Automatic Speech Recognition and its Applications

Pratiksha.C.Raut[1] , Seema.U.Deoghare[2]

[1]*PG Student [VLSI& Embedded],Dept. of ETC, PCCOE , Pune,Maharashtra, India*
[2]*Assistant professor, Dept. of ETC, PCCOE , Pune,Maharashtra India*

---***---

Abstract: *With the technology advancements in signal processing domain, speech recognition has over- come many challenges like speaker and language variability, vocabulary size, noise etc. Speech Recognition differs from Speaker recognition as it detects the person behind the speech signal and is language independent. Automatic Speech Recognition is a technology that allows human beings to use their voices to speak with a computer interface in a way that, resembles normal human conversation. It allows users of information systems to speak entries rather than punching numbers on a keypad. The ASR system can be applied to various applications like converstional IVR's, Dictations, Air Traffic Control, Automated Car Environment, Biomedical Applications, etc.*

Keywords: Automatic Speech Recognition, Air Traffic Management, Car Environment, Minimum Variance Distortionless Response

1.INTRODUCTION

In recent years, automatic speech recognition technology has advanced to the point where it is used by millions of individuals to automatically create documents from dictation. Speech Signal is created at the vocal cords which travels through the vocal tract and is produced at the speaker's mouth. It gets to the listener's ears as a pressure wave. Vowels and consonants form the two major classes of speech. The main difference between speech and audio signal is that speech is narrow band while audio is upto 20kHz. Speech can be quantized with 8 bit log quantizer while high quality audio requires 16 bit. Speech recognition is a computer science term and is also known as automatic speech recognition. It is a feature that turns speech into text. One of the main advantages for speech recognition services is the cut back on misspelled words that some typists may suffer from when typing. The service cuts down on the amount of time editing and fixing spelling corrections. It is also a big advantage to people who may suff.er from disabilities that affect their writing ability but can use their speech to create text on computers or other devices. The speech production mechanism in humans converts the words into speech. An ASR system aims to infer those original words given the observable signal.

1.1 RELATED WORK

In [1] , author has overviewed the car environment and conditions to be considered with ASR view. Author has described the in-route navigation system. The text data obtained from speech recognition system can be used for hands-free applications using smartphones etc. The author in [2] has provided a thorough overview of modern noise-robust techniques for ASR. The pros and cons of using noise robust ASR in practical applications and future research in this field. As ASR is applied to various applications, author gives an overview of types and phases of ASR. The speech recognition process alongwith the Hidden Neural Networks are explained[3]. Applying ASR to Air traffic management system is a challenging task. In [4], author describes characteristics of ASR process, problems in ATC and process of applying ASR system for ATC.

2.AUTOMATIC SPEECH RECOGNITION

An automatic speech recognition system works by pattern matching digitized audio of spoken words against computer models (i.e., computer representations) of speech patterns to generate a text transcription. Speech recognition or more commonly known as automatic speech recognition (ASR), is the process of interpreting human speech in a computer. An ideal automatic speech recognition system would process speech audio into an error-free, word-for-word transcription of that speech. However, due to a number of factors, including the huge variations in normal human speech, perfect, verbatim transcription is not currently feasible.Fig.1 shows block diagram of ASR system.

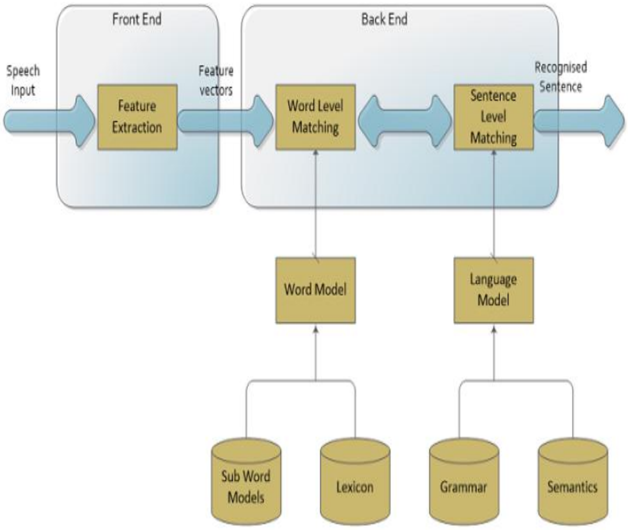

Fig 1: Block diagram of Automatic Speech Recognition System

To account for this reality, state-of-the-art speech recognition systems return one or more text strings, where each string is associated with a confidence level. The goal of an ASR system is to accurately recognise the words in speech spoken by any person in any environment, in the same way humans can. The front-end of an ASR converts speech signals into sequences of observation vectors. Feature extraction techniques include Mel-frequency cepstral coefficients (MFCC), linear prediction coefficients (LPC) and perceptual linear prediction (PLP). Both the training and classification phase use feature extraction. The second step is a combined word-level/sentence-level matching procedure. A set of word models is created by concatenating each of the sub-word models as specified by the word lexicon. A word grammar and semantics is used for the language model. The grammar is the syntax of the system and the semantics specify valid sentences in the task language. The word-level match procedure provides scores for individual words as specified by the sentence-level match procedure. The output is the sentence that provides the best match to the speech input. Applications where ASR is used, vary from simple tasks to more complex ones. Some examples are speech-to-text input, air traffic control, security and biometric identification, gaming, and home automation.

3. APPLICATION OF ASR IN AIR TRAFFIC MANAGEMENT

Automated speech recognition, the technological capability to translate speech to text, has advanced significantly over the past decade. Air Traffic Control (ATC) remains a domain that is ripe for the application of automatic speech recognition technology, particularly in the areas of safety and efficiency. Automatic speech recognition is a continuously improving technology that can be used to tap into this information source for potential system benefits in a variety of ATC applications, such as monitoring live operations for safety benefit, conducting analysis on large quantities of recorded controller-pilot speech, or enabling automated simulation pilots to facilitate training and Human-in-the-Loop (HITL) simulation experiments A significant portion of an air traffic controller's tasking involves voice communications, which represent a vital pipeline for exchanging information between controllers and pilots, as well as between controllers themselves, on the current and future movement of aircraft..

In the areas of ATM automation planning and prediction, controller-pilot voice communications are an untapped resource; instead, surveillance data and surveillance trends are heavily relied upon to deduce the air traffic situation. The common uses of ASR in ATC include the following: ASR of Controller utterances for simulator training and pseudo pilots, Offline transcription of simulated and field data for research and forensic purposes, Onboard ASR for Avionic Instruments, Controller workload estimations. The ATC Speech Recognition prototype developed works according to the block diagram in4.3.An initial preprocessing block performs done Low-pass filtering, restricting the information to twice the vocal frequency band [0-4 KHz], to remove high-frequency noise. Obtained signal is then segmented by removing silences, and individual ATC voice communications are extracted, so that small audio files are obtained, stored and labelled by a reference index. Several parameters in the system can be modified for better segmentation if needed as silence, pause or mumbling time. The speech recognizer module then provides a suggested transcription, based in an enhanced Speech recognizer, trained with around 300 transcript hours of real ATC communications.

Determination of language of the communication (English/Spanish) is also done here. The Speech Recognition is done by a multi-mode Hidden Markov Model (HMM), trained according to the syntax and structure of ATC voice communications. This speech recognition module is subdivided into two models which perform effective recognition as: Language Model: It defines the phraseology to be used (in Spanish and English), with syntax and grammar usual structure (i.e., callsign composition rules). Logical relationships, probabilistic modeling and language restrictions are included here, allowing the reuse of the model in different ATC centers or individuals, making the system almost speaker independent (in an order). Acoustic model: It takes into account physical characteristic of the speech as a sound, characterizing particularities of the speaker accent, speed, etc., as well as models the physical channel so that changes between channels in different centers are compensated and do not affect the speech recognition module. Figure.2 shows block diagram of ASR for ATC environment[5].

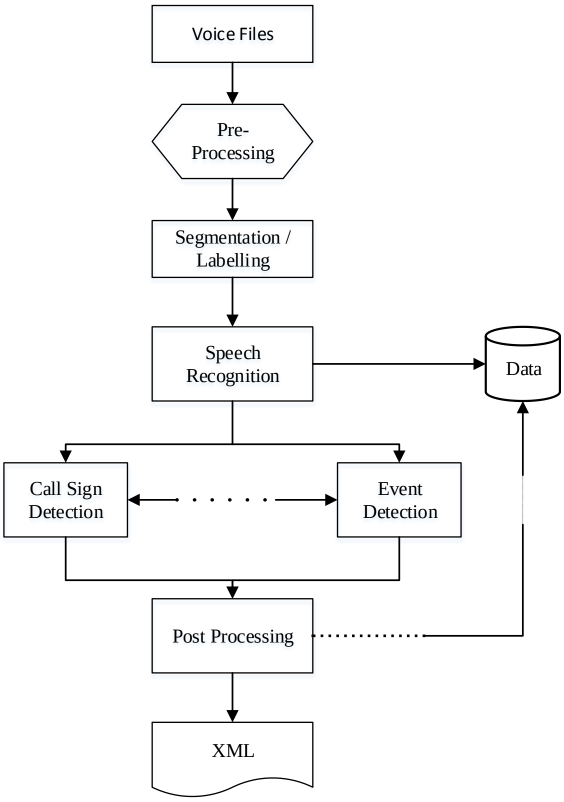

Fig 2 : Block Diagram for ASR for ATC

It is consequently necessary to update the acoustic model whenever using the prototype in new physical site. The speech recognition module outputs a transcription of the audio segment, and stores it in an internal database. The callsign detection module looks into the transcription for a determined segment and starts to look for possible callsign structures, which can be quite diverse.

After a callsign and a control event (if existing in the phrase) have been found, they are included in an output XML file that enters the post-processing module. Here, information is checked in order to fill the gaps that might have not been detected, callsigns specially, correlating them with other detected, trying to find relationships according to established typical behaviors of a light inside a sector. Additional filtering is done, checking if a same event for a same callsign in a short time is in fact a new event or a repetition (what has an effect on effective controller workload calculation). After this, a final output XML file is obtained for a sector in a period of time (typically one hour), that can be used for further analysis. Final results are also stored in a database, with the audio files.

4.APPLICATION OF ASR FOR CAR ENVIRONMENT

A number of studies have been carried out on automatic speech recognition (ASR) systems in a car environment. It has been widely applied to vehicle operation command systems, speech-controlled navigation systems and speech-controlled cabin systems. speech. In this study, we emulate the signal characteristics of primary user signals. The block diagram consists of formulation of a new microphone array and multi-channel noise suppression front-end, environmental (sniffer) classification for changing in-vehicle noise conditions, and a back-end navigation information retrieval task. Figure 3 shows block diagram of speech recognition system for car environment.[1]

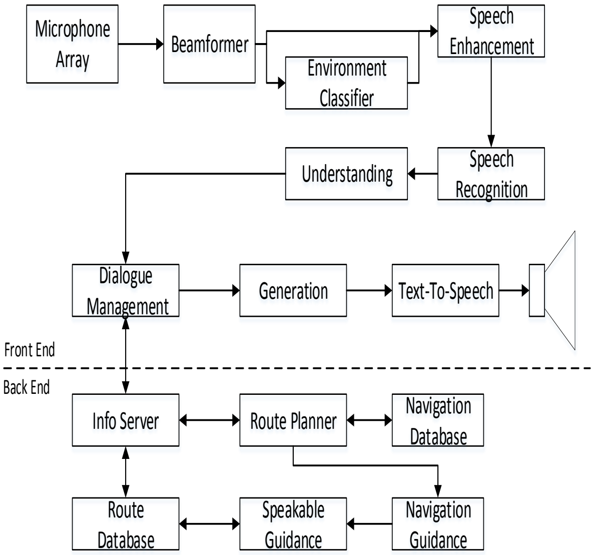

Fig 3: Block diagram of speech recognition system in car environment

If a speech source is detected, the beamformer is used to adjust the beam pattern and enhance the desired speech. Environmental Sniffing is to detect, classify and track acoustic environmental conditions in the car environment. The first goal of the framework is to seek out detailed information about the environmental characteristics instead of just detecting environmental changes. The second goal is to organize this knowledge in an effective manner to allow smart decisions to direct other speech systems. Speech recognition includes robust acoustic feature representation and built-in speaker normalization. Capturing the vocal tract transfer function (VTTF) from the speech signal while eliminating other extraneous information, such as speaker dependent characteristics and pitch harmonics, is a key requirement for robust and accurate speech recognition. The vocal tract transfer function is mainly encoded in the short-term spectral envelope.

Traditional MFCCs use the gross spectrum obtained as the output of a non-linearly spaced filterbank to represent the spectral envelope. MFCCs are known to be fragile in noisy conditions, requiring additional compensation for acceptable performance in realistic environments. Minimum Variance Distortionless Response (MVDR) spectrum has a long history in signal processing but recently applied successfully to speech modeling. It has many desired characteristics for a spectral envelope estimation method, most important being the fact it estimates the spectral powers accurately at the perceptually important harmonics, thereby providing an upper envelope which has strong implications for robustness in additive noise. Since the upper envelope relies on the high-energy portions of the spectrum, it will not be affected substantially by additive noise.

5.CONCLUSION

Automatic Speech Recognition Systems provide the ability to convert speech into well-understandable words. Due to its ability of real time speech conversion, it application to Air traffic control and Automated car environment has been studied. The ASR system for Air traffic control uses the Hidden Markov model in feature extraction while its phraseology is based on the commands used in air applications. In the car environment, speech recognition is done used for route navigation application.

REFERENCES

[1] John H.L. Hansen, Xianxian Zhang, Murat Akbacak, Umit H. Yapanel, Bryan Pellom, Wayne Ward, Pongtep Angkititrakul, Chuan-Kai Yang, " DSP for In-Vehicle and Mobile Systems, Springer Publication ", IEEE Global Conference on Consumer Electronics, 2015.

[2] Jinyu Li, Li Deng, Yifan Gong, R. Haeb-Umbach,"An Overview of Noise-Robust Automatic Speech Recognition", IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014.

[3] Anchal Katyal, Amanpreet Kaur, Jasmeen Gill, "Automatic Speech Recognition : A Review", International Journal of Engineering and Advanced Technology, 2014

[4] Hunter D. Kopald, Ari Chanen, Ph.D., Shuo Chen, Elida C. Smith, and Robert M. Tarakan,"Applying Automatic Speech Recognition Technology to Air Traffic Management.", IEEE Conference of Digital Avionics Systems (DASC), 2013.

[5] Jose Manuel Cordero, Manuel Dorado, Jose Miguel de Pablo", Automated Speech Recognition in ATC Environment", International Conference on Application and Theory of Automation in Command and Control Systems,2012 .