# SPEECH RECOGNITION SYSTEM

Speech is a natural mode to interact with others. Speech recognition allows machining system to turn the incoming speech signals into commands through the process of identifying and understanding. There are four classes of speech.

**Isolated words**
It accepts single word or single utterance at a time. This system has “listen/non-listen” states, where they require the speaker to wait between utterances.

**Connected words**
They allow separate utterances to be run together with minimal pause between them.

**Continuous speech**
In this users speak naturally and recognisers determine the content, but it is difficult to determine the utterance boundaries.

**Spontaneous speech**
An ASR system with spontaneous speech ability should be able to handle a variety of natural speech features such as words being run together.

## BASIC PRINCIPLE OF SPEECH RECOGNITION

The speech recognition system is essentially a pattern recognition system, including feature extraction, pattern matching and the reference model library [11]. A speech recognition engine (or speech recognizer) takes an audio stream as input and turns it into a text transcription. The speech recognition process can be thought of as having a front end and a back end.

i) **Convert Audio Input**
The front end processes the audio stream, isolating segments of sound that are probably speech and converting them into a series of numeric values that characterize the vocal sounds in the signal.

ii) **Match Input to Speech Models**
The back end is a specialized search engine that takes the output produced by the front end and searches across three databases: an acoustic model, a lexicon, and a language model.

The quality of a recognizer is determined by how good it is at refining its search, eliminating the poor matches, and selecting the more likely matches. This depends in large part on the quality of its language and acoustic models and the effectiveness of its algorithms, both for processing sound and for searching across the models.

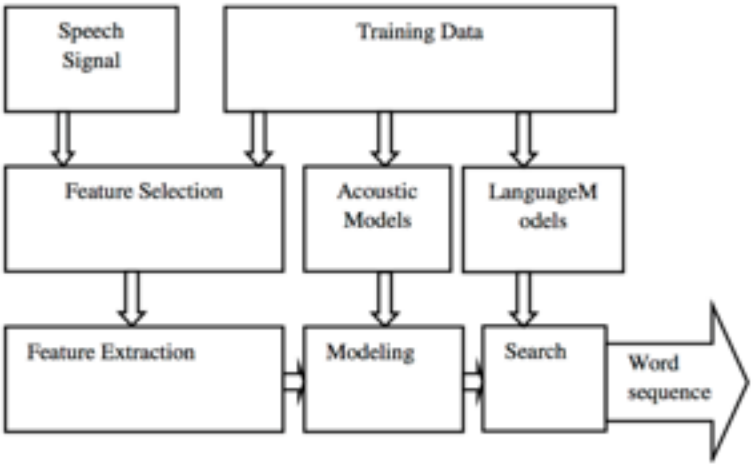

**Figure 2: Principle of Speech Recognition**

1. **SPEECH SIGNAL**
   The signal requires some preprocessing steps to be followed spectral estimation method based on the linear predictive coding principle. This model is built on the idea of prediction error minimization. The signal spectrum obtained forms the basis for the analysis of Toeplitz Matrix (TM). Then TM is applied to describe the speech signal in the form of a feature vector for each signal.

2. **FEATURE SELECTION**
   There are several choices of features for speech recognition, amplitude, zero-crossing rate, and spectral content. Amplitude is a primary source for endpoint information and also for vowel/consonant discrimination in phonetic recognition. Zero-crossing rate and spectrum balance provides for the characterization of fricatives and sibilants. High-resolution spectral information like pitch and format, LP parameters or filter-bank output provides formant and formant transition information. The patterns used in recognition may be represented either as time functions which span the entire word, or as a set of characteristic feature values which represent the subdivisions of the word.

3. **FEATURE EXTRACTION**
   It should be possible to recognize speech directly from the digitized waveform. However, because of the large variability of the speech signal, it is better to perform some feature extraction that would reduce that variability. Particularly, eliminating various source of information, such as whether the sound is voiced or unvoiced and, if voiced, it eliminates the effect of the periodicity or pitch, amplitude of excitation signal and fundamental frequency etc [19]. Various feature extraction methods are:
   i. Principal Component Analysis
   ii. Linear Discriminant Analysis
   iii. Wavelet Analysis
   iv. Linear Predictive Coding.

4. **ACOUSTIC MODELS**
   An acoustic model is a file that contains statistical representations of each of the distinct sounds that makes up a word. Each of these statistical representations is

assigned a label called a phoneme. The English language has about 40 distinct sounds that are useful for speech recognition, and thus we have 50 different phonemes. An acoustic model is created by taking a large database of speech (called a speech corpus) and using special training algorithms to create statistical representations for each phoneme in a language.

## DIFFERENT APPROACHES TO SPEECH RECOGNITION

### 1. ACOUSTIC PHONETIC APPROACH

There exist finite, distinctive phonetic units (phonemes) in spoken language and that these units are broadly characterized by a set of acoustics properties that are manifested in the speech signal over time. The first step is the spectral analysis of the speech combined with a feature detection that converts the spectral measurements to a set of features that describe the broad acoustic properties of the different phonetic units. The speech signal is segmented into stable acoustic regions, followed by attaching one or more phonetic labels to each segmented region, resulting in a phoneme lattice characterization of the speech.

### 2. PATTERN RECOGNITION APPROACH

A speech pattern representation can be in the form of a speech template or a statistical model (e.g., a HIDDEN MARKOV MODEL) and can be applied to a sound (smaller than a word), a word, or a phrase.

#### I. Template based approach

A collection of prototypical speech patterns are stored as reference patterns which represents the dictionary of candidate words. An unknown spoken utterance is matched with each of these reference templates and a category of the best matching pattern is selected.

#### II. Stochastic approach

Stochastic modelling [20] entails the use of probabilistic models to deal with uncertain or incomplete information. The most popular stochastic approach today is hidden Markov modeling. A hidden Markov model is characterized by a finite state markov model and a set of output distributions.

### 3. DYNAMIC TIME WARPING

It is an algorithm for measuring similarity between two temporal sequences which may vary in time or speed. DTW is a method that calculates an optimal match between two given sequences. DTW has been applied to temporal sequences of video, audio, and graphics data — indeed, any data which can be turned into a linear sequence can be analyzed with DTW.

### 4. ARTIFICIAL INTELLIGENCE APPROACH

Artificial Intelligence approach [20] is a hybrid of the acoustic phonetic approach and pattern recognition approach. In its pure form, knowledge engineering design involves the direct and explicit incorporation of expert’s speech knowledge into a recognition system. This form of knowledge based system enhancement has contributed considerably to the design of all successful strategies reported. It plays an important role in the selection of a suitable input representation, the definition of units of speech, or the design of the recognition algorithm itself.

## PRODUCT AVAILABLE COMMERCIALLY

1. **Alexa - Amazon Echo**
   Amazon Echo (known in-development as Doppler or Project D and shortened and referred to as Echo) is a smart speaker developed by Amazon.com. The device consists of a 9.25-inch (23.5 cm) tall cylinder speaker with a seven-piece microphone array. The device is capable of voice interaction, music playback, making to-do lists, setting alarms, streaming podcasts, playing audiobooks, and providing weather, traffic and other real time information. It can also control several smart devices using itself as a home automation hub. Even when it's alseep, Echo is still listening for the trigger phrase that will turn it on and start streaming your voice to the cloud.

   **Applications**
   In a few years, Amazon is planning to fill homes with "smart" clocks, refrigerators, TVs and security systems. Together, they'll help voice assistants break out of single devices to be available anywhere in houses, cars or workplaces.

2. **GoVivace**
   GoVivace offers a reliable and robust Automatic Speech Recognition engine. It provide the speech recognition software as a Software Developer Kit (SDK) library as well as the websocket with bidirectional streaming to use the software as a service. These include interfaces for Independent Software Vendors and developers, even those working on cloud-based applications.The speech to text engine is designed to work 24 X 7 X 365 and can be used to build mobile, web as well as high volume application. The engine compares the spoken input with a number of pre-specified possibilities. The entire set of pre-specified possibilities constitute the application's grammar, which powers the interface between the dialogue-speaker and the back-end processing. The GoVivace speech recognition solution needs only very simple grammar, but also supports very large grammars for complex tasks such as dates, complex commands and yellow pages styled complex directory lookups. GoVivace additionally offers consulting services for the construction of complex grammars. Performance tuning is another service, whereby GoVivace troubleshoots poorly performing grammars.

   The GoVivace's speech recognition software can work with both pre-compiled grammars that can be referenced by name, and on-the-fly grammars that evolve as the client uses the application and which can be detected if reused. Both kinds of grammars are stored on the server after compilation, to ensure fast processing.

   A key feature of the GoVivace Automatic Speech Recognition engine is it uses a statistical language model to understand natural language, which means it is not limited to understanding speech which matches its grammar. Developers integrating the engine into speech recognition systems will appreciate being able to create advanced intuitive natural language processing interfaces boasting of high linguistic intelligence quotients.

   The GoVivace automatic speech recognition solution supports a distributed client / server architecture for easy scaling and to support an ever growing list of client devices. A load balancer can be used as the front end, and servers added to the system at the back end to allow for redundancy, reliability and scalability.

3. **Windows Speech Recognition**

Windows Speech Recognition is a speech recognition component developed by Microsoft and introduced in the Windows Vista operating system that enables the use of voice commands to perform operations, such as the dictation of text, within applications and the operating system itself.

**Working**

A speech recognition engine (or speech recognizer) takes an audio stream as input and turns it into a text transcription. The speech recognition process can be thought of as having a front end and a back end.

The front end processes the audio stream, isolating segments of sound that are probably speech and converting them into a series of numeric values that characterize the vocal sounds in the signal.

The back end is a specialized search engine that takes the output produced by the front end and searches across three databases: an acoustic model, a lexicon, and a language model.

The acoustic model represents the acoustic sounds of a language, and can be trained to recognize the characteristics of a particular user's speech patterns and acoustic environments.

The lexicon lists a large number of the words in the language, and provides information on how to pronounce each word.

The language model represents the ways in which the words of a language are combined.

For any given segment of sound, there are many things the speaker could potentially be saying. The quality of a recognizer is determined by how good it is at refining its search, eliminating the poor matches, and selecting the more likely matches. This depends in large part on the quality of its language and acoustic models and the effectiveness of its algorithms, both for processing sound and for searching across the models.

Grammars

While the built-in language model of a recognizer is intended to represent a comprehensive language domain (such as everyday spoken English), a speech application will often need to process only certain utterances that have particular semantic meaning to that application. Rather than using the general purpose language model, an application should use a grammar that constrains the recognizer to listen only for speech that is meaningful to the application. This provides the following benefits:

1. Increases the accuracy of recognition
2. Guarantees that all recognition results are meaningful to the application
3. Enables the recognition engine to specify the semantic values inherent in the recognized text

The Microsoft Speech Platform SDK 11 provides processes for authoring grammars programmatically, and also supports grammars authored using industry-standard markup language

**Applications**

Windows Speech Recognition allows a user to control a computer, including the operating system desktop user interface, through voice commands. Applications, including most of those that are bundled with Windows, can also be controlled through voice commands. By using speech recognition, users can dictate text within documents and e-mail messages, fill out forms, control the operating system user interface, perform keyboard shortcuts, and move the mouse cursor.

4. **S-voice**

S Voice is an intelligent personal assistant and knowledge navigator which is only available as a built-in application for the SamsungGalaxy S III, S III Mini (including NFC Variant), S4, S4 Mini, S4 Active, S5, S5 Mini, S II Plus, Note II, Note 3, Note 4, Note 10.1, Note 8.0, Stellar, Mega, Grand, Avant, Core, Ace 3, Tab 3 7.0, Tab 3 8.0, Tab 3 10.1, Galaxy Camera, and other 2013 or later Samsung Android devices. The application uses a natural language user interface to answer questions, make recommendations, and perform actions by delegating requests to a set of Web services.

Some of the capabilities of S Voice include making appointments, opening apps, setting alarms, updating social network websites such as Facebook or Twitter and navigation. S Voice also offers multitasking as well as automatic activation features, for example, when the car engine is started.

5. **TrulyNatural**

TrulyNatural™ is the first embedded large vocabulary continuous speech recognizer system which can provide the same state-of-the-art performance as cloud based technologies, but on the device. Many applications don't need, or don't want to rely on, a cloud based connection to do high performance speech recognition. Connections can frequently fail or cause fatal delays, and recent concerns have been expressed about sending personal data to the clouds where it could be stolen or used for purposes undesirable to the owner.

Sometimes a client/cloud distributed system with optimal performance is desired or a standalone embedded solution that can perform complex or natural language tasks without being too slow or lacking accuracy. TrulyNatural is the solution for these needs.

Sensory uses a unique form of a neural network with deep learning to achieve acoustic models an order of magnitude smaller than the present state-of-the-art. These neural networks employ the most recent breakthroughs in speech feature extraction to produce superior accuracy in real world noise. Combining this with a small Finite State Transducer (FST) enables storage and accurate processing of multiple large search domains as required in today's market. Applications using TrulyNatural include Voice Dial

6. **CORTANA**

Cortana is an intelligent personal assistant created by Microsoft for Windows 10, Windows 10 Mobile, Windows Phone 8.1 (where it now supersedes Bing Mobile), Microsoft Band, Xbox One, iOS and Android. Cortana can set reminders, recognize natural voice without the requirement for keyboard input, and answer questions using information from the Bing search engine (e.g., current weather and traffic conditions, sports scores, biographies). When a user presses the "Search" button on their device. Cortana includes a music recognition service. Cortana can simulate rolling dice and flipping a coin. Cortana's "Concert Watch" monitors Bing searches to determine which bands or musicians the user is interested in. It integrates with the Microsoft Band watch band for Windows Phone devices if connected via Bluetooth, it can make reminders and phone notifications. The natural language processing capabilities of Cortana are derived from Tellme Networks (bought by Microsoft in 2007) and are coupled with a Semantic search database called Satori.

7. **SIRI**

Siri is a computer program that works as an intelligent personal assistant and knowledge navigator, part of Apple Inc.'s iOS, watchOS, macOS, and tvOS operating systems. The feature uses a natural language user interface to answer questions, make recommendations, and perform actions by delegating requests to a set of Web services. The software, both in its original version and as an iOS feature, adapts to the user's individual language usage and individual searches (preferences) with continuing use, and returns results that are individualized. Siri was originally introduced as an iOS application available in the App Store by Siri Inc., which was acquired by Apple.

## REFERENCES:

1. http://ieeexplore.ieee.org/document/237532/?arnumber=237532&tag=1
2. https://en.wikipedia.org/wiki/List_of_speech_recognition_software
3. https://en.wikipedia.org/wiki/Amazon_Echo
4. https://www.govivace.com/products/automatic-speech-recognition-asr
5. https://en.wikipedia.org/wiki/Windows_Speech_Recognition
6. https://en.wikipedia.org/wiki/S_Voice
7. http://www.sensory.com/products/technologies/
8. http://www.sensory.com/products/applications/
9. https://en.wikipedia.org/wiki/Speech_recognition#Applications
10. https://en.wikipedia.org/wiki/Siri
11. https://en.wikipedia.org/wiki/Cortana_(software)
12. http://www.ijana.in/Special%20Issue/C13.pdf
13. https://www.ijarcsse.com/docs/papers/Volume_4/8_August2014/V4I8-0362.pdf
14. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.259.445&rep=rep1&type=pdf
15. http://ijsetr.org/wp-content/uploads/2014/08/IJSETR-VOL-3-ISSUE-8-2192-2196.pdf
16. https://www.resna.org/sites/default/files/JournalArticle22.2a.pdf
17. https://arxiv.org/pdf/1001.2267.pdf
18. http://shodhganga.inflibnet.ac.in/bitstream/10603/38340/10/10_chapter%202.pdf
19. 18. Urmila Shrawankar, Vilas Thakar, “Techniques for Feature Extraction in Speech Recognition System : A Comparative Study”.
20. 19. Jianliang Meng, Junwei Zhang and Haoquan Zhao, “Overview of the Speech Recognition Technology”, 2012 Fourth International Conference on Computational and Information Sciences, 978-0- 7695-4789-3/12$26.00©2012 IEEE.
21. 20. R.K.Moore, Twenty things we still don t know about speech , Proc.CRIM/ FORWISS Workshop on Progress and Prospects of speech Research an Technology , 1994.