Applied GIS

ISSN: 1832-5505
Vol-5 Issue-01 Jan 2017

Integrating CTC Speech Recognition Into Real-Time Integrated Development Environment

Prof. Manjushai Tatiya[1] Nikita Kangane[2] Aarti Korade[3] Gauri Jaydeokar[4] Juilli Kedari[5]

Computer Department, Savitribai Phule Pune University

***Abstract*-— Automatically identifying a spoken word and its speaker from the unique data contained in their voice waves is the goal of speech recognition technology. This method allows for the implementation of speech-based biometrics, database access services, voice-based calling, voice mail, and remote access to computers, among other applications. Due to its central role in human communication, voice has the potential to be an effective HCI tool. The development of effective speech recognition methods has become more important as the use of wireless networks continues to explode. Attaching speech apps would strengthen portable and wearable computer devices, which are required components for being computationally powerful, since voice may facilitate invisible connection with a computer device in a natural manner. A voice recognition system's feature set must first be extracted from the signal processing front end. Despite extensive investigation, the optimal feature set has not yet been determined. There is a wide variety of traits that may be obtained in various ways and have a positive effect on the accuracy of identification. For the purpose of voice recognition systems, this project demonstrates one method for extracting the feature set from a spoken signal.**

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

INTRODUCTION

Computers are becoming one of the most ubiquitous household items thanks to constant innovation in the field. However, as demand for more sophisticated and feature-rich devices grows, manufacturers are responding by decreasing the form factors of these gadgets without sacrificing functionality. There is now effectively no distinction between computers and PDAs, since many computers come equipped with a phone, a personal directory, a note capability, an alarm clock, a scheduler, a camera, games, and a variety of programs that were previously only available on PDAs.

While designers have been working hard to reduce the size of computers, this has come at the expense of the devices' ability to provide users with the full range of features they need. Their usefulness and use are severely constrained, hence scientists often seek out other routes of interaction. Human computer interaction (HCI) experts have spent the last several years trying to perfect voice-based interfaces that make it easier and faster to get things done. The function of speech in interfaces has been largely overlooked despite the fact that voice technology has been investigated for application in desktop computer and telephone information systems.

Historically, speech recognition has had abysmal accuracy, and its use in a system has been questioned due to room for ambiguity and mistake. However, speech technology has matured to the point where it can be used profitably, and it is currently being used by a variety of computer devices to improve user experiences. Voice interaction needs simply audio I/O devices such a microphone and speaker, both of which are currently relatively tiny and cheap, allowing the interface size to be scaled down. Since voice is the most fundamental means of human communication, voice interfaces on a computer device are sufficient to replace graphical user interfaces for accessing all information and content without the use of keywords, buttons, and touch screens.

OBJECTIVE

Goal 1: Develop a voice-activated technology for use by the physically challenged.

The second goal is to create electronics that can recognize human voices and react accordingly.

Third, to identify the speaker by examining the speech signal.

To accelerate the rate at which a processor can carry out a given job.

In order to compare the time required for regular access versus voice-based access, 5.

LITERATURE SURVEY

Temporal Connectionism: A Classification Scheme There are additional suggestions for lattices, economical and effective modular speech recognition techniques, second pass rescoring for big vocabulary continuous speech recognition, and keyword identification through phone [1]. For the purpose of voice casting, the suggested voice casting system [2] investigates acoustic models based on the Gaussian mixture model and multilabel identification of pre-received paralinguistic content. A very effective speaker comparison algorithm [3] has emerged as a consequence of recent developments in the area of

ISSN: 1832-5505
Vol-5 Issue-01 Jan 2017

speaker recognition. Applying a super vector machine classifier based on the GMM algorithm. Two new SVM kernels were developed [4] that use distance matrices to compare generalized linear mixed-models. Traditional approaches based on the usage of a keyboard and mouse do not prove adequate for current human system interaction. We must seek the simplest and most pleasant approach for human system interaction [5], especially when considering the impoverished and ill-qualified members of the Information Society. To accommodate the wide range of unplanned outcomes in human-to-human conversations, researchers have developed novel noise models that characterize both human and non-human noise, as well as fragments. It is shown that the identification ability is much improved by both auditory and linguistic modelling of the noise. Experiments include doing a clustering of the noise classes and comparing the resultant cluster variations in order to find the optimal balance between sensitivity and trainability of the models [6]. Here, we revisit the concept of using Hidden Markov Models (HMMs) for this purpose, detailing the improvements and optimizations made to a speech-based emotion recognizer.

working in tandem with IVR (automated voice recognition) [7]. In this study, we'll look at how to fix some problems that arise when comparing human and machine voice recognition. Experimental findings with humans, in particular, imply that the resultant inaccuracy is a function of the component streams [8]. An intriguing technology, emotional speech recognition can identify a speaker's mood just by analyzing their voice. Intelligent robots are being trained in emotion detection for use in Human-robot interaction (HRI) [9] so that they can respond to human emotions. Provides details such as laboratory layout and personnel, as well as departmental addresses and contact numbers [10].

I. PROPOSED AND IMPLEMENTED SYSTEMARCHITECTURE

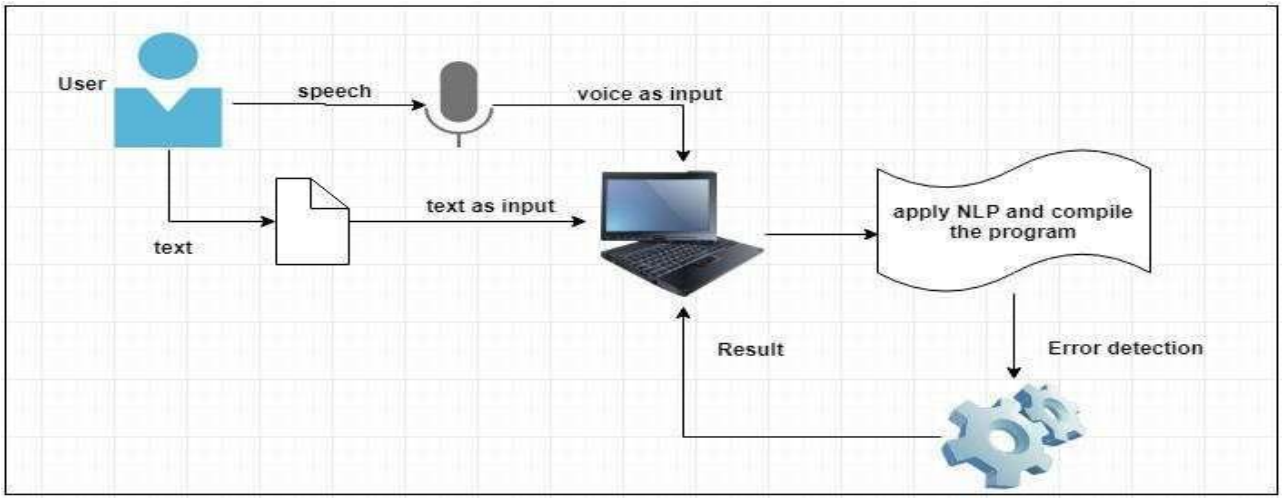

Fig.1: System Architecture

ISSN: 1832-5505
Vol-5 Issue-01 Jan 2017

For computers to be able to understand and respond to human speech, researchers in the interdisciplinary subject of computational linguistics have had to create new approaches and technologies known as speech recognition. Automatic speech recognition (ASR), computer speech recognition, and speech to text (STT) are all names for the same thing. It draws on studies and findings from the domains of languages, computing, and electrical engineering.

For certain voice recognition systems, "training" entails the user reading text or a set of isolated words into the program. The user's unique voice is analyzed by the system, and this information is then used to fine-tune the speech recognition for that user. A "speaker independent" system is one that does not need training, whereas a "system dependent" one is one that does. What is being spoken is less of a concern when using speech recognition or speaker identification. Systems that have been trained on a particular voice may employ speaker recognition to make the translation process easier, and it can also be used for security purposes to verify or validate the speaker's identity.

A system's architecture is its underlying conceptual model, outlining its physical make-up, behavioral characteristics, and other perspectives. A formal description and representation of a system, structured to facilitate reasoning about the system's underlying structures and behaviors, is known as an architectural description.

The user's identity and information must be established as a starting point in the system design. Then they are prepared to start using the system. After the first information entry, the user may choose between voice input and text input. If the user chooses to enter data by voice, the computer will transcribe the user's spoken words into text. The software will be finished when the input is provided and Natural Language Processing (NLP) is used. The program is compiled, and if any errors are found, the code is re-run through the input form and the compilation process is repeated until no more problems are found. The primary goal of the system is to accept voice as input, conduct natural language processing, and provide the results to the user. Converting spoken language into text and vice versa is the core functionality of the system.

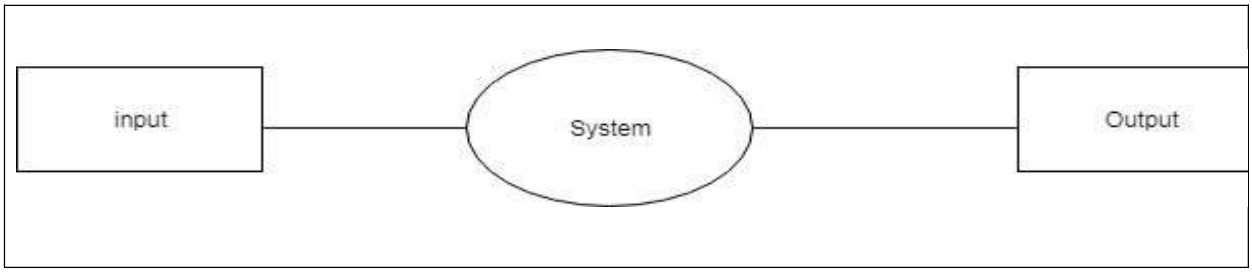

Fig2. Components of the System

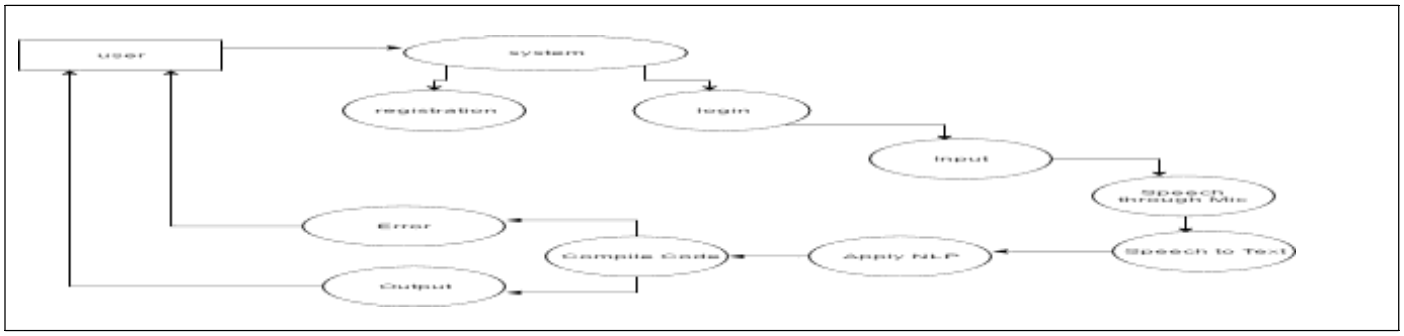

Fig.3: Internal and External Process of the system

CONCLUSION

The CTC model's output is rather peaky, therefore decoding often results in many blank frames. These blank spaces do not contribute to the searchable lexical domain and so should not be searched. By removing the silent frames from the linguistic search space, we demonstrate how to transform traditional frame synchronous decoding (FSD) into phone synchronous decoding (PSD). PSD may be seen as a hybrid decoding framework of beam search and A* search due to the fact that blank frames from the CTC-trained model are deleted automatically from the decoding period. Using PSD, a compact and granular CTC lattice may be generated at the phone scale. We describe two CTC lattice-based modular search techniques, one for the LVCSR task and the other for the

ISSN: 1832-5505
Vol-5 Issue-01 Jan 2017

KWS job.

ACKNOWLEDEMENT

Our greatest appreciation goes out to Prof. Manjusha Tatiya, who provided us with invaluable direction and oversight during the duration of our research. She has always inspired us to go beyond the box and pursue novel research questions. The contribution of our team is mostly due to her. Please accept my sincere appreciation for everyone who has helped in any way with this endeavor. Their helpful cooperation was crucial to meeting the deadline for this work.

Reference

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

[2] Similarity Search of Acted Voices for Automatic Voice Casting. Vol. 3, 2016, by Nicolas Obin and Axel Roebel.

[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] In 1994, at the IEEE international conference on acoustics, voice, and signal processing, P.C. Woodland, J.J. odell, V.Valtchev, and S.J.Young published "large vocabulary continuous speech recognition using HTK" (Vol. 2, pp. 125-128).

[5] "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; X.Liu, X. Chen, Y. Wang, M.J. Gales, and P.C. Woodland.

[6] According to [6] "Multiframe deep natural networks for acoustic modelling" by V. Vanhoucke, M. Devin, and G. Heigold in Proceedings of the 2013 IEEE International Conference on Acoustics and Speech Signal Processing, pages 7582–7585.

[7] In Proc.2016 IEEE int.conf.Acoust,Speech Signal Process, 2016,pp.2284-2288, Y.miao,j.Li,Y.Wang,S-X.Zhang,and Y.Gong write about simplifying lengthy short-team memory acoustic models for quick training and decoding.

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

ISSN: 1832-5505
Vol-5 Issue-01 Jan 2017

[9] Speaker verification using support vector machines and Gaussian mixture models," by W.M.Campbell, D.E. Sturim, and D.A. Reynolds [9].