Volume 2, Issue 10, October 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Speech Recognition as Emerging Revolutionary Technology

Parwinder pal Singh
Computer science &Engg.
IGCE, PTU Kapurthala

Er. Bhupinder singh
Computer science &Engg
IGCE, PTU Kapurthala

Abstract— ***Speech recognition is the new emerging technology in the field of computer and artificial intelligence. It has changed the way we communicate with computer and other intelligent devices of same calibre like smart phones. It is a major area of interest for research in this field which is related to artificial intelligence. In this paper the overview of this technology and its current implementations are listed and introduced.***

Keywords— ***Speech recognition, Phonetics, Acoustic, Utterance, DTW.***

I. INTRODUCTION

Speech recognition is the translation of spoken words into text. It is also known as "automatic speech recognition", "ASR", "computer speech recognition", "speech to text", or just "STT". Speech Recognition is technology that can translate spoken words into text. Some SR systems use "training" where an individual speaker reads sections of text into the SR system. These systems analyze the person's specific voice and use it to fine tune the recognition of that person's speech, resulting in more accurate transcription.

II. TYPES OF SPEECH RECOGNITION SYSTEMS

A. *Speaker dependant*- A number of voice recognition systems are available on the market. The most powerful can recognize thousands of words. However, they generally require an extended training session during which the computer system becomes accustomed to a particular voice and accent. Such systems are said to be speaker dependent. A speaker dependent system is developed to operate for a single speaker. These systems are usually easier to develop, cheaper to buy and more accurate, but not as flexible as speaker adaptive or speaker independent systems. Speaker–dependent software works by learning the unique characteristics of a single person's voice, in a way similar to voice recognition. New users must first "train" the software by speaking to it, so the computer can analyze how the person talks. This often means users have to read a few pages of text to the computer before they can use the speech recognition software

B. *Speaker independent* - A speaker independent system is developed to operate for any speaker of a particular type (e.g. American English). These systems are the most difficult to develop, most expensive and accuracy is lower than speaker dependent systems. However, they are more flexible. Speaker–independent software is designed to recognize anyone's voice, so no training is involved. This means it is the only real option for applications such as interactive voice response systems — where businesses can't ask callers to read pages of text before using the system. The downside is that speaker–independent software is generally less accurate than speaker–dependent software.

C. *Speaker adaptive* - A third variation of speaker models is now emerging, called speaker adaptive. Speaker adaptive systems usually begin with a speaker independent model and adjust these models more closely to each individual during a brief training period.

III. SPEECH RECOGNITION SYSTEMS DISTINGUISHED ACCORDING TO THE INPUTS

A. *IWR: Isolated word recognition* - Isolated word recognizers usually require each utterance to have quiet (lack of an audio signal) on BOTH sides of the sample window. It doesn't mean that it accepts single words, but does require a single utterance at a time. Often, these systems have "Listen/Not−Listen" states, where they require the speaker to wait between utterances (usually doing processing during the pauses).Isolated Utterance might be a better name for this class.

B. *CWR: Connected word recognition* - Connect word systems (or more correctly 'connected utterances') are similar to Isolated words, but allow separate utterances to be 'run−together' with a minimal pause between them.

C. *CSR: Continuous speech recognition* - Continuous recognition is the next step. Recognizers with continuous speech capabilities are some of the most difficult to create because they must utilize special methods to determine utterance boundaries. Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content. Basically, it's computer dictation.

D. *SSR: Spontaneous speech recognition* - There appears to be a variety of definitions for what spontaneous speech actually is. At a basic level, it can be thought of as speech that is natural sounding and not rehearsed. An ASR system with spontaneous speech ability should be able to handle a variety of natural speech features such as words being run together, "ums" and "ahs", and even slight stutters.

E. *Voice Verification/Identification* – Some ASR systems have the ability to identify specific users. This document doesn't cover verification or security systems.[4]

IV. WORKING OF SPEECH RECOGNITION SYSTEM

Basically, the microphone converts the voice to an analog signal. This is processed by the sound card in the computer, which takes the signal to the digital stage. Input from user is also known as utterance (Spoken input from the user of a speech application. An utterance may be a single word, an entire phrase, a sentence, or even several sentences.)[3]This is the binary form of "1s" and "0s" that make up computer programming languages. Computers don't "hear" sounds in any other way.

Sound-recognition software has acoustic models (An acoustic model is created by taking audio recordings of speech, and their text transcriptions, and using software to create statistical representations of the sounds that make up each word. It is used by a speech recognition engine to recognize speech) [1] convert the voice sounds to one of about four dozen basic speech elements (called phonemes). The latest versions of speech technology have been refined so that they eliminate the noise and useless information that is not needed to let the computer work. The words we speak are transformed into digital forms of the basic speech elements (phonemes).

Once this is complete, a second sector of the software begins to work. The language is compared to the digital "dictionary" that is stored in computer memory. This is a large collection of words, usually more than 100,000. When it finds a match based on the digital form it displays the words on the screen. This is the basic process for all speech recognition systems and software. [2]

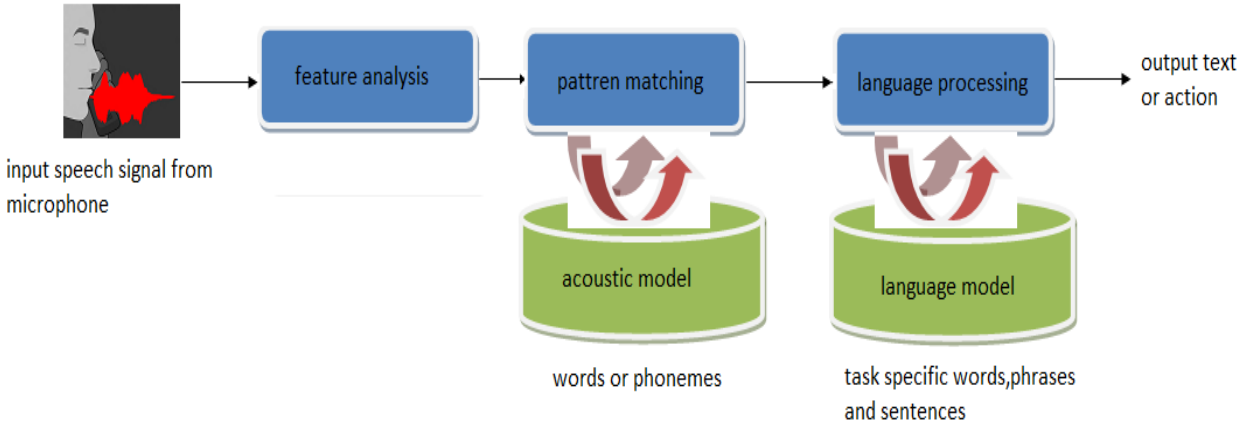

Fig. 1 Flowchart of Simple Speech Recognition System

V. ALGORITHMS USED FOR SPEECH RECOGNITION SYSTEM

Both acoustic modelling and language modelling are important parts of modern statistically-based speech recognition algorithms. Hidden Markov models (HMMs) are widely used in many systems. Language modelling has many other applications such as smart keyboard and document classification.

Hidden Markov models- Modern general-purpose speech recognition systems are based on Hidden Markov Models. These are statistical models that output a sequence of symbols or quantities. HMMs are used in speech recognition because a speech signal can be viewed as a piecewise stationary signal or a short-time stationary signal. In a short time-scales (e.g., 10 milliseconds), speech can be approximated as a stationary process. Speech can be thought of as a Markov model for much stochastic purposes.Another reason why HMMs are popular is because they can be trained automatically and are simple and computationally feasible to use. In speech recognition, the hidden Markov model would output a sequence of n-dimensional real-valued vectors (with n being a small integer, such as 10), outputting one of these every

10 milliseconds. The vectors would consist of cepstral coefficients, which are obtained by taking a Fourier transform of a short time window of speech and decorrelating the spectrum using a cosine transform, then taking the first (most significant) coefficients. The hidden Markov model will tend to have in each state a statistical distribution that is a Fig 1 shows the general speech recognition system. Each word, or (for more general speech recognition systems), mixture of diagonal covariance Gaussians, which will give likelihood for each observed vector. Each phoneme will have a different output distribution; a hidden Markov model for a sequence of words or phonemes is made by concatenating the individual trained hidden Markov models for the separate words and phonemes. Described above are the core elements of the most common, HMM-based approach to speech recognition. Modern speech recognition systems use various combinations of a number of standard techniques in order to improve results over the basic approach described above. A typical large-vocabulary system would need context dependency for the phonemes (so phonemes with different left and right context have different realizations as HMM states); it would use cepstral normalization to normalize for different speaker and recording conditions; for further speaker normalization it might use vocal tract length normalization (VTLN) for male-female normalization and maximum likelihood linear regression (MLLR) for more general speaker adaptation. The features would have so-called delta and delta-delta coefficients to capture speech dynamics and in addition might use heteroscedastic linear discriminant analysis (HLDA); or might skip the delta and delta-delta coefficients and use splicing and an LDA-based projection followed perhaps by heteroscedastic linear discriminant analysis or a global semitied covariance transform (also known as maximum likelihood linear transform, or MLLT). Many systems use so-called discriminative training techniques that dispense with a purely statistical approach to HMM parameter estimation and instead optimize some classification-related measure of the training data. Examples are maximum mutual information (MMI), minimum classification error (MCE) and minimum phone error (MPE).Decoding of the speech (the term for what happens when the system is presented with a new utterance and must compute the most likely source sentence) would probably use the Viterbi algorithm to find the best path, and here there is a choice between dynamically creating a combination hidden Markov model, which includes both the acoustic and language model information, and combining it statically beforehand (the finite state transducer, or FST, approach).A possible improvement to decoding is to keep a set of good candidates instead of just keeping the best candidate, and to use a better scoring function (rescoring) to rate these good candidates so that we may pick the best one according to this refined score. The set of candidates can be kept either as a list (the N-best list approach) or as a subset of the models (a lattice). Rescoring is usually done by trying to minimize the Bayes risk (or an approximation thereof): Instead of taking the source sentence with maximal probability, we try to take the sentence that minimizes the expectancy of a given loss function with regards to all possible transcriptions (i.e., we take the sentence that minimizes the average distance to other possible sentences weighted by their estimated probability). The loss function is usually the Levenshtein distance, though it can be different distances for specific tasks; the set of possible transcriptions is, of course, pruned to maintain tractability. Efficient algorithms have been devised to rescore lattices represented as weighted finite state transducers with edit distances represented themselves as a finite state transducer verifying certain assumptions.

Dynamic time warping (DTW)-based speech recognition- Dynamic time warping is an approach that was historically used for speech recognition but has now largely been displaced by the more successful HMM-based approach. Dynamic time warping is an algorithm for measuring similarity between two sequences that may vary in time or speed. For instance, similarities in walking patterns would be detected, even if in one video the person was walking slowly and if in another he or she were walking more quickly, or even if there were accelerations and decelerations during the course of one observation. DTW has been applied to video, audio, and graphics – indeed, any data that can be turned into a linear representation can be analysed with DTW.A well-known application has been automatic speech recognition, to cope with different speaking speeds. In general, it is a method that allows a computer to find an optimal match between two given sequences (e.g., time series) with certain restrictions. That is, the sequences are "warped" non-linearly to match each other. This sequence alignment method is often used in the context of hidden Markov models.

Neural networks-Neural networks emerged as an attractive acoustic modelling approach in ASR in the late 1980s. Since then, neural networks have been used in many aspects of speech recognition such as phoneme classification, isolated word recognition, and speaker adaptation. In contrast to HMMs, neural networks make no assumptions about feature statistical properties and have several qualities making them attractive recognition models for speech recognition. When used to estimate the probabilities of a speech feature segment, neural networks allow discriminative training in a natural and efficient manner. Few assumptions on the statistics of input features are made with neural networks. However, in spite of their effectiveness in classifying short-time units such as individual phones and isolated words, neural networks are rarely successful for continuous recognition tasks, largely because of their lack of ability to model temporal dependencies. Thus, one alternative approach is to use neural networks as a pre-processing e.g. feature transformation, dimensionality reduction, for the HMM based recognition. [1]

VI. CURRENT AND FUTURE USES OF SPEECH RECOGNITION SYSTEM

Currently speech recognition is used in many fields like Voice Recognition System for the Visually Impaired [5] highlights the Mg Sys Visi system that has the capability of access to World Wide Web by browsing in the Internet, checking, sending and receiving email, searching in the Internet, and listening to the content of the search only by giving a voice command to the system. In addition, the system is built with a translator that has the functionality to convert html codes to voice; voice to Braille and then to text again. This system comprises of five modules namely: Automatic Speech Recognition (ASR), Text-to- Speech (TTS), Search engine, Print (Text-Braille) and Translator (Text-to-Braille and

Braille-to -Text) module, was originally designed and developed for the visually impaired learners, can be used for other users of specially needs like the elderly, and the physically impaired learners. Speech Recognition in Radiology Information System. The Radiology report is the fundamental means by which radiologists communicate with clinicians and patients. The traditional method of generating reports is time consuming and expensive. Recent advances in computer hardware and software technology have improved Speech Recognition systems used for radiology reporting. [6] Integration of Robust Voice Recognition and Navigation System on Mobile Robot [7] and there are many other fields in which speech recognition can be used.

VII. CONCLUSIONS

This paper introduces the basics of speech recognition technology and also highlights the difference between different speech recognition systems. In this paper the most common algorithms which are used to do speech recognition are also discussed along with the current and its future use.

ACKNOWLEDGMENT

The author expresses appreciation to Er. Bhupinder Singh for his extensive support.

REFERENCES

[1] en.wikipedia.org/wiki/Acoustic_model
[2] www.thegeminigeek.com/how-speech-recognition-works.
[3] www.lumenvox.com/resources/tips/tipsGlossary.aspx.
[4] Lawrence R. Rabiner,AT&T Labs Florham Park, New Jersey07932,APPLICATIONS OF SPEECH RECOGNITION IN THE AREA OF TELECOMMUNICATI0NS,1997 IEEE.
[5] Halimah B.Z.Dep. of Info. Science,UKM, Selangor, Malaysia.hbz@ftsm.ukm.my,Azlina A.Dep. of Indus. Comp.UKM, Selangor,Malaysia.aa@ftsm.ukm.my Behrang P. Dep. of Info. Science, UKM, Selangor, Mlaysia.hani_p114@yahoo.com Choo W.O.UTAR, Kampar,Perak, Malaysia.kenny@yahoo.com Voice Recognition System for the Visually Impaired: Virtual Cognitive Approach, IEEE2008
[6] Xinxin Wang1,Feiran Wu1,Zhiqian Ye11College of Biomedical Engineering& Instrument Science, Zhejiang University, Hangzhou, China meaita2009@gmail.com com, yezhiqian@hzcnc, *The Application of Speech Recognition in Radiology Information System*,IEEE2010.
[7] Huu-Cong Nguyen, Shim-Byoung , Chang-Hak Kang, Dong-Jun Park and Sung-Hyun Han Division of Mechanical System Eng., Graduate School, Kyungnam University, Masan, Korea *Integration of Robust Voice Recognition and Navigation System on Mobile Robot*, ICROS-SICE International Joint Conference 2009
[8] X., Huang, A., Acero, and H.W., Hon, "Spoken Language Processing: A Guide to Theory, Algorithm, and System Development". Prentice Hall, Upper Saddle River, NJ, USA, 2001.
[9] M., Ursin, "Triphone Clustering in Finnish Continuous Speech Recognition". Master Thesis, Department of Computer Science, Helsinki University of Technology, Finland, 2002.
[10] O. Khalifa, S. Khan, M.R. Islam, M. Faizal and D. Dol, "Text Independent Automatic Speaker Recognition", 3rd International Conference on Electrical & Computer Engineering,, Dhaka, Bangladesh, 28-30 December 2004, pp. 561-564.
[11] C.R. Buchanan, "Informatics Research Proposal – Modeling the Semantics of Sound", School of Informatics, University of Edinburgh, United Kingdom, March 2005.
[12] http://ozanmut.sitemynet.com/asr.htm, Retrieved in November 2005.
[13] X., Huang, A., Acero, and H.W., Hon, "Spoken Language Processing: A Guide to Theory, Algorithm, and System Development". Prentice Hall, Upper Saddle River, NJ, USA, 2001.
[14] Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantizer Design", IEEE Transactions on Communications, VOL. COM-28, No. 1, pp. 84 - 95, January 1980.
[15] D., Jurafsky, "Speech Recognition and Synthesis: Acoustic Modeling", winter 2005.
[16] S.K., Podder, "Segment-based Stochastic Modelings for Speech Recognition". PhD Thesis. Department of Electrical and Electronic Engineering, Ehime University, Matsuyama 790-77, Japan, 1997.
[17] S.M., Ahadi, H., Sheikhzadeh, R.L., Brennan, and G.H., Freeman, "An Efficient Front-End for Automatic Speech Recognition". IEEE International Conference on Electronics, Circuits and Systems (ICECS2003), Sharjah, United Arab Emirates, 2003.
[18] M., Jackson, "Automatic Speech Recognition: Human Computer Interface for Kinyarwanda Language". Master Thesis, Faculty ofComputing and Information Technology, Makerere University, 2005.
[19] M.R., Hasan, M., Jamil, and M.G., Saifur Rahman, "Speaker Identification Using Mel Frequency Cepstral Coefficients". 3rd International Conference on Electrical and Computer Engineering, Dhaka, Bangladesh, 2004, pp. 565-568.
M.Z., Bhotto and M.R., Amin, "Bangali Text Dependent Speaker Identification Using MelFrequency Cepstrum Coefficient and VectorQuantization". 3rd International Conference on Electrical and Computer Engineering, Dhaka, Bangladesh, 2004, pp. 569-572.

