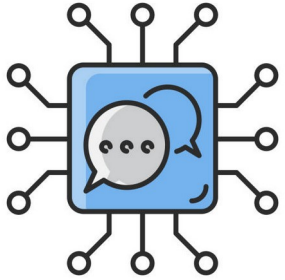
Speech Processing

Tushar B. Kute,
http://tusharkute.com



Articulatory Phonetics

- The production of speech involves 3 processes:
- Initiation: Setting air in motion through the vocal tract.
- Phonation: The modification of airflow as it passes through the larynx (related to voicing).
- Articulation: The shaping of airflow to generate particular sound types (related to manner)

Articulatory Phonetics

- Articulatory phonetics refers to the “aspects of phonetics which looks at how the sounds of speech are made with the organs of the vocal tract” Ogden (2009:173).
- Articulatory phonetics can be seen as divided up into three areas to describe consonants. These are voice, place and manner respectively.
- Each of these will now be discussed separately, although all three areas combine together in the production of speech.

Articulatory Phonetics: Voice

- In English we have both voiced and voiceless sounds. A sound fits into one of these categories according to how the vocal folds behave when a speech sound is produced.
- Voiced: Voiced sounds are sounds that involve vocal fold vibrations when they are produced. Examples of voiced sounds are /b,d,v,m/.
- If you place two fingers on either side of the front of your neck, just below your jawbone, and produce a sound, you should be able to feel a vibrating sensation. This tells you that a sound is voiced.
- Voiceless: Voiceless sounds are sounds that are produced with no vocal fold vibration. Examples of voiceless sounds in English are /s,t,p,f/.

Articulatory Phonetics: Place

- The vocal tract is made up of different sections, which play a pivotal role in the production of speech. These sections are called articulators and are what make speech sounds possible. They can be divided into two types.
- The active articulator is the articulator that moves towards another articulator in the production of a speech sound. This articulator moves towards another articulator to form a closure of some type in the vocal tract (i.e open approximation, close, etc – define)
- The passive articulator is the articulator that remains stationary in the production of a speech sound. Often, this is the destination that the active articulator moves towards (i.e the hard palate).

Articulatory Phonetics: Place

- Bilabial: Bilabial sounds involve the upper and lower lips. In the production of a bilabial sound, the lips come into contact with each other to form an effective constriction. In English, /p,b,m/ are bilabial sounds.
- Labiodental: Labiodental sounds involve the lower lip (labial) and upper teeth (dental) coming into contact with each other to form an effective constriction in the vocal tract. Examples of labiodental sounds in English are /f,v/. Labiodental sounds can be divided into two types.
  - a) Endolabial: sounds produced where the upper teeth are pressed against the inside of the lower lip.
  - b) Exolabial: sounds produced where the upper teeth are pressed against the outer side of the lower lip.

Articulatory Phonetics: Place

- Dental: Dental sounds involve the tongue tip (active articulator) making contact with the upper teeth to form a constriction.
  - Examples of Dental sounds in English are / θ, ð/. If a sound is produced where the tongue is between the upper and lower teeth, it is attributed the term ‘interdental’.

Articulatory Phonetics: Place

- Alveolar: First of all, before I explain what an alveolar sound is, it's useful to locate the alveolar ridge itself.
- If you place your tongue just behind your teeth and move it around, you'll feel a bony sort of ridge. This is known as the alveolar ridge.
- Alveolar sounds involve the front portion of the tongue making contact with the alveolar ridge to form an effective constriction in the vocal tract.
- Examples of alveolar sounds in English are /t,d,n,l,s/.

Articulatory Phonetics: Place

- Postalveolar: Postalveolar sounds are made a little further back ('post') from the alveolar ridge.
- A postalveolar sound is produced when the blade of the tongue comes into contact with the post-alveolar region of your mouth. Examples of post-alveolar sounds in English are / ʃ, ʒ /.
- Palatal: Palatal sounds are made with the tongue body (the big, fleshy part of your tongue). The tongue body raises up towards the hard-palate in your mouth (the dome shaped roof of your mouth) to form an effective constriction.
- An example of a palatal sounds in English is /j/, usually spelt as <y>.

Articulatory Phonetics: Place

- Velar: Velar sounds are made when the back of the tongue (tongue dorsum) raises towards the soft palate, which is located at the back of the roof of the mouth.
- This soft palate is known as the velum. An effective constriction is then formed when these two articulators come into contact with each other. Examples of velar sounds in English are /k,g ŋ /.

Articulatory Phonetics: Manner

- In simple terms, the manner of articulation refers to the way a sound is made, as opposed to where it’s made.
- Sounds differ in the way they are produced. When the articulators are brought towards each other, the flow of air differs according to the specific sound type.
- For instance, the airflow can be completely blocked off or made turbulent.

Stop Articulations

- Stop articulations are sounds that involve a complete closure in the vocal tract. The closure is formed when two articulators come together to prevent air escaping between them.
- Stop articulations can be categorized according to the kind of airflow involved.
- The type of airflow can be oral (plosives) or nasal (nasals).

Stop Articulations

- 1a) Plosives: are sounds that are made with a complete closure in the oral (vocal) tract. The velum is raised during a plosive sound, which prevents air from escaping via the nasal cavity. English plosives are the sounds /p,b,t,d,k,g/. Plosives can be held for quite a long time and are thus also called ‘maintainable stops’.
- 1b) Nasals are similar to plosives in regards to being sounds that are made with a complete closure in the oral (vocal) tract. However, the velum is lowered during nasal sounds, which allows airflow to escape through the nasal cavity. There are 3 nasal sounds that occur in English /m,n, ŋ/

Fricatives

- Fricative sounds are produced by narrowing the distance between the active and passive articulators causing them to be in close approximation.
- This causes the airflow to become turbulent when it passes between the two articulators involved in producing a fricative sound. English fricatives are sounds such as / f,v, θ,ð, s,z, ʃ,ʒ /

Approximants

- Approximant sounds are created by narrowing the distance between the two articulators.
- Although, unlike fricatives, the distance isn't wide enough to create turbulent airflow. English has 4 approximant sounds which are /w,j,r,l/.

Vowels

- When it comes to vowels, we use a different specification to describe them. We look at the vertical position of the tongue, the horizontal position of the tongue and lip position.
- Vowels are made with a free passage of airflow down the mid-line of the vocal tract.
- They are usually voiced and are produced without friction.

Vowels

- Vertical tongue position (close-open): vertical tongue position refers to how close the tongue is to the roof of the mouth in the production of a vowel.
- If the tongue is close, it is given the label close. However, if the tongue is low in the mouth when a vowel is produced, it's given the label open. + close-mid/open mid (see below).
  - Some examples of open vowels: ɪ, ʊ
  - Some examples of close vowels: æ, ɒ,

Vowels

- Horizontal tongue position (front, mid, back): Horizontal tongue refers to where the tongue is positioned in the vocal tract in terms of ‘at the front’ or ‘at the back’ when a vowel is produced.
- If the tongue is at the front of the mouth it’s given the label front, if the tongue is in the middle of the mouth it’s given the label mid and if the tongue is at the back of the mouth it’s given the label back.
  - Some examples of front vowels: ɪ , e, æ
  - Some examples of mid vowels: ə
  - Some examples of back vowels: ʌ,ɒ

Vowels

- Lip position: As is inferred, lip position concerns the position of the lips when a vowel is produced.
- The lips can either be round, spread or neutral.
  - Examples of round vowels: u, o
  - Examples of spread vowels: ɪ, ɛ

Vowels

- There are also different categories of vowels, for example: monophthongs and diphthongs.
- Monophthongs: Monophthongs are vowels that are produced by a relatively stable tongue position.
- Monophthongs can be divided into two categories according to their duration.
- These are long and short vowels and their duration is mirrored in their names.
  - Examples of short vowels: e, æ, ɪ, ʊ
  - Examples of long vowels: ɔ: ɜ:, i:, u:

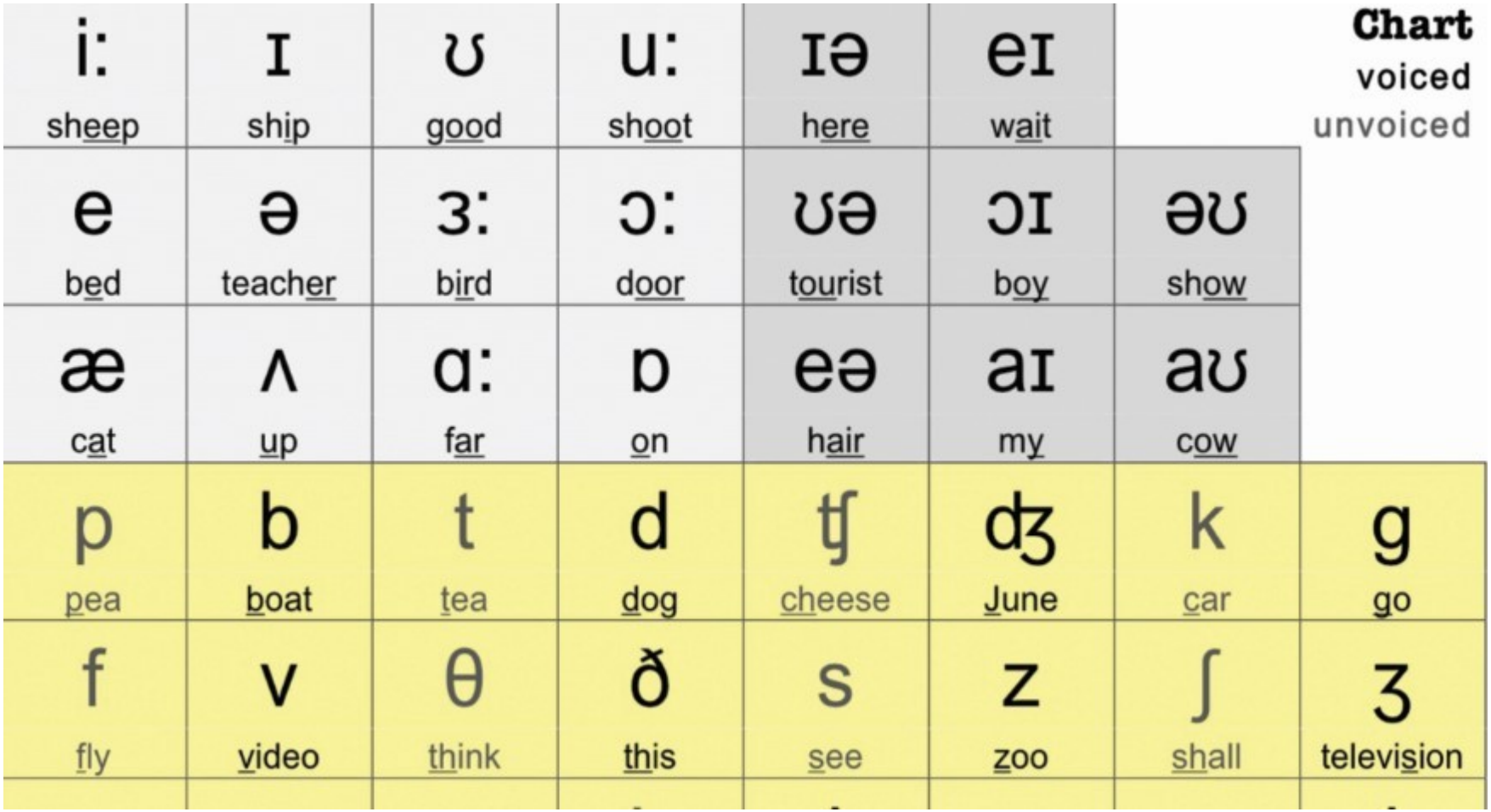
Vowels

- There are also different categories of vowels, for example: monophthongs and diphthongs.
- Diphthongs: Diphthongs are vowels where the tongue moves from one part of the mouth to another. They can be seen as starting of as one vowel and ending as a different vowel.
  - Here are some examples: /aʊ, ɪə, ɔɪ, əʊ/

Phonetic Transcription

- Phonetic transcription (also known as phonetic script or phonetic notation) is the visual representation of speech sounds (or phones) by means of symbols.
- The most common type of phonetic transcription uses a phonetic alphabet, such as the International Phonetic Alphabet.

Phonetic Transcription

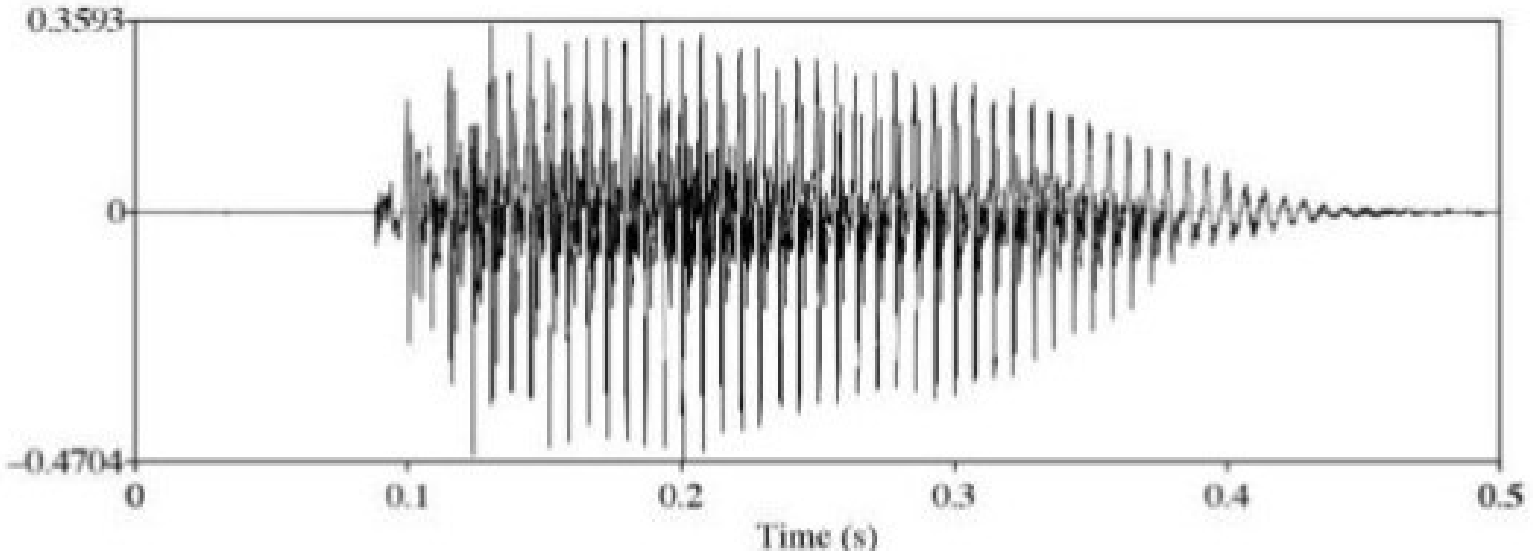


iː sheep	ɪ ship	ʊ good	uː shoot	ɪə here	eɪ wait		**Chart** voiced unvoiced
e bed	ə teacher	ɜː bird	ɔː door	ʊə tourist	ɔɪ boy	əʊ show	
æ cat	ʌ up	ɑː far	ɒ on	eə hair	aɪ my	aʊ cow	
p pea	b boat	t tea	d dog	ʧ cheese	ʤ June	k car	g go
f fly	v video	θ think	ð this	s see	z zoo	ʃ shall	ʒ television

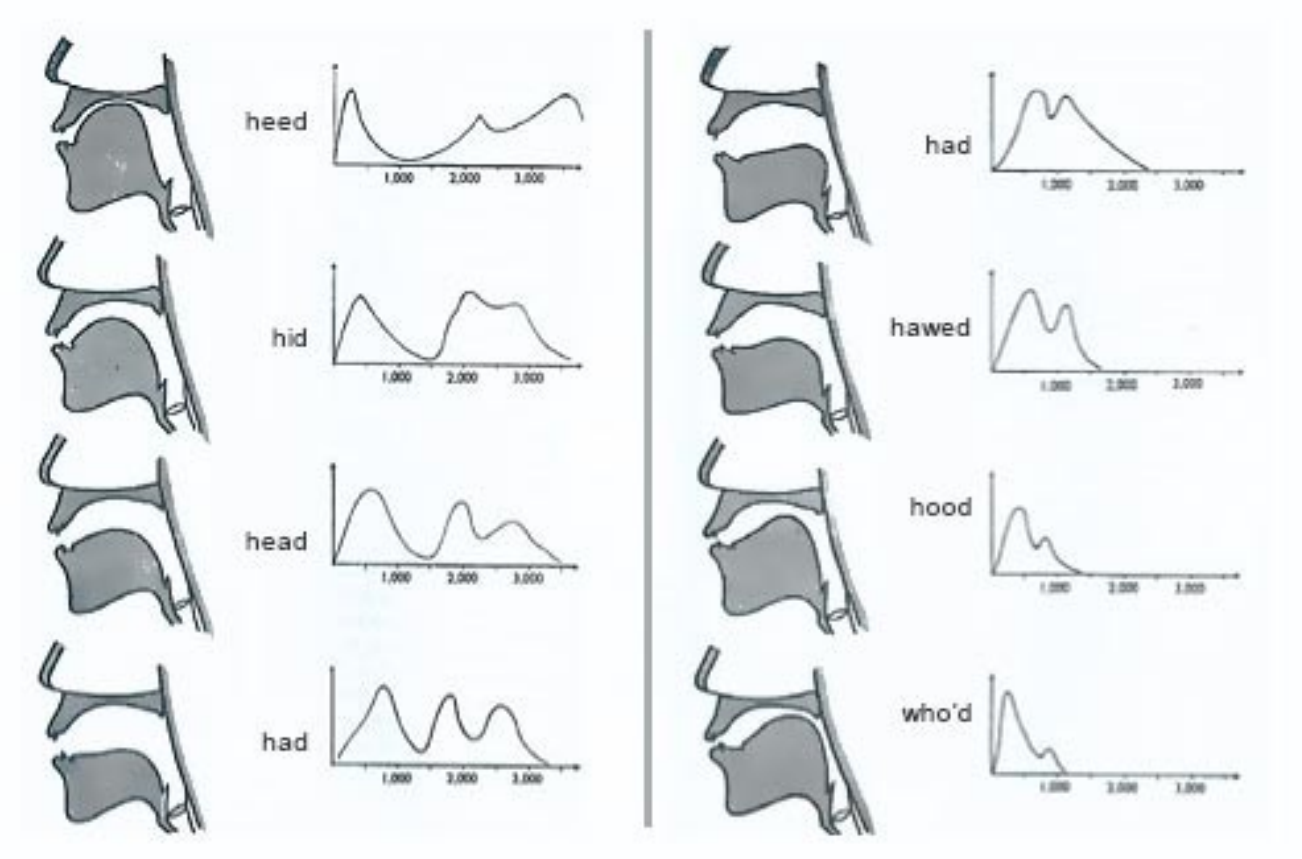
Acoustic Phonetics

- Acoustic phonetics is the study of the physical properties of speech, and aims to analyse sound wave signals that occur within speech through varying frequencies, amplitudes and durations.
- One way we can analyse the acoustic properties of speech sounds is through looking at a waveform.
- Pressure changes can be plotted on a waveform, which highlights the air particles being compressed and rarefied, creating sound waves that spread outwards.
- A tuning fork being struck can provide an example of the pressure fluctuations in the air and how the air particles oscillate (move in one direction rhythmically) when we perceive sound.

Acoustic Phonetics



Acoustic Phonetics



Phonology

- Phonology refers to the study of the sounds of a language.
- Every language has its own inventory of sounds and logical rules for combining those sounds to create words.
- The phonology of a language essentially refers to its sound system and the processes used to combine sounds in spoken language.

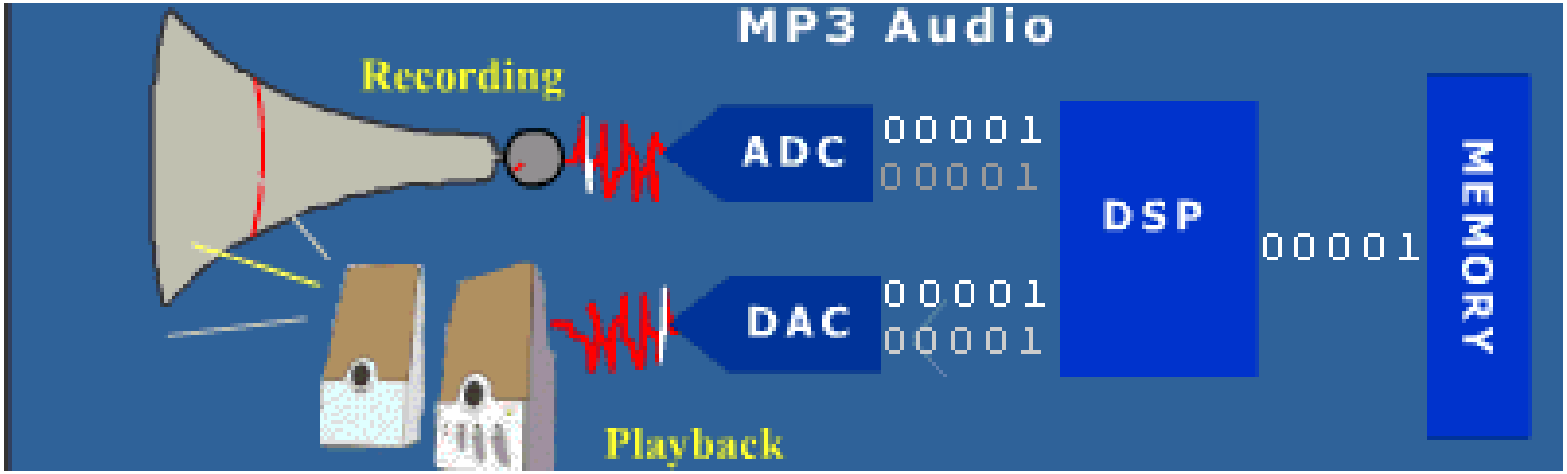
Computational Phonology

- Computational phonology is the application of formal and computational techniques to the representation and processing of phonological information. It "deals with the changes in sound patterns that take place when words are put together".
- While phonological analysis defines formal models and systematically tests it against data, computer science would speed up task, i.e. by software which induces (computational) models.

Digital Signal Processing

- Digital Signal Processors (DSP) take real-world signals like voice, audio, video, temperature, pressure, or position that have been digitized and then mathematically manipulate them.
- A DSP is designed for performing mathematical functions like "add", "subtract", "multiply" and "divide" very quickly.
- Signals need to be processed so that the information that they contain can be displayed, analyzed, or converted to another type of signal that may be of use.

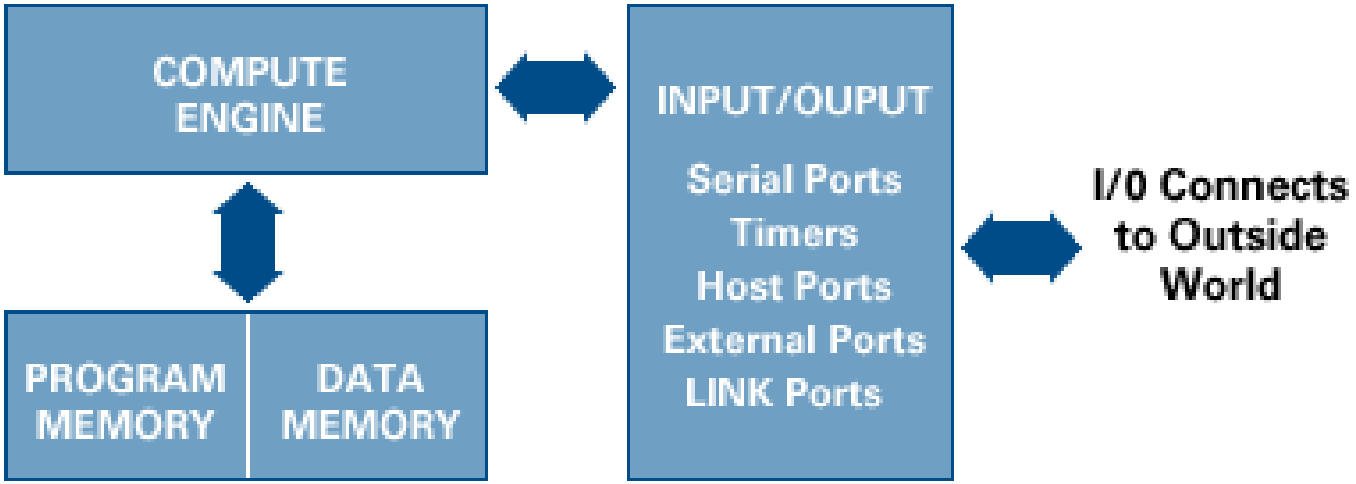
Digital Signal Processing



Inside Digital Signal Processing

- A DSP contains these key components:
  - Program Memory: Stores the programs the DSP will use to process data
  - Data Memory: Stores the information to be processed
  - Compute Engine: Performs the math processing, accessing the program from the Program Memory and the data from the Data Memory
  - Input/Output: Serves a range of functions to connect to the outside world.

Inside Digital Signal Processing



Automatic Speech Recognition

- Automatic Speech Recognition, or ASR, is the use of Machine Learning or Artificial Intelligence (AI) technology to process human speech into readable text.
- The field has grown exponentially over the past decade, with ASR systems popping up in popular applications we use every day such as TikTok and Instagram for real-time captions, Spotify for podcast transcriptions, Zoom for meeting transcriptions, and more.

How ASR works?

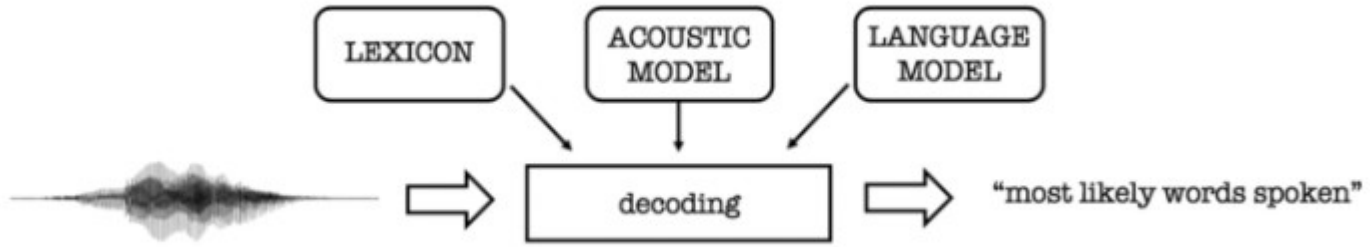
- Today, there are two main approaches to Automatic Speech Recognition:
  - a traditional hybrid approach and
  - an end-to-end Deep Learning approach.

How ASR works?

- Traditional Hybrid Approach
  - The traditional hybrid approach is the legacy approach to Speech Recognition and has dominated the field for the past fifteen years.
  - Many companies still rely on this traditional hybrid approach simply because it’s the way it has always been done--there is more knowledge around how to build a robust model because of the extensive research and training data available, despite plateaus in accuracy.

Traditional HMM and GMM

- Traditional HMM (Hidden Markov Models) and GMM (Gaussian Mixture Models) require forced aligned data.
- Force alignment is the process of taking the text transcription of an audio speech segment and determining where in time particular words occur in the speech segment.



Drawbacks

- Each model must be trained independently, making them time and labor intensive.
- Forced aligned data is also difficult to come by and a significant amount of human labor is needed, making them less accessible.
- Finally, experts are needed to build a custom phonetic set in order to boost the model's accuracy.

End-to-end Deep Learning

- With an end-to-end system, you can directly map a sequence of input acoustic features into a sequence of words.
- The data does not need to be force-aligned. Depending on the architecture, a Deep Learning system can be trained to produce accurate transcripts without a lexicon model and language model, although language models can help produce more accurate results.

End-to-end Deep Learning

- CTC, LAS, and RNNTs are popular Speech Recognition end-to-end Deep Learning architectures.
- These systems can be trained to produce super accurate results without needing force aligned data, lexicon models, and language models.

Advantages

- End-to-end Deep Learning models are easier to train and require less human labor than a traditional approach.
- They are also more accurate than the traditional models being used today.
- The Deep Learning research community is actively searching for ways to constantly improve these models using the latest research as well, so there’s no concern of accuracy plateaus any time soon--in fact, we’ll see Deep Learning models reach human level accuracy in the next few years.

ASR Applications

- Telephony
- Video Platforms
- Media Monitoring
- Virtual Meetings

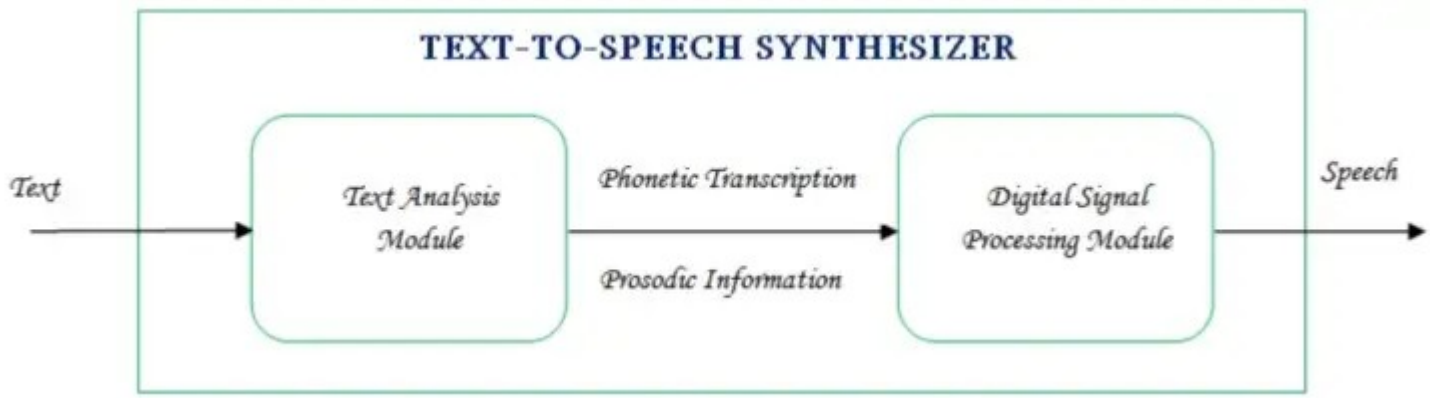
Text to Speech Systems

- The goal of TTS is the automatic conversion of written text into corresponding speech.
- The speech synthesis field has witnessed much advancement in the past few decades.
- A general TTS synthesizer comprises a text analysis module and digital signal processing (DSP) module.

Text to Speech Systems

- Figure shows the functional diagram of a very general TTS synthesizer.
- The text analysis module produces a phonetic transcription of the text read, together with the desired intonation and rhythm (i.e., prosody).
- The DSP module produces the synthetic speech corresponding to the transcription produced by the text analysis module.

Text to Speech Systems



Conversational AI

- As the name suggests, conversational AI is a broad term directed towards the use of AI and automation technologies to build equipment that can convert text to speech. Amazon's Alexa is one of the most conspicuous examples serving this category.
- Conversational AI is aimed at providing a personalized experience to the consumer. Most importantly, they offer two-way interaction. This is just like a Chabot. Instead, it is more of an audio-bot, which can satisfy your daily requirements.
- According to the various surveys, it is messaging apps that are fuelling conversational AI. These include WhatsApp, Messenger, We Chat, Skype, Instagram, Telegram, Snapchat, Line, etc.

Google Conversational AI

- Google also provides google cloud text to speech that can convert the text into a human-like voice in more than 180 sounds and 30+ languages variants.
- In addition to that, this cloud API can be implemented in any application or device like phones, tablets that are able to send REST or gRPC requests.

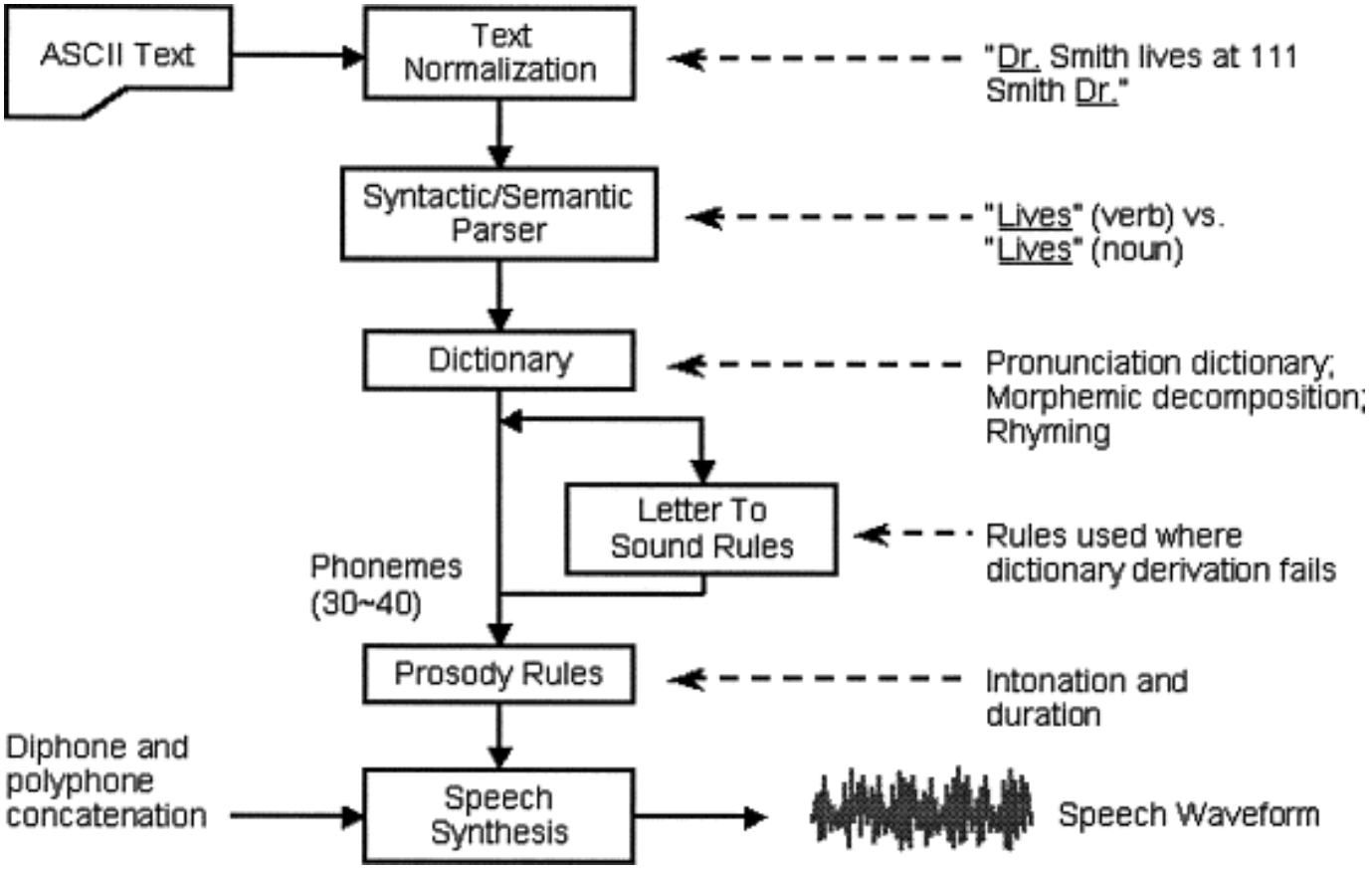
Speech Synthesis

- Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer, and can be implemented in software or hardware products.
- A text-to-speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech.
- The reverse process is speech recognition.

Speech Synthesis

- Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database.
- Systems differ in the size of the stored speech units; a system that stores phones or diphones provides the largest output range, but may lack clarity.
- For specific usage domains, the storage of entire words or sentences allows for high-quality output.
- Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output

Speech Synthesis



Language Models

- Language Model
  - The language model (LM) models the statistics of language. It learns which sequences of words are most likely to be spoken, and its job is to predict which words will follow on from the current words and with what probability.

Thank you

This presentation is created using LibreOffice Impress 7.4.1.2, can be used freely as per GNU General Public License

@mitu_skillologies

@mITuSkillologies

@mitu_group

@mitu-skillologies

@MITUSkillologies

kaggle

@mituskillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

@mituskillologies

contact@mitu.co.in

tushar@tusharkute.com