Process Speech Recognition System using Artificial Intelligence Technique

Anupam Choudhary, Ravi Kshirsagar

Abstract: This paper describes the detail process of speech recognition using artificial intelligence technique. It includes coustic model, Language model,Trigram model, Class model ,Source channel model .Speech recognition or natural language processing referred to artificial intelligence methods of communicating with a computer in natural language like English. The objective of NLP Program is to understand the IP and initiate the action. Method gives theoretical conceptual view to process the speech recognition, a acoustic model need sto be able to interface with telephony system because there is no GUI it needs to manage a spoken dialogue with user.

Keywords: NLP,GUI,IP,channel model

I. INTRODUCTION

Process speech recognition using artificial intelligence technique using natural languge processing uses IP words. The IP words are scanned and matched against internally stored known words. The identification of a keyword causes some action to be taken. Thus speech recognition allows you to provide IP to an application with your voice and in your own language no need to enter the program in special language for crating software[4][2]. You can provide the I/p to an application with your voice just like clicking the mouse & typing your keyboard. The acoustic signal captured by microphone is converted to a set of words and recognized words are the first results of the application like command and control, data entry, document preparation this is you can give the commands like switch to calculator, open a notepad by voice instead of using mouse or keyboard. Thus you can write your mail while driving a car by dictating to computer. Many speech recognition systems are used Isolated word speech recognition systems require that speaker must paused briefly between the words while continuous speech recognized system does not spontaneous speech may have disfluesncies and is difficult to recognize[1]. Some speech recognizes systems require speech enrollment that is user has to provide sample of his/her speech before using the system while some systems are speaker independent. That system does not require the sample of the speech .Generally speech recog. Is difficult when vocabulary is large or the system has many similar sounding words. When speech is generated as a sequence of words artificial grammar or language models are used to restrict the combination of words. The vocal tract system including coupling of nasal tract accurately described by the position of articulator like tounge, jaws etc. The vocal tract system is described by the acoustic features like frequency response resonance and antiresonance. The vocal tract system is excited by voice source, unvoiced source and plosive source[6]

Voice source: The voice source is produced when vocal cards vibrate open& close.

The rate of opening and closing and air pressure determines pitch period. The voices source show long term periodicity at pitch period. Voice source includes vowels and no. of consonants.

Unvoiced source: Unvoiced source is due to turbulence flow at air it shows little large term periodicity at pitch period ex. F, S, SH Plosive source: Closure of vocal card and air pressure is built behind and suddenly released. Some sounds do not fall in above categories. They are mixtures many possible but as the shape of vocal tract system and mode of excitation changes relatively slowly due to their quasi stationary nature. They show a high degree of predictability[3][4]. This high degree of predictability is used by speech codes for producing high degree of voice quality.The lowest frequency of oscillation is called pitch frequency.

II. DIGRAMATIC REPRESENTATION OF SPEECH RECOGNITION

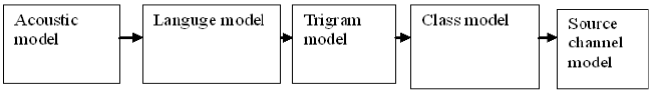

Fig. 1 Speech recognition process

1. Acoustic model represents the acoustic sounds of a language and can be trained to recognize the char of a particular user's speech patter and acoustic environment. Lexical model gives a list of large no. of words in a language along with how to pronounce each word.
2. Language model gives the way in which different words of a language are combined.
   In order to recognized a word the recognizer chooses it is guess from a finite vocabulary as the word is uniquely identified by it is spelling different models are used for this purpose.
3. Trigram model: In a trigram model the concept is the probability of next word depends upon the history of previous words that have been spoken i.e.

Probability a next word w depends upon his (previous) Where I = 1 + 0n but as number of previously spoken words increase the complexity of model increase in order to have a practical model trigram model i.e. where n = 3 is used it means that most recent only two words at the history are used to obtain the condition probability of next word.

Manuscript Received on November 21, 2012.

Anupam Choudhary, Assistant Professor, Gurunanak College of Institute of Engineering and Technology, Nagpur. India.

Ravi Kshirsagar, Vice-Principal & Professor,Priydarshini College of Engineering & Technology, Nagpur,India.

The term perplexity is used to determine the performance perplexity is defined as size of set of the words from which the next words is choose to we use the history at previously spoken words for 5 diagram model the perplexity in difference domain is

Domain	Perplexity
Domain Radiology	26
Emergency medicine	60
Journalism	105
General English	247

When two language model are given one need to compare them one method used to model in recognize and select the one which provide minimum recognizer error rate

Or we can determine the best language method by talking long probability per word basis on a new text which is not being used for building the language model which will give you perplexity.

Again show diagram major components the digital speech signal is first transformed

To a set measurement or future at fixed rates typically at 10-20 msec & used to search the most likely words condition depending upon the contains imposed by lexical, language & acoustic model. Throughout this process training data is used to determine the valued of modeled parameters.

4. Class model: Instant at using separate words set at words i.e. Class is used There may be overlapping i.e. one word may belong to many classes the classes are made deepening upon morphological analysis it words & segmentation information at the word.

5.Source Channel Model: this model is pioneered by IBM continues speech recognition group it use statistical model at joint distribution p (w, o) w- Sequence of spoken words & 0 is the corresponding sequence of observed acoustic information[7][1] .It determine estimation W identity of spoken words from observed acoustic in order to minimize the error rat the recognize choose that sequence with max posterior distribution.

III. ACOUSTIC MODELING IN SPEECH RECOGNITION SYSTEM

At the level of signal representation the researcher have developed representation & emphasize of the perceptually imp speaker independent features & deemphasize speaker independent char. At acoustic phonetic level the speaker variability is modeled by different adoption algorithm that will adopt speakers' independent system the speaker variability is handled by statistical modeling which will operate on large no. of data the effect of linguistics context on phoneme at acoustic phonetic level is handled by training separate model for phoneme in different acoustic modeling.The word level variability is modeled by different pronunciation n/w which will handled common pronunciation of words through speech algorithm different statically technique are used to find most probable words sequence depending upon the frequency of accuracy of the words.The most dominate model used HMM which is statically given by the rule of probability mode in which underling phoneme an d frame by frame acoustic realization is probability representative as mark or process. The speech segment are identify during the search process rather than identity explicitly alternative approach is to the first find speech segment then classify speech according to segment score recognize words[4][3] . This approached produced competitive approach in several task in the speech segments and modeled according there means, variance and shape it reduces error rate up to 34 %. Different technologies are appropriate for different task. When vocabulary is mall the word can be considered as single unit search as approached is not appropriate when vocabulary are large in search case words must be modeled by sub word units.

IV. FACTOR AFFECTING THE PERFORMANCE OF SPEECH RECOGNIZATION SYSTEM:

There are many external factors that affect the performance of speech recognization system like noise environment condition, placement of IPPhone etc.

Parameter	Range
Speaking mode	Isolated to continuous speech
Speaking style	Read speech or spontaneous speech
Transducer	IPPhone to telephone
Enrollment	Speaker dependent & speaker independent
Vocabulary	small < 20 large > 20000 words
Perplexity	small < 10 large > 100
Language Model	finite state to context dependent

Vocabulary is the most dominant feature which affects the performance of the speech recognition system as the error rate is recognizer is no less that the % of spoken words that are not in recognizer vocabulary therefore building a language model vocabulary is the most important factor for determining vocabulary corpus (collection) of text along with directories is used. When the recognition is restricted for a particular application then more personalized vocabulary is useful rather than general vocabulary.

Table shows the static coverage of unseen text depending upon.

V Size	Static coverage
20,000	94.1%
64000	98.7%
100000	99.3%
200000	99.4%

The next imp parameter is the acoustic representation of the phonemes i.e. the smallest sound units from which word are composed of this phonemes are highly dependent on the context in which they are used in for example the acoustic representation of phoneme /t/ in 'true' & 'two'.Another factor that affects the performance of speech recognition system is quality & placement of microphone, speakers, emotional & physical condition, speech rate, voice quality, size & shape of vocal tract system. So it is very difficult to specify how speech sound like, moreover human speech rarely follows strict & formal grammar rules & word cannot be said exactly in the same way twice therefore speech recognition is never going to find it's match but the quality of recognition is depends upon how good it is refining it's search i.e. eliminated poor matches & selecting most likely matches[7].

The accuracy of the recognition depends upon good language & acoustic model as well as on algorithms for both search & sound processing better is the models & algorithms fewer errors are made & result are quickly found.When general language model is used it will have comprehensive language domain i.e. consist of general day to day spoken English, but if recognition is to be used for the particular application then instead of using general language model.It is beneficial to use the model in which only restricted words that are required for the application are used. It has several benefits it increases accuracy, few errors are made quicker search and each search result is meaningful as due to restricted Vocal the recognizer will listen only that speech which is.

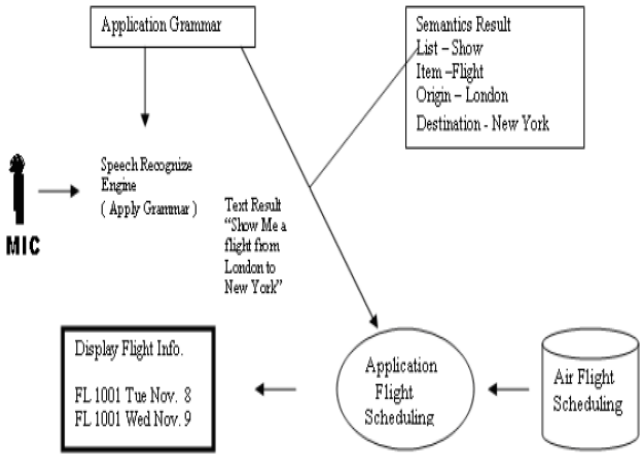

Fig.2. Process the I/P audio stream

Required for that application. Using speech recognition for application I/P (using windows vista). The speech recognition system can be said as consisting at a front end and at back end. The front end process the I/P audio stream, isolated sound segments i.e. probably speech into series of numeric values that are characterized by vocal sounds at the signal. The back end is a search engine which searches through 3 databases 1) Acoustic 2) lexicon 3) language model (Explain diagram) Windows Vista speech technology has built-in dictation capability. It has edit controls for deleting, inserting. You can correct disorganized words by redictating like for New your spell N like Nest e-Element etc, choosing alternatives.

Microsoft speech server 2004R2 is also used where as MSS 2004 only supported English for automated speech recognition and text to speech generation. R2 adds French and Spanish recognition and generation.There are tremendous forces driving the development technology in many countries touch tone penetration are low and voice is the only option for controlling automated services. Another application is home voice. It uses latest voice recognition to give you control at your home. It uses sound blaster compatible sound card and your computer to add voice recognisation capabilities too many popular home automation controller you have to only select command phrases. You want to use and associate them with the action you want to perform. The action may be infrared command, macro or even relay closure or X10 command. X10 is a communication language that allows computable products to talk to each other using the existing electrical wiring in the home. Installation is simple and requires a transmitter plug in at one location and sets its control signal like on, off, dim etc to receiver into another location t home There are many methods to build language grammar simplest method is semantic grammar.

List --- Show me, I want, Can I see

For all this options it will select semantic word list. As user can ask question can I see a flight from……to…..or I can…….or showme……

Departure time – after, before, morning, afternoon, evening.

Hour – one, two, three

Item—a flight, flights

Origin – from city

Destination – to city

City—Boston, Santorin, New, London

Date & day – 1-2-2008……..Sun-Mon

Semantics for a sentence

List Item Origin

Show me Flight from Boston

Destination Date & day

To New Tuesday 19

Show me flight from London to newyork on Tuesday 19 at 10PM.

V. FUTURE DIRECTIONS

There are many challenges in the area at human language technology although there have been significant recent gains in spoken language understanding, current technology is far from human like. Only system in limited domains can be envisioned in the near term and portability at existing tech is rather limited there are many challenges like

Robustness – In a robust system performance degrades gracefully rather than catastrophically as conditions become more different from those under which it is trained.

Portability – Portability to goal of rapidly designing, developing and deploying system for new application. At present system tend to suffer significant degradation when moved to a new task.

Adaption – How can system adapt to changing conditions (new speakers, phone, tasks etc) & improve through use? Such adaption can occur at many levels in systems, sub word models, pronunciation, language models etc.

Language Modeling – Current systems use a satisfied language models to help reduce the search space and resolve acoustic ambiguity. As vocabulary size grows and other constraints are relaxed to create more habitation system it will be increasingly imp to set as much constraint as possible from language models.

Perhaps incorporating syntactic and semantic constraints that cannot be captured by purely statistical models.

Confidence Measures-- Most speech recognition systems assign scores to hypotheses for the purpose the range ordering them. These scores do not provide a good identification of whether a hypothesis is correct or not just that it is better than other hypotheses.Out of vocabulary words: System are designed for use with a particular set of words but system users may not know exactly which words are in the system vocabulary. This leads to a certain percentage of out of vocabulary words in natural conditions[4]. System must have some methods of detecting such out of vocabulary words, or they will end up mapping a word from the vocabulary onto the unknown word causing an error.

Spontaneous speech -- Systems that are deployed for real use deal with a variety of spontaneous speech phenomena such as filled pause, hesitation ungrammatical construction so development in this area is require. Prosody-- Prosody reference to acoustic structure that extends over several segments of the word stress, irritation and rhythm convey important information for word recognition and the user's intentions (e.g. anger). Current system does not capture prosodic structure. Modeling dynamics-- System assume a sequence of I/p frames which are treated as if they were independent, but it is known that perceptual clues for words and phenomena require the integration of features that reflect the movement of articulators which are dynamic in nature and incorporate this information into recognition systems is an unsolved problem.Prosody can be defined as the information that cannot be localized by a specific sound segment ass lexical stress; rythms conveys imp information about speaker line anger speaker's intention current system does not capture this structure.

VI. CONCLUSION

Speech recognition performance systems are now being deployed within telephone and cellular network. Within next few years speech recognisation will be pervasive in telephone network around the world telephone needs completely different acoustic model it need to be able to interface with telephony system because there is no GUI it needs to manage a spoken dialogue with user.

REFERENCES

1. "Hidden Markov models for speech recognition; X.D. Huang, Y. Ariki, M.A. Jack. Recognition", The Complete Practical Reference Guide; T. Schalk, P. J. Foster: Telecom Library Inc, New York; ISBN O-9366648-39-2.
2. "Automatic speech recognition: the development of the SPHINX system",Kai-Fu Lee; Boston; London: Kluwer Academic, c1989
3. "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition", S. E. Levinson, L. R. Rabiner and M. M. Sondhi; in Bell Syst. Tech. Jnl. v62(4), pp1035--1074, April 1983
4. "Review of Neural Networks for Speech Recognition, R. P. Lippmann; in Neural Computation", v1(1), pp 1-38, 1989.
5. "Automatic Speech and Speaker Recognition: Advanced Topics", C.H. Lee, F.K. Soong and K.K. Paliwal (Eds.), Kluwer, Boston, 1996.
6. 6. "Fiction database for emotion detection in abnormal situations." Clavel, C., Vasilescu, I., Devillers, L., Ehrette, T.,In: Proc. Int. Conf. Spoken Language Process.(ICSLP '04). Korea, pp. 2277–2280, 2005.
7. Modifications of phonetic labial targets in emotive speech: effects of the co-production of speech and emotions. Speech Communication, Caldognetto, E. M., Cosi, P., Drioli, C., Tisato, G., Cavicchio, pp173–185, 2003.
8. "You stupid tin box- children interacting with the AIBO robot: A crosslinguistic emotional speech " Batliner, A., Hacker, C., Steidl, S., N¨oth, E., D' Archy,S., Russell, M., Wong, M., In: Proc. Language Resources and Evaluation (LREC '04). Lisbon, 2004.