

determine the best trade-off between the sensitivity and trainability of the models [6]. In this article we pick up the idea of using Hidden Markov Models (HMMs) to recognize emotions from speech signals and we describe the enhancements and optimizations of a speech-based emotion recognizer jointly operating with automatic speech recognition [7]. This paper focuses on resolving a number of issues that appear when the performance of human speech recognition is compared to that of automatic speech recognition. In particular human experimental data suggest that the resulting error is a product of the individual streams [8]. Emotional speech recognition is an interesting application that is able to recognize different emotional states from speech signal. In Human-robot interaction (HRI) , emotion recognition is being applied on intelligent robots so that they can understand emotional states of user and interact in a more human-like manner [9]. It offers information about the structure and staff of laboratories , the location and phones of departments and employees of the institutions [10].

IV. PROPOSED AND IMPLEMENTED SYSTEMARCHITECTURE

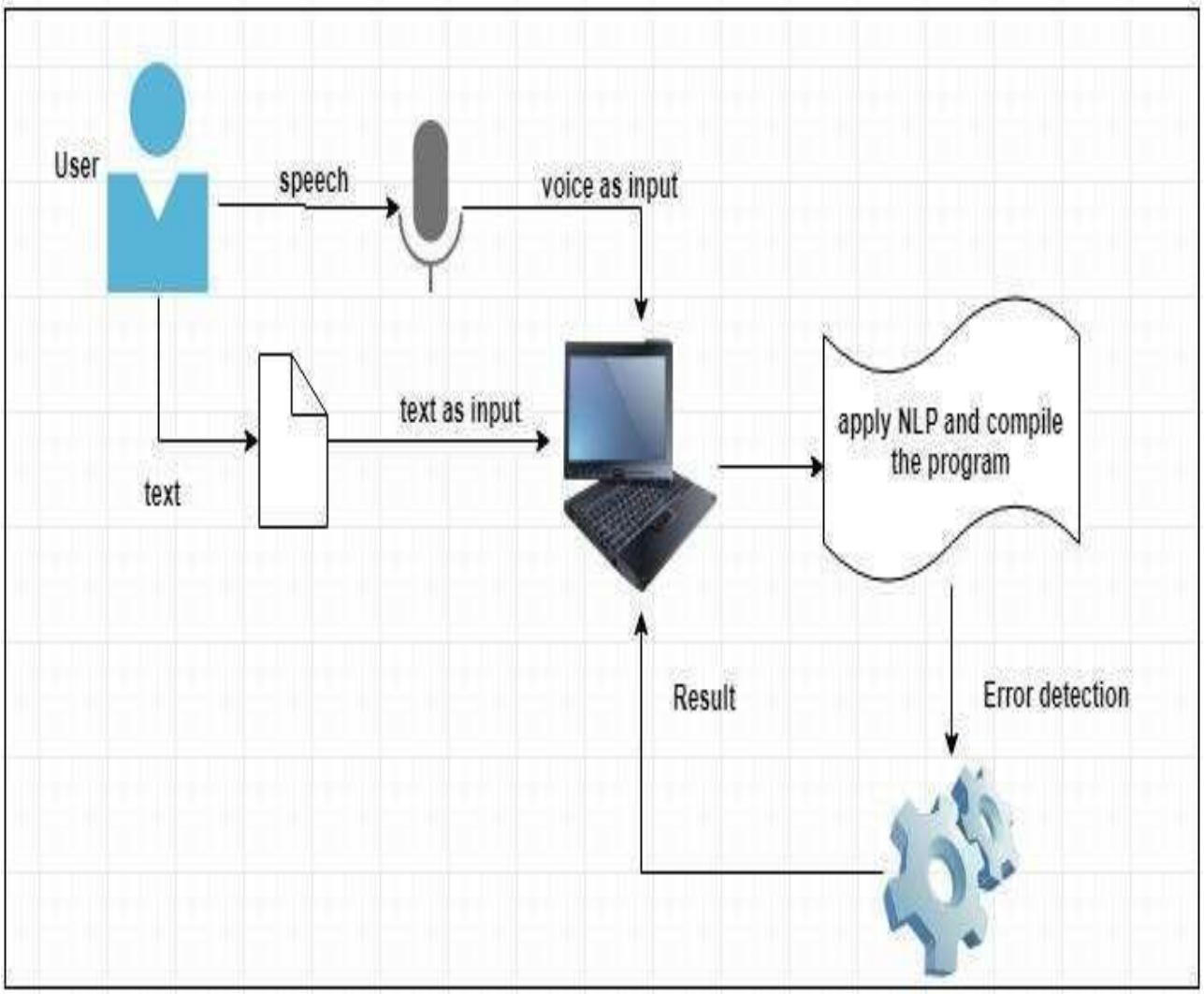

Fig.1: System Architecture

Speech recognition is the inter-disciplinary sub-field of computational linguistics that develops the methodologies and the technologies that enables the recognition and the translation of the spoken language into text by computers. It is also known as automatic speech recognition (ASR), computer speech recognition or speech to text (STT). It incorporates knowledge and research in the linguistics, computer science and electrical engineering fields.

Some speech recognition systems require "training" where an individual speaker reads text or isolated vocabulary into the system. The system analyses the person's specific voice and uses it to fine-tune the recognition of that person's speech, resulting in increased accuracy. Systems that do not use training are called "speaker independent" systems and the systems that use training are called "system dependent". The term voice recognition or speaker identification refers to identifying the speaker, rather than what they are saying. Recognizing the speaker can simplify the task of translating speech in systems that have been trained on a specific person's voice or it can be used to authenticate or verify the identity of a speaker as a part as a security process.

A system architecture is the conceptual model that defines the structure, the behaviour and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviours of the system.

In the system architecture, the user has to fulfil his/her personal details first. Then they are ready to use this system. After filling the details, the user can use any mode out of the two modes of input i.e. speech or the text as input. If the user selects speech as input then whatever the input is there given by speech is converted into text in the PC. After giving the input, Natural Language Processing (NLP) will be applied and he program will be completed. Then the program is compiled and if the errors are detected it will go to the input form it resolves the error then again compile the program till there exists no error and then gives us the output of the program. Our main aim of the system is to give input in the speech form, apply NLP and give the output of the program. The main system is all about speech-to-text conversion and text-to-speech conversion.

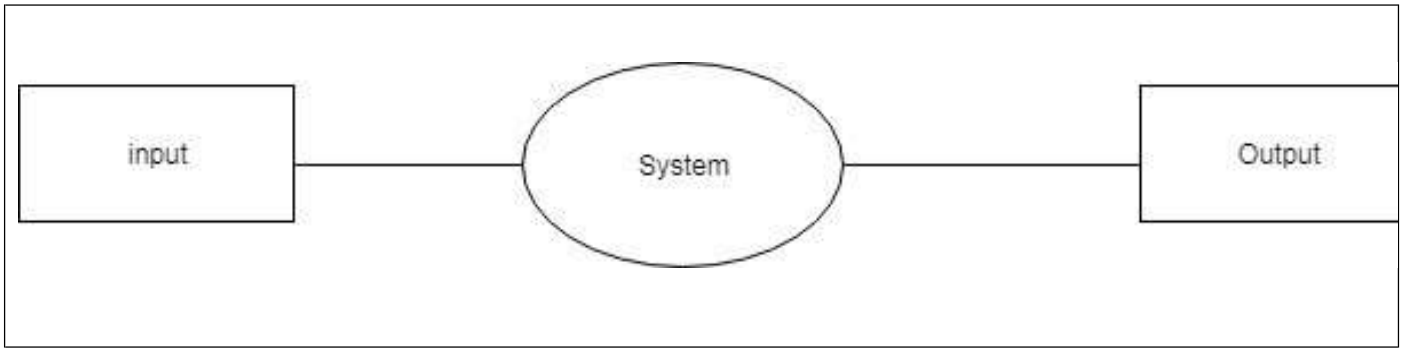

Fig2. Components of the System

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 8, August : 2024

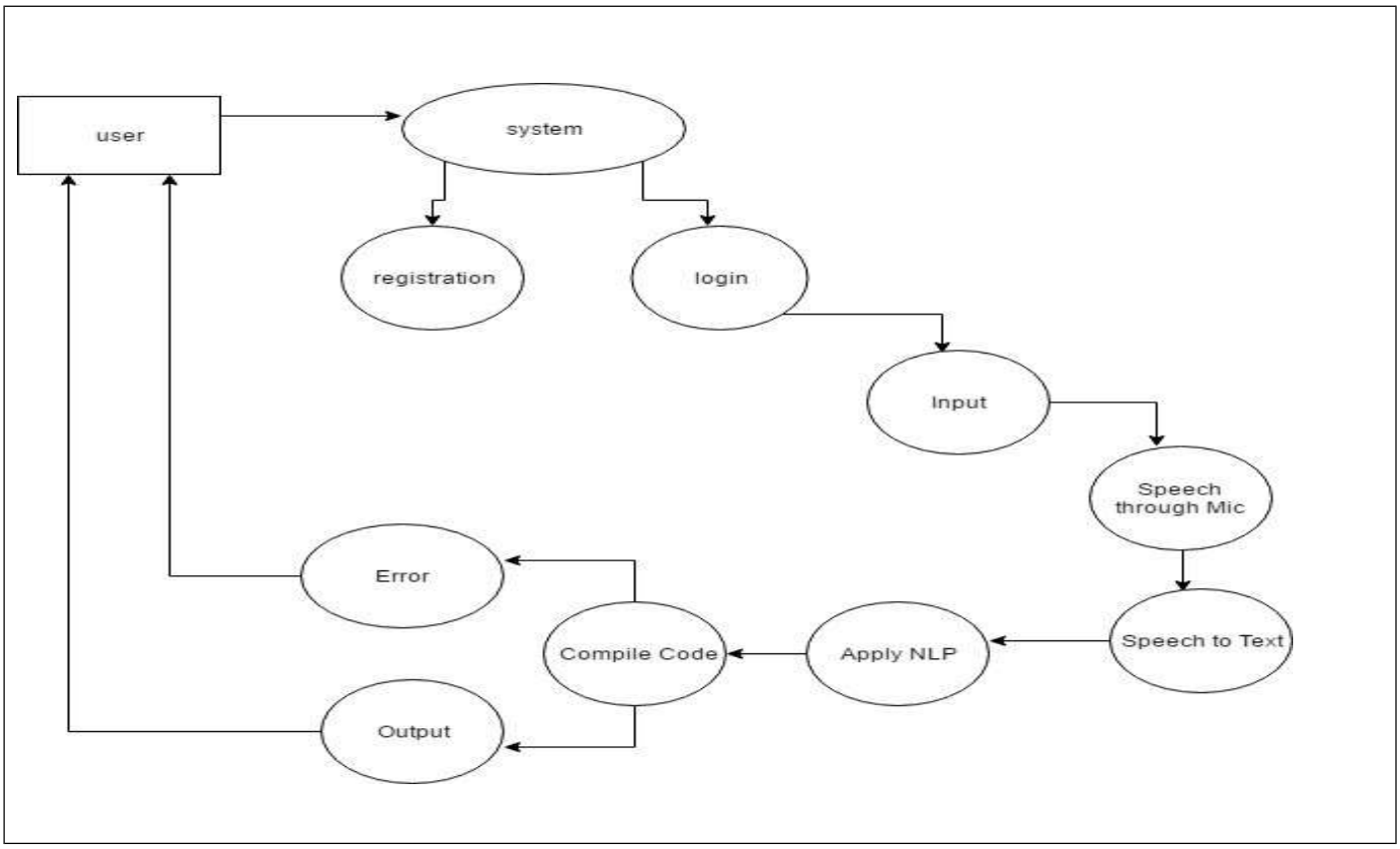

Fig.3: Internal and External Process of the system

V. CONCLUSION

The output of CTC model is usually peaky and this results in large number of blank frames during decoding. These blank frames do not change the linguistic search space and search at these frames are therefore redundant. In this paper by removing the blank frames from the linguistic search space, traditional frame synchronous decoding (FSD) can be transformed into phone synchronous decoding (PSD). PSD can be viewed as a hybrid decoding framework of beam search and A* search because of its self-adjusted decoding interval to remove tremendous search redundancy due to blank frames from CTC-trained model. With PSD, compact and precise phone-level CTC lattice can be produced. Two modular search approaches based CTC lattice are proposed for LVCSR and KWS tasks respectively.

VI. ACKNOWLEDEMENT

We would like to express our sincere gratitude towards our guide Prof. Manjusha Tatiya for her valuable guidance and supervision that helped us in our project work. She has always encouraged us to explore new concepts and pursue new research problems. I credit our project contribution to her. I take this opportunity to thank all those who are

directly or indirectly involved in this project. Without their active co-operation, it would not been possible to complete this paper on time.

Reference

[1] Zhehuai Chen, Yiemeng Zhuang, Yanmin Quan, "Phone Synchrounous Speech Recognition With CTC Lattices" IEEE Student Member May 2017.

[2] Nicolas Obin, Axel Roebel, "Similarity Search of acted voices for automatic voice casting", 2016 Vol 3.

[3] F. Jelink, L. R. Bahl and R. L. Mercer, "Design of a linguistic statistical decoder for the recognition of continuous speech", IEEE trans. Inf. Theory,vol 21,no. 3,pp. 250-256, May 1975.

[4] P.C Woodland,J.J. odell,V.Valtchev,and S.J.Young,"large vocabulary continuous speech recognition using HTK",in Proc.1994 IEEE int.conf Acoust.speech,signal process.,1994,Vol,2,pp.125-128.

[5] X.Liu, X. Chen, Y . Wang,M.J. Gales,and P.C Woodland,"Two efficient lattice rescoring methods using recurrent neural network language models,"IEEE/ACM trans.Audio,Speech,Lang,Process.vol. 24,no. 8, pp. 1438-1449,Aug. 2016.

[6] V. Vanhoucke,M.Devin,and G.Heigold, "Multiframe deep natural networks for acoustic modelling" in Proc. 2013 IEEE int.conf.Acoust,Speech Signal Process,2013, pp 7582-7585.

[7] Y.miao,j.Li,Y.Wang,S-X.Zhang,and Y.Gong,"Simplifying long short-team memory acoustic models for fast training and decoding."in Proc.2016 IEEE int.conf.Acoust,Speech Signal Process, 2016,pp. 2284-2288.

[8] Zahi N.Karam,William M.Campbell,Najim Dehak,"Graph Relational Features for Speaker Recogntion and Minning ,IEEE Aug2017.

[9] W.M.Campbell , D.E Sturim and D.A Reynolds, "Support Vector machines using GMM supervectors for speaker Verification."