Features for Automatic Speech Recognition

S. R. M. Prasanna

Dean (Faculty Welfare, Research & Development)
Professor, Dept of Electrical Engineering
Indian Institute of Technology Dharwad

prasanna@iitdh.ac.in

July 19, 2021

Some Good Books in Speech Processing

- Rabiner, Jhuang and Yegnanarayana, ”Fundamentals of Speech Recognition”, Pearon LPE, 2006.
- L.R. Rabiner and R.W. Schafer, ”Digital Processing of Speech Signals”, Pearson Education, Delhi, India, 2004
- J. R. Deller, Jr., J. H. L. Hansen and J. G. Proakis, ”Discrete-Time Processing of Speech Signals”, Wiley-IEEE Press, NY, USA, 1999.
- D. O'Shaughnessy, ”Speech Communications: Human and Machine”, Second Edition,University Press, 2005.
- Dong Yu, Li Deng, ”Automatic Speech Recognition: A Deep Learning Approach”, Springer, 2015

Outline

- Introduction
- Speech Processing : Human vs Computing Machine
- Speech Recognition : Human vs Automatic
- Traditional Framework for Automatic Speech Recognition
- Speech Analysis
- Feature Extraction
- Deep Learning Framework for Automatic Speech Recognition
- Representation learning
- Handcrafted vs representation learning
- Summary

Introduction

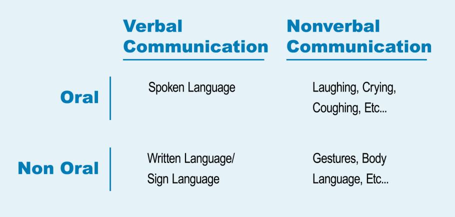
- Speech processing is the study of speech signals and associated methods for processing them.

- Extract and model information from speech signals

- Information: Message, language, speaker, emotion, health, etc

- Task: Speech recognition, language identification, speaker recognition, emotion recognition, health condition recognition, etc

Human - Human Communication



	Verbal Communication	Nonverbal Communication
Oral	Spoken Language	Laughing, Crying, Coughing, Etc...
Non Oral	Written Language/ Sign Language	Gestures, Body Language, Etc...

Figure: Verbal vs Non-Verbal Communication[1]

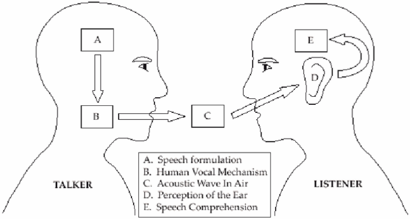
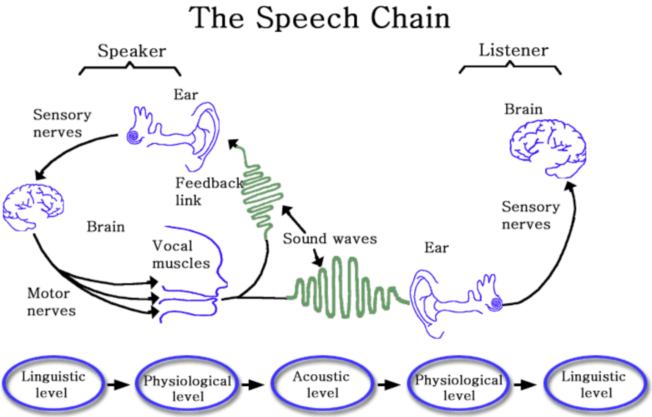


Figure: Speech production, transmission, perception, comprehension[2]

1.[https://mytext.cnm.edu/lesson/5-1-0-defining-verbal-communication/]

2.[https://towardsdatascience.com/speech-recognition-is-hard-part-1-258e813b6eb7]

Human Speech Communication Chain



[http://indra-bohara.blogspot.com/2010/10/brief-critical-review-of-speech-chain.html]

Speech Processing vs Communication

- Speech Signal: Electrical Communication vs Speech Processing
- Communication $\Longrightarrow$ Exchanging information w/o looking what is inside or opaque.
- Digitization and compression in electrical communication $\Longrightarrow$ speech as correlated signal
- Extracting and modeling information in speech processing

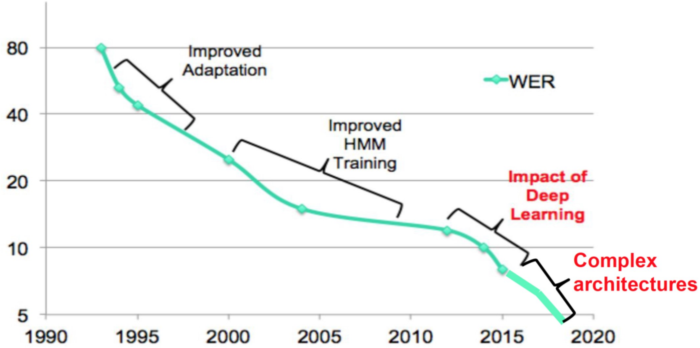
Speech Processing: Human vs Computing Machine

- Acoustic to mechanical to electrical in human ears.
- Electrical: bio-evoked potential on auditory nerve.
- Human cognitive system is good at modeling information in speech.
- Computing machine is trying to mimic these activities for decades.
- Computing machine approaches based on pattern recognition
- Pattern recognition through machine learning and deep learning (DL)
- Latest trends using deep learning in most tasks.

Speech Processing: Deep Learning vs Earlier

- Data Driven : More data, complex models, more computing (S/W, H/W) infrastructure, better performance.
- Domain Knowledge : Not mandatory hence proliferation of speechtech startups and companies. Domain to Domain agnostic
- S/W & H/W Requirements : Mostly open source toolkits. GPU infra on rent.
- Industry vs Academia Data driven vs domain
- Way Forward : Collaborate, share, mentor.

Automatic Speech Recognition Trends

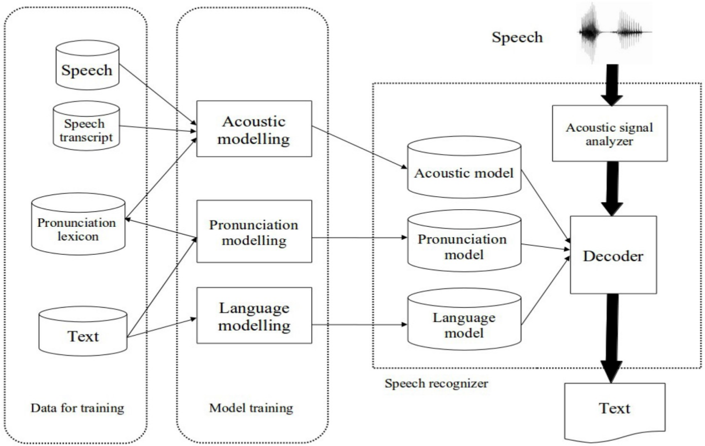


[taken from public domain]

Speech Recognition: Human vs Automatic

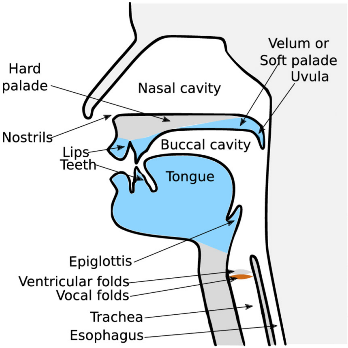
- Cognitive vs computing
- Spoken vs written language
- Human Speech Recognition exploits only spoken language.
- Labelled speech database, dictionary, language models.
- Mobile networks and internet makes life easy.
- Deep learning provides models that can learn features
- Transfer learning, end-to-end system.
- Build speech recognition exploiting more spoken language cues.
- Domain to domain agnostic

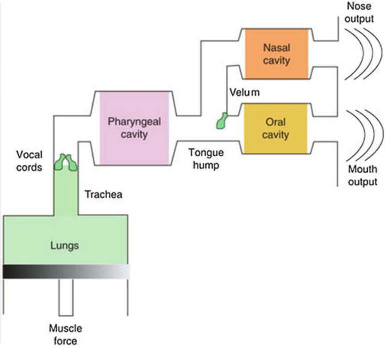
Traditional Framework for Automatic Speech Recognition



[taken from public domain]

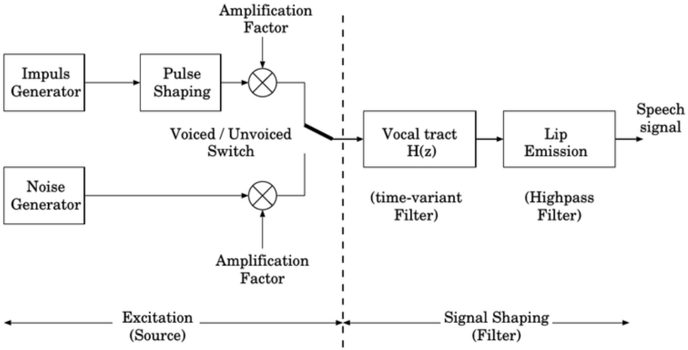
Feature Extraction for Acoustic Modeling





[taken from public domain]

Feature Extraction for Acoustic Modeling

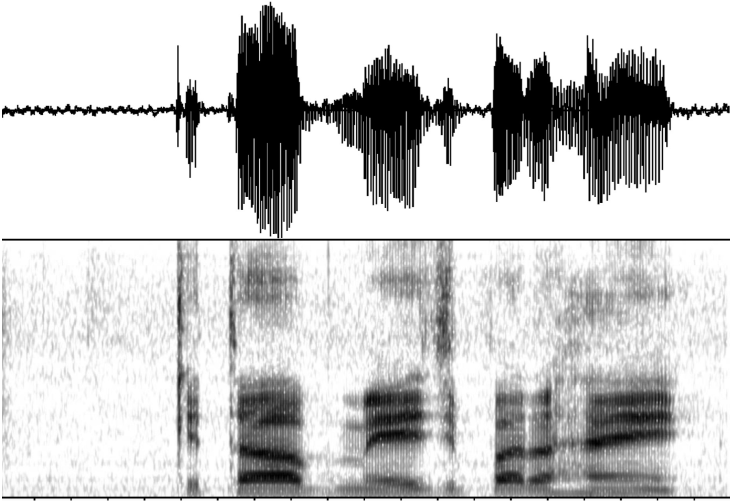


[taken from public domain]

Speech Analysis

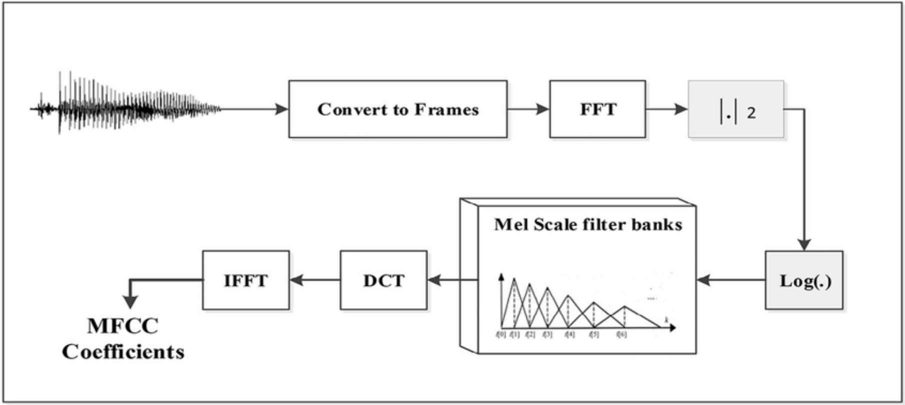
- Non-Stationary : Short term processing (10-30 ms)
- Time Domain : Amplitude variation as a function of time.
- Frequency Domain : Amplitude variation as a function of frequency (Spectrum).
- Vocal tract information as feature vectors for speech recognition.
- Spectrogram : Amplitude variation as a function of time and frequency.

Speech Analysis



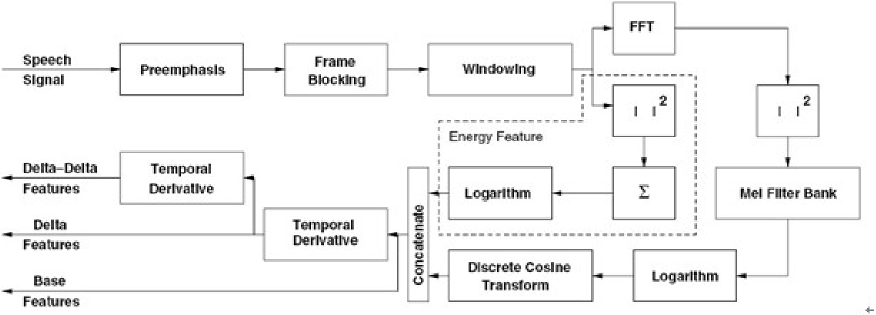
[taken from public domain]

Mel Frequency Cepstral Coefficients (MFCCs)



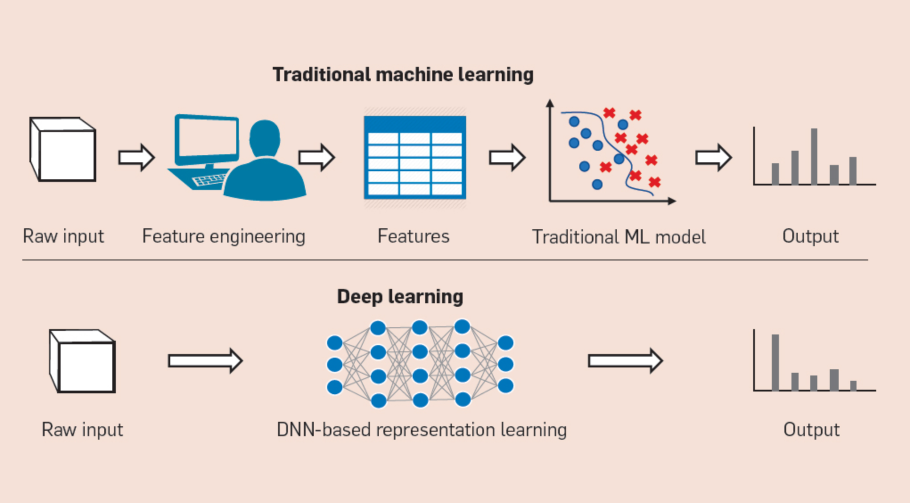
[taken from public domain]

Delta, Delta-Delta MFCCs



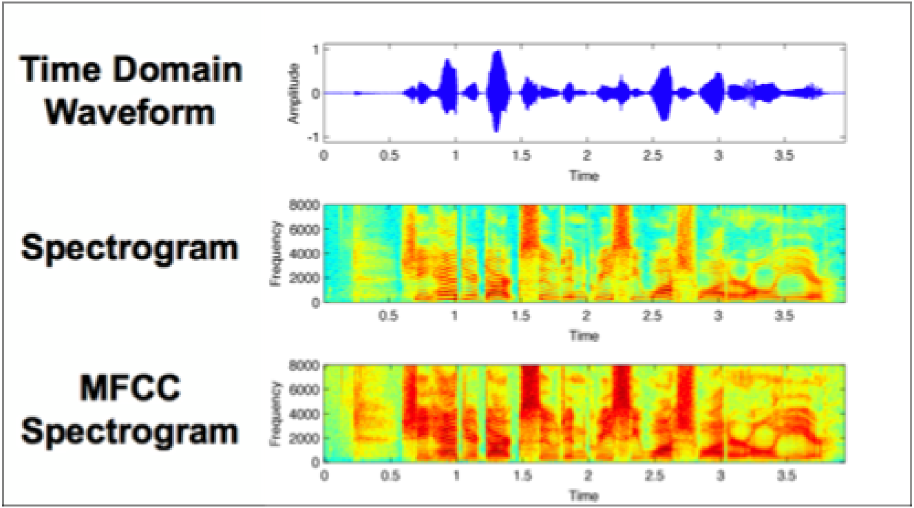
[taken from public domain]

Traditional ML vs DL



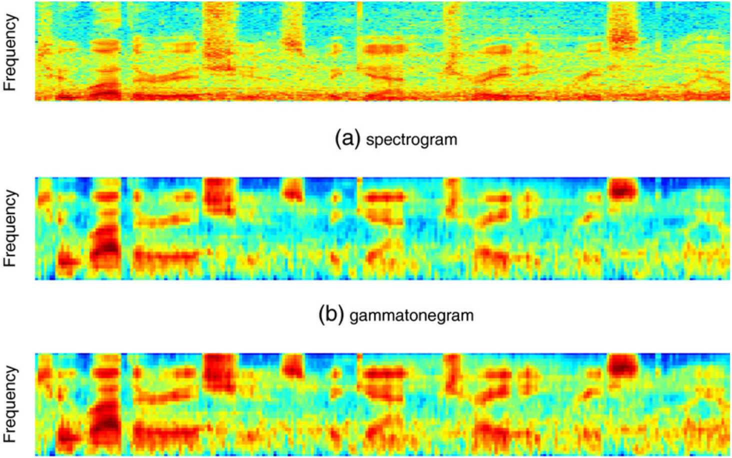
[taken from public domain]

Spectrogram vs MelSpectrogram



[taken from public domain]

Spectrogram vs Gammatone Spectrogram



(a) spectrogram

(b) gammatonegram

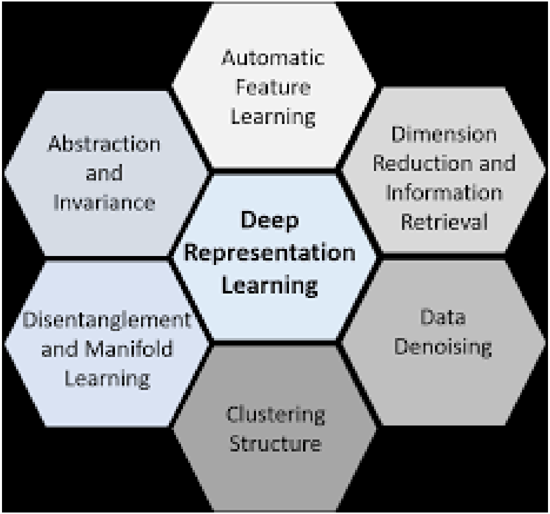
(c) gammatonegram (after non linearity)

[taken from public domain]

Deep Learning based Expert System

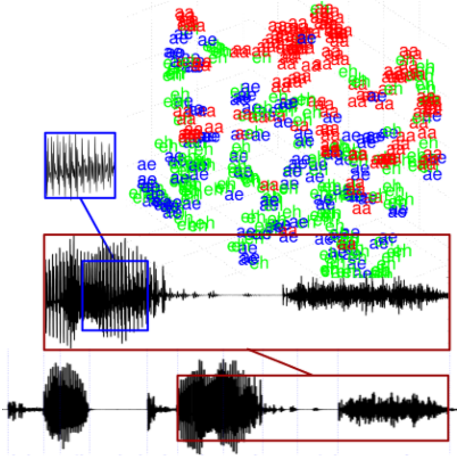
- Expert System:
  - Human expert experience is coded as set of rules.
  - Humans are spectrogram reading experts
- Deep Learning based expert system:
  - Deep learning models derive representation and then recognize patterns.

Deep Representation Learning


Automatic Feature Learning
Abstraction and Invariance
Dimension Reduction and Information Retrieval
Deep Representation Learning
Disentanglement and Manifold Learning
Data Denoising
Clustering Structure

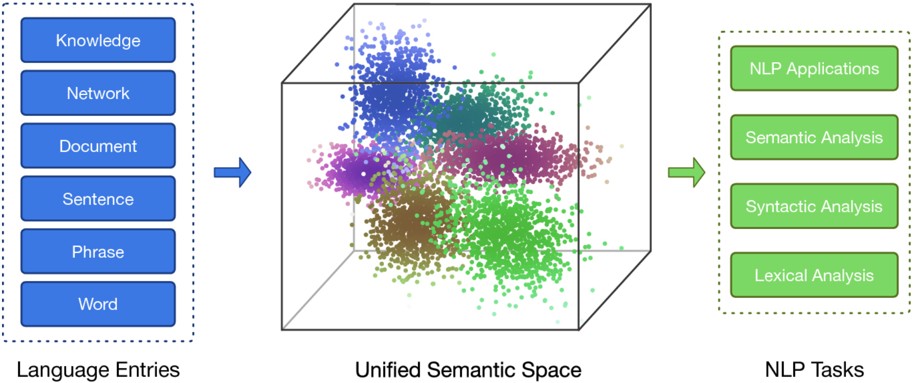
[taken from public domain]

Invariant Representation Learning



[taken from public domain]

Universal Acoustic Space



Knowledge

Network

Document

Sentence

Phrase

Word

Language Entries

Unified Semantic Space

NLP Applications

Semantic Analysis

Syntactic Analysis

Lexical Analysis

NLP Tasks

[taken from public domain]

Summary

- Introduction to speech processing
- Human approach for speech processing
- Handcrafted features and machine learning for speech processing
- Representation learning and deep learning for speech processing
- Way forward for feature extraction

Thank You