Speech Processing: Handcrafted to Deep Representations

S. R. M. Prasanna

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

prasanna@iitdh.ac.in

॥ सा विद्या या विमुक्तये ॥

Some References

- Siddique Latif, et al., "Deep Representation Learning in Speech Processing: Challenges, Recent Advances, and Future Trends", http://arxiv.org/abs/2001.00378v1
- Y. Bengio, A. Courville, and P. Vincent "Representation Learning: A Review and New Perspectives", IEEE Trans. on Software Engineering, August 2013.
- Dong Yu, Li Deng, "Automatic Speech Recognition: A Deep Learning Approach", Springer, 2015
- 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.

Outline

- Introduction
- Speech Processing : Human vs Computing Machine
- Speech Processing :
  - Time domain and Frequency domain processing
  - Cepstral and linear prediction analysis
  - Time-Frequency domain processing
    - Spectrogram, Filterbank energies, Modulation spectrum
- Representation learning: NN and Deep Learning for feature extraction
- Handcrafted vs representation learning
- Summary

Motivation: Feature Engineering vs Representation Learning

- Weakness of ML algorithms is their inability to extract and organize the discriminative information from the data.
- Feature engineering is a way to take advantage of human ingenuity and prior knowledge to compensate for that weakness.
- To expand the scope and ease of applicability of machine learning, make learning algorithms less dependent on feature engineering.
- Novel applications could be constructed faster, and make progress towards Artificial Intelligence (AI).

Motivation

- Speech processing $\implies$ Designing hand crafted acoustic features + designing efficient machine learning (ML) models to make prediction and classification decisions.
- Drawbacks
  - Manual feature engineering is cumbersome and needs domain knowledge.
  - Designed features might not be best for the objective at hand.
- Motivation for representation learning, learn intermediate representation of speech automatically that better suits the task and hence improved performance.
- Part 1 Traditional or hand crafted features
- Part 2 Representation learning

Part I: Traditional or Hand Crafted Features

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

	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]

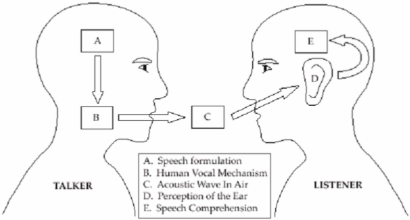
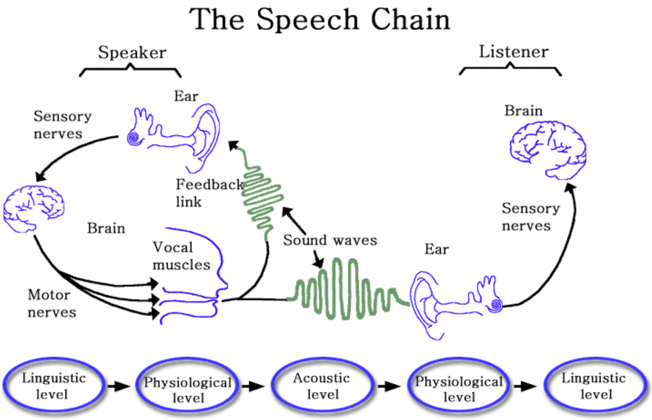


Figure: Speech production, transmission, perception, comprehension[2]

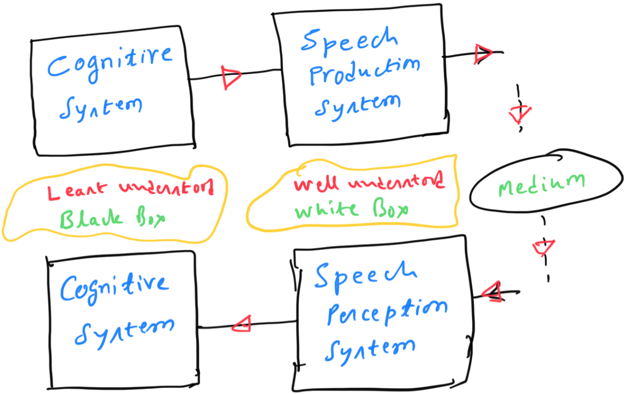
1.[https://mytext.cnm.edu/lesson/5-1-0-defining-verbal-communication/]
2.[https://towardsdatascience.com/speech-recognition-is-hard-part-1-258e813b6eb7]

Human Speech Communication Chain



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

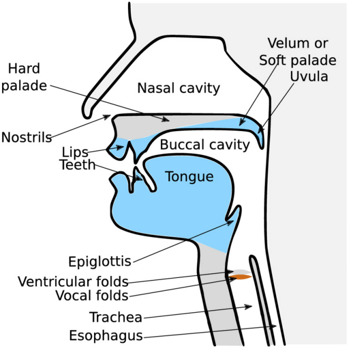
Feature Extraction for Speech Processing

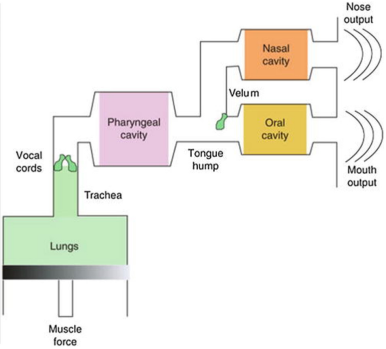


Speech Processing: Deep Learning vs Earlier

- 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 : Open source toolkits. GPU infra.
- Industry vs Academia Data driven vs domain
- Data driven vs domain may complement each other
- Part I: Better knowledge about feature extraction may help in better understanding and interpretation of DL systems
- Part II: Data driven approach may yield better features and hence improved performance

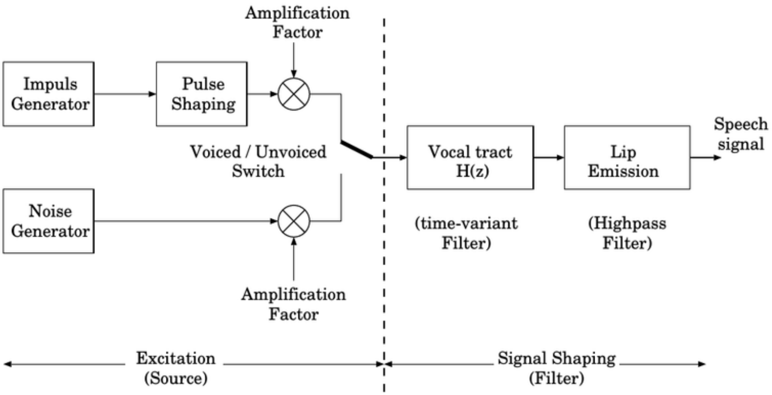
Speech Production Process





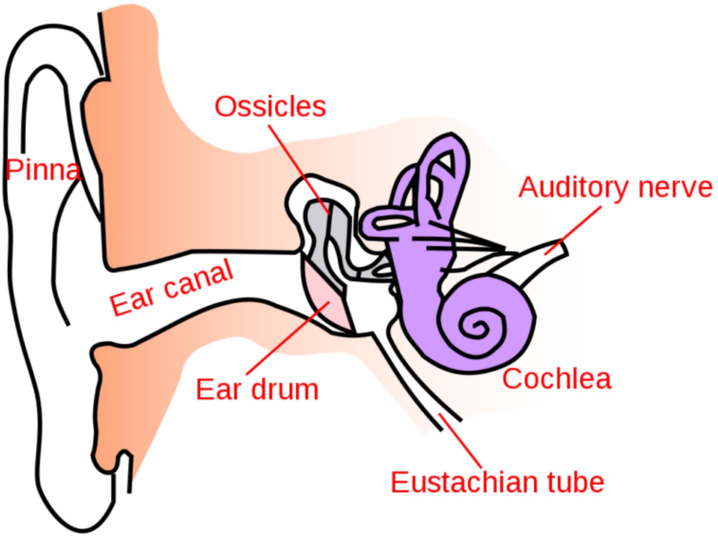
[Taken from public domain and copyright belongs to original authors]

Two State Speech Production Model



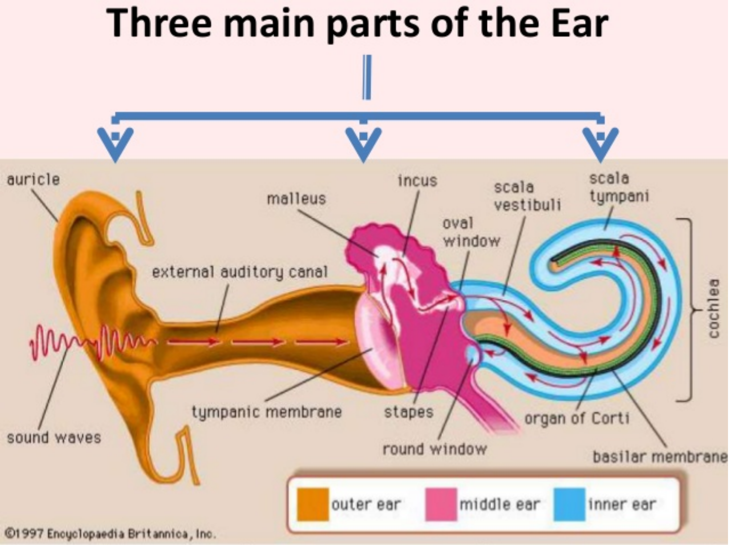
[Taken from public domain and copyright belongs to original authors]

Speech Perception Process: Ear



[Taken from public domain and copyright belongs to original authors]

Speech Perception Process: Acoustic to BEP



[Taken from public domain and copyright belongs to original authors]

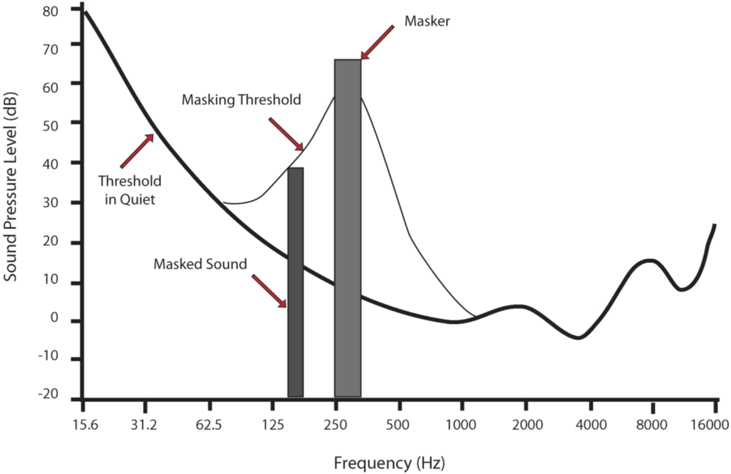
Speech Processing: Human vs Computing Machine

- Acoustic to mechanical to electrical in human ears.
- Electrical: bio-evoked potential on auditory nerve.
- Human ear is good at processing speech signal.
- Bio-evoked potential (1D) to spatial representation (2D) ?
- Human cognitive system is good at modeling information in speech.
- Computing machine is trying to mimic these activities for decades.
- Approaches based on signals processing & pattern recognition
- Pattern recognition through machine learning and deep learning (DL)
- Latest trends using deep learning in most tasks.

Speech Processing: Psychoacoustics

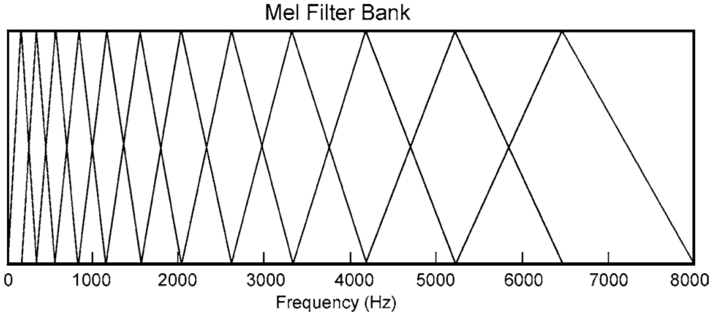
- Scientific study of sound perception.
- Ear and the brain are involved in a person's listening experience.
- Human ear is good at processing speech signal.
- Inner ear does significant signal processing in converting sound waveforms into neural stimuli.
- Certain differences between waveforms may be imperceptible.
- In addition, the ear has a nonlinear response to sounds of different intensity levels; this nonlinear response is called loudness.

Speech Perception Process: Masking Phenomenon



[Taken from public domain (wikipedia) and copyright belongs to original authors]

Speech Perception Process: Mel Filterbank



[Taken from public domain and copyright belongs to original authors]

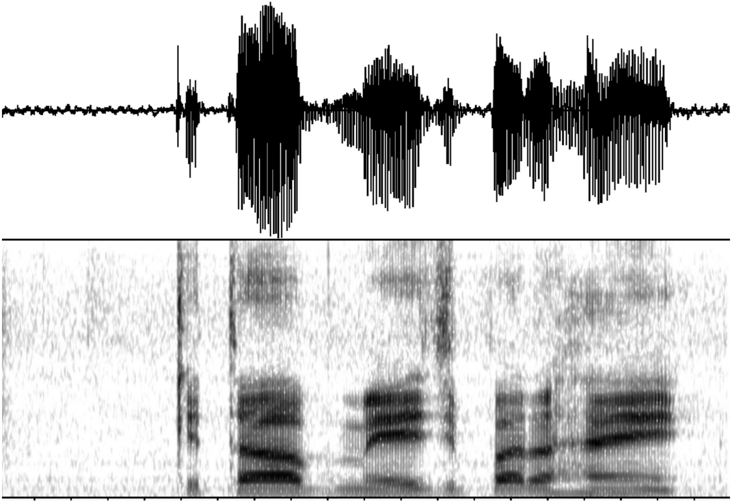
Nature of Speech Signal

- One dimension non-stationary with multiple information sources
- Information sources at different levels: Subsegmental, segmental and suprasegmental.
- Subsegmental: 1-3 ms, naturalness, closed phase region, burst, ...
- Segmental: 10-30 ms vocal tract shape, excitation source (pitch, shimmer, jitter), ...
- Suprasegmental: $>$ 100 ms pitch contour, voiced/univoiced/silence regions, energy contour, loudness, tone, duration, ...
- Cognitive system exploiting information from all levels for modelling
- Further based on selective attention

Speech Analysis

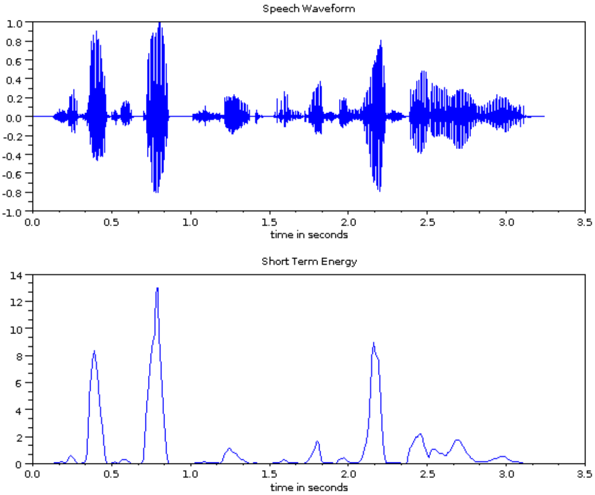
- Time Domain : Amplitude variation as a function of time.
- Frequency Domain : Amplitude vs frequency (spectrum).
- Time-Frequency Domains: Amplitude vs time and frequency.
- Vocal tract information as feature vectors for speech recognition.
- Spectrogram : Amplitude vs time and frequency.
- Excitation Source information Voiced, unvoiced, pitch, gci, rmfcc, mpdss, mpdss as features.
- Phase Information : Group delay, Hilbert phase, instantaneous phase
- Modulation Features : Modulation functions
- Representation Learning : Neural network and deep learning approaches

Speech Analysis: Waveform vs Spectrogram



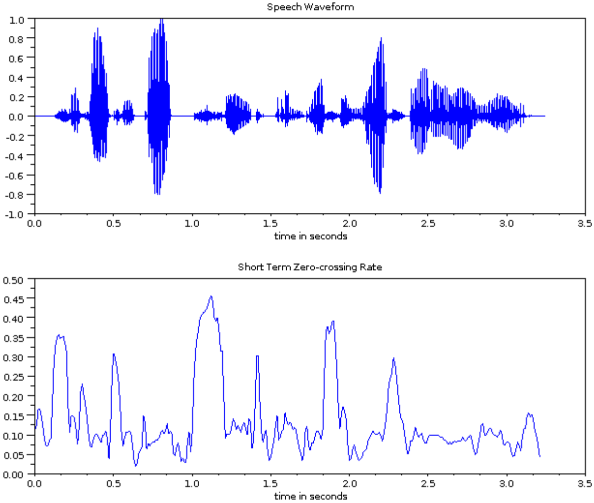
[taken from public domain]

Time Domain Speech Analysis: Short term Energy



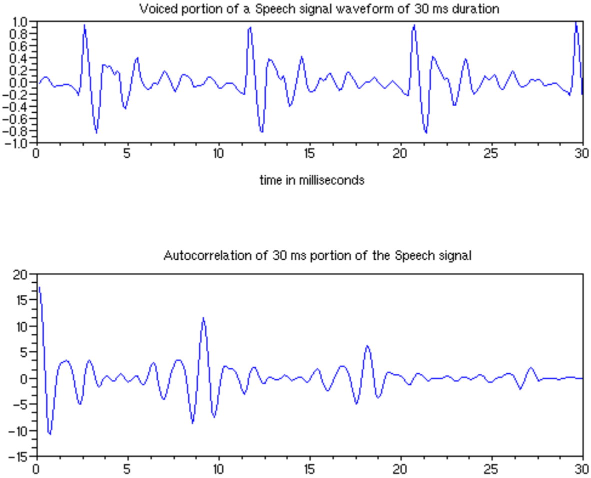
[taken from public domain]

Time Domain Speech Analysis: Short term ZCR



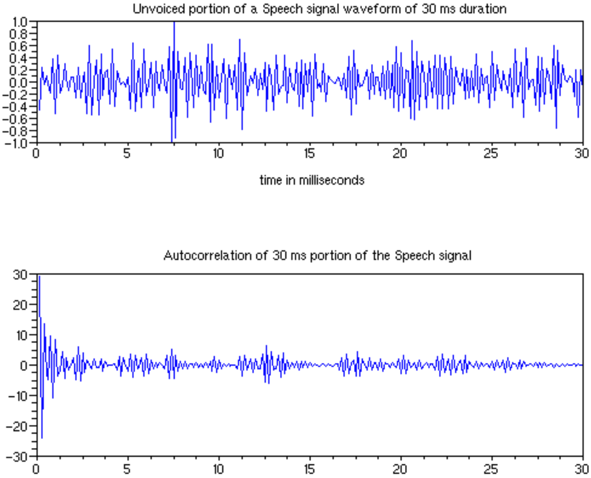
[taken from public domain]

Time Domain Speech Analysis: Short term ACR Voiced



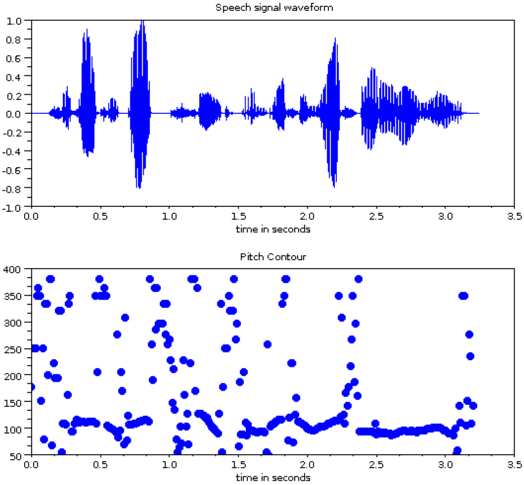
[taken from public domain]

Time Domain Speech Analysis: Short term ACR Unvoiced



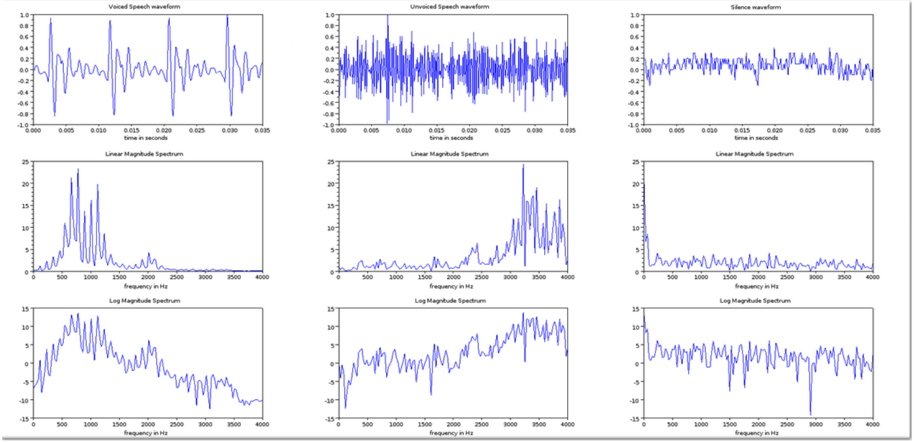
[taken from public domain]

Time Domain Speech Analysis: Short term Pitch



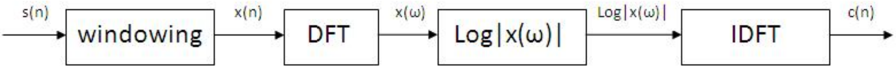
[taken from public domain]

Frequency Domain Speech Analysis: Spectra



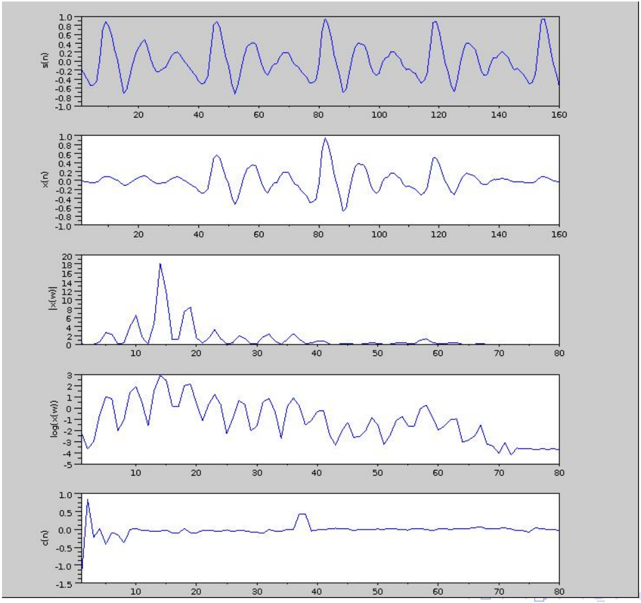
[taken from public domain]

Cepstrum Analysis

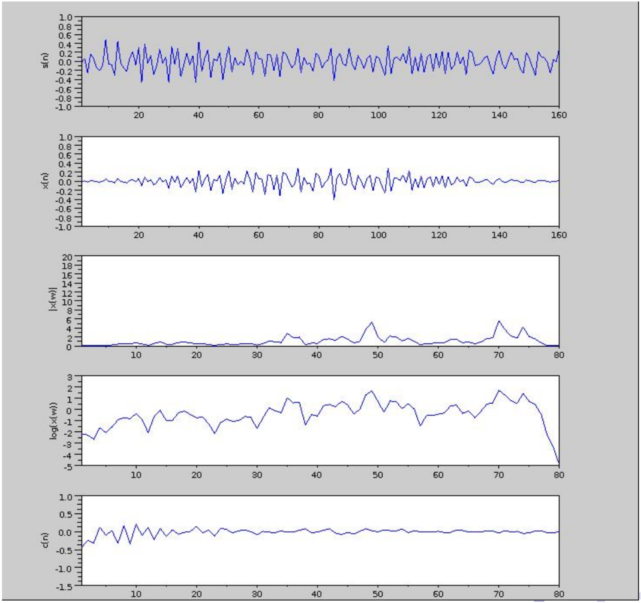


[taken from public domain]

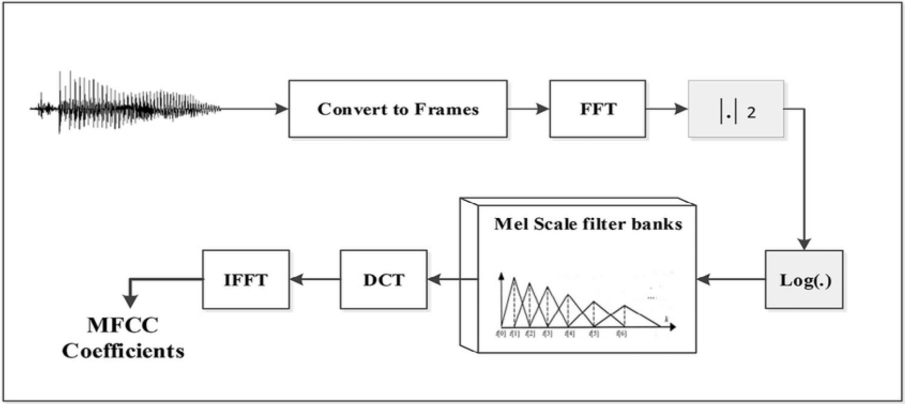
Cepstrum Analysis: Voiced Speech



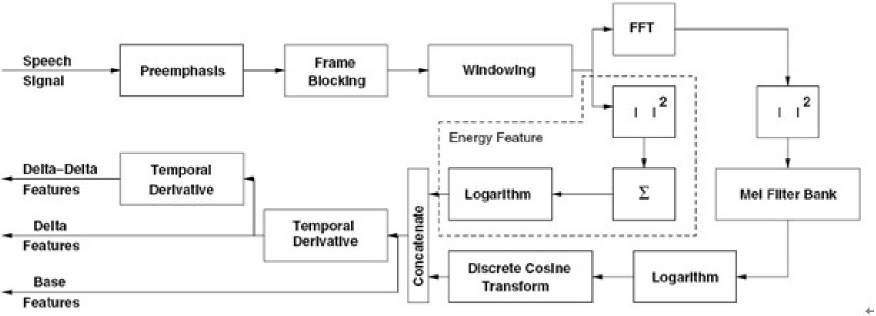
Cepstrum Analysis: Unvoiced Speech



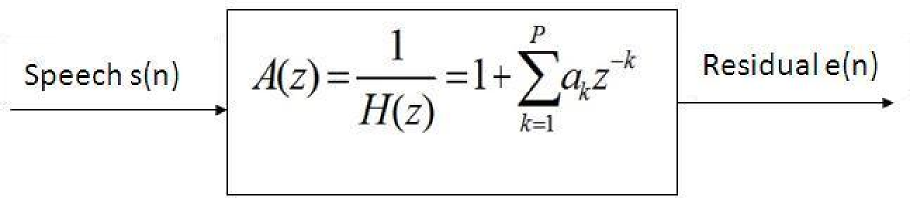
Cepstrum Analysis: Mel Frequency Cepstral Coefficients



[taken from public domain]

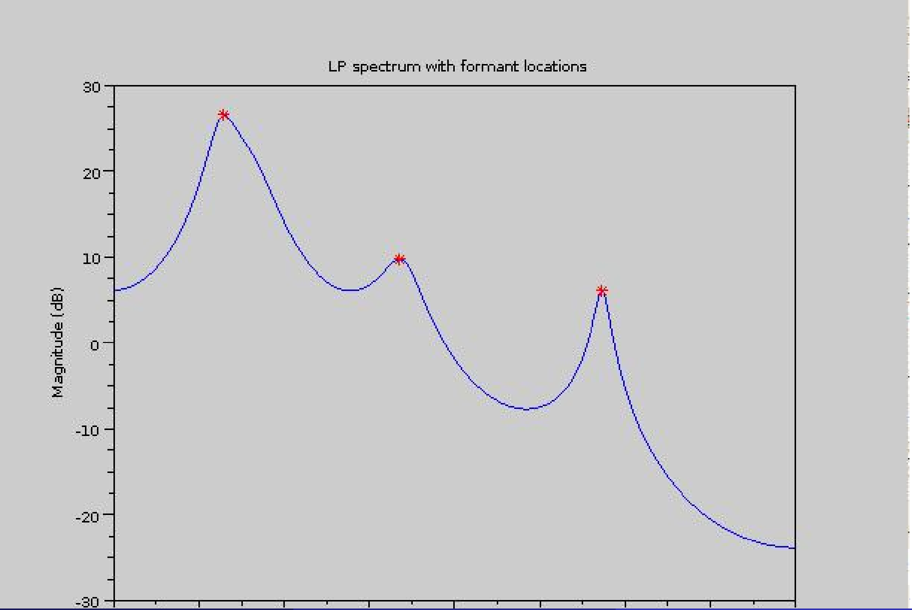


[taken from public domain]

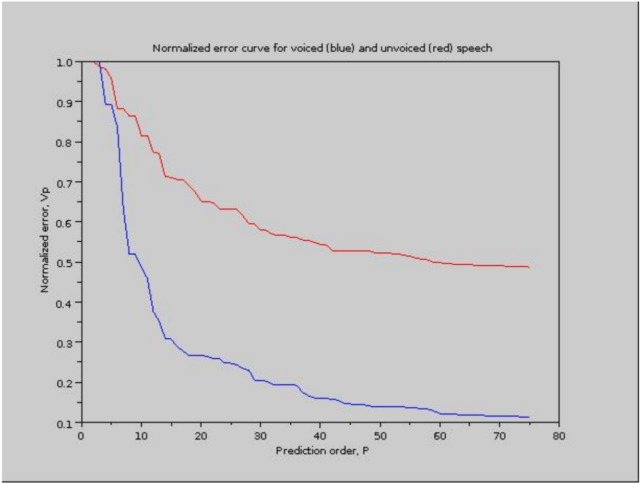


[taken from public domain]

Linear Prediction Analysis: LP Spectrum

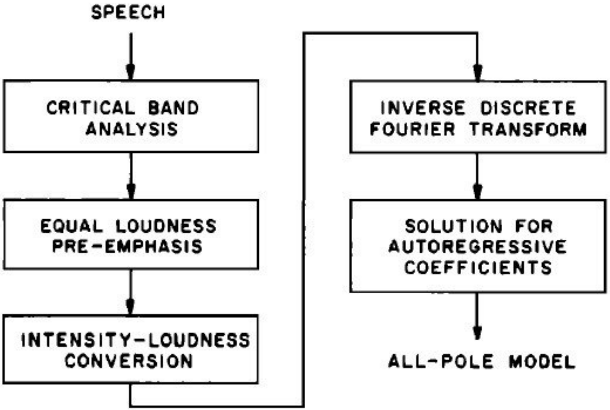


Linear Prediction Analysis: Norm Error



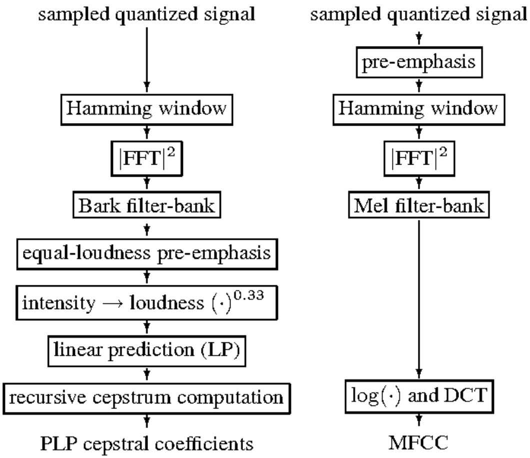
[taken from public domain]

Perceptual Linear Prediction Analysis



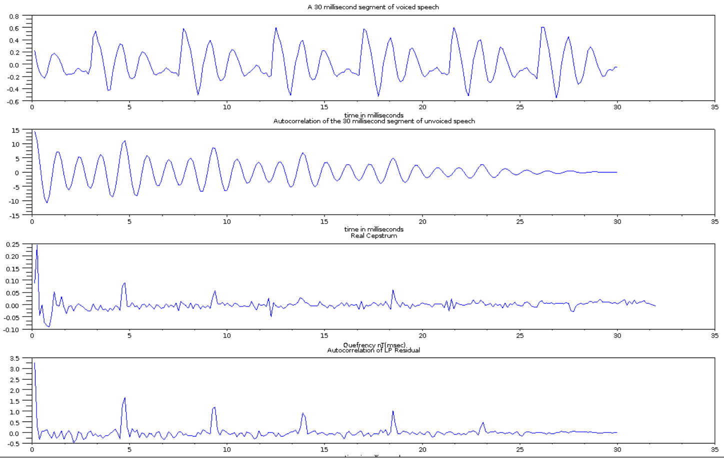
[taken from public domain]

Perceptual Linear Prediction Analysis: PLPCC



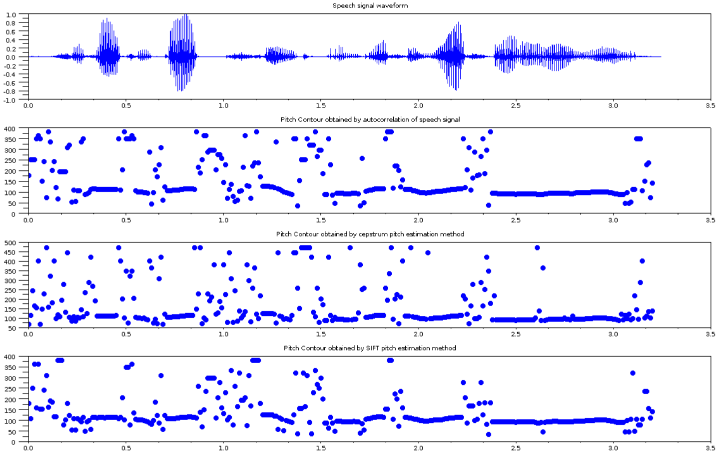
[taken from public domain]

Estimation of Pitch



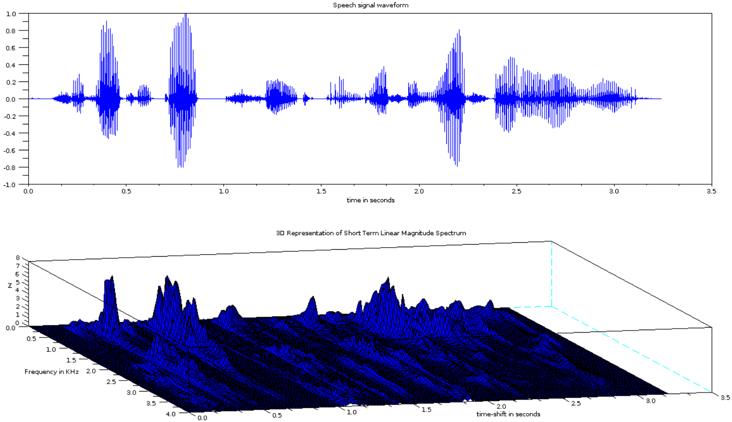
[taken from public domain]

Estimation of Pitch Contours



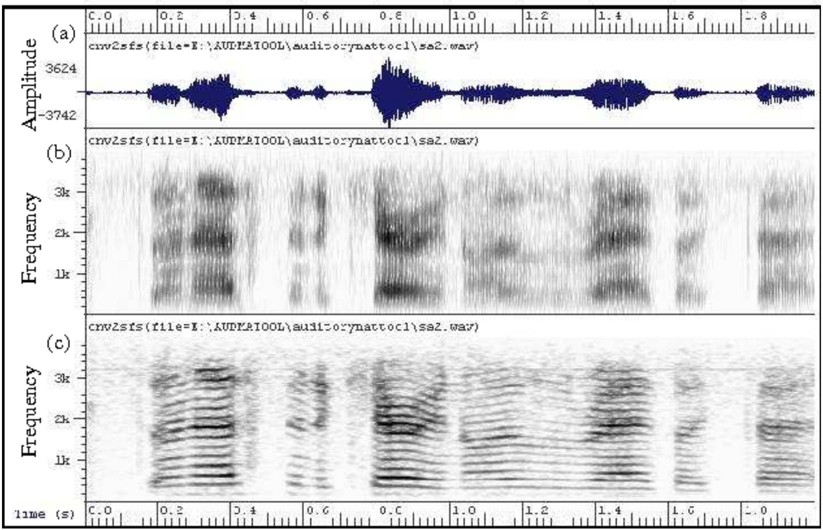
[taken from public domain]

Time-Frequency Analysis: Short term Fourier Transform



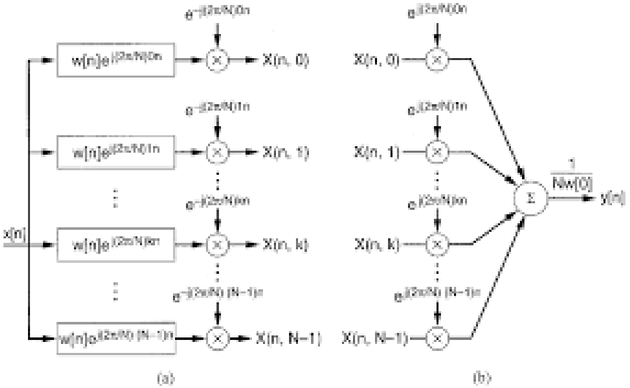
[taken from public domain]

Time-Frequency Analysis: Spectrogram



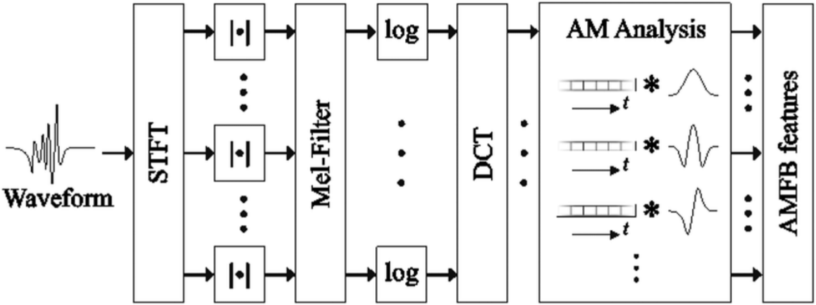
[taken from public domain]

Time-Frequency Analysis: STFT Filterbank

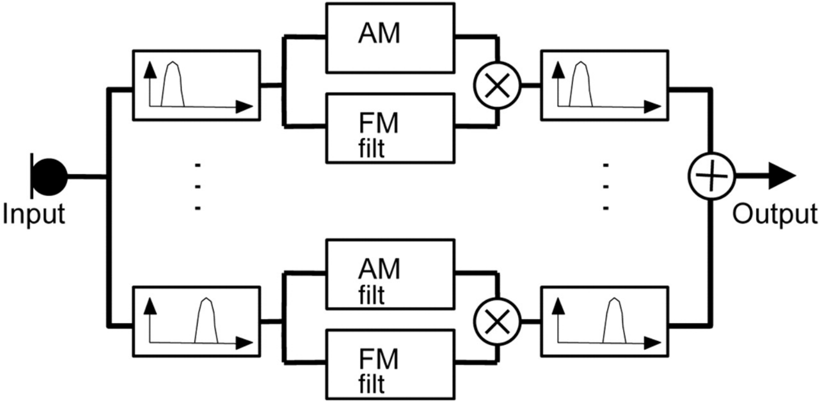


[taken from public domain]

Time-Frequency Analysis: AM Filterbank



[taken from public domain]



[taken from public domain]

Part II: Representation Learning

Representation Learning: Motivation

- Technique of learning representations of input data that yields in abstract and useful representations.
- What to learn? both image (spectrogram) and sequence informn.
- Traditionally efficiency of ML algorithms in speech processing relied heavily on the quality of hand-crafted features.
- A good set of features often leads to better performance compared to a poor speech feature set.
- Feature engineering led to lots of research and has been an important field for a long time.
- DL models can learn feature representation automatically and thereby give better performance.

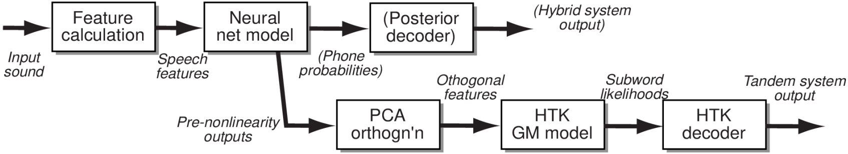
Representation Learning: Linear Ones

- PCA & LDA Traditional or shallow feature learning algorithms.
- Both PCA and LDA are linear data transformation techniques.
- LDA is a supervised method that requires class labels to maximise class separability.

Representation Learning: Non-Linear Ones

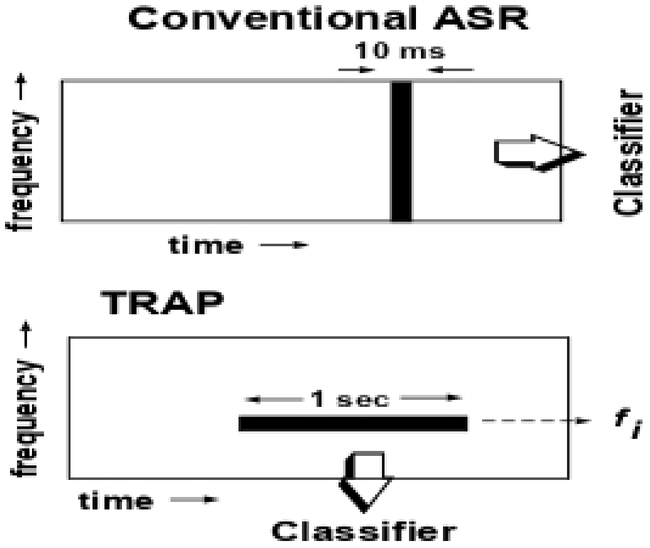
- Kernel version of some linear feature mapping algorithms like kernel PCA (KPCA) and generalised discriminant analysis (GDA).
- Non-negative Matrix Factorisation (NMF).
- Neural newtworks.
- In contrast to kernel based methods, non-linear feature representation algorithms like neural networks directly learn the mapping functions.
- Traditional representation algorithms have been widely used for transforming the speech representations to more informative features.

Representation Learning using NN: Tandem Features



[taken from public domain]

Representation Learning using NN: TRAP Features



Representation Learning using NN: Trap Features

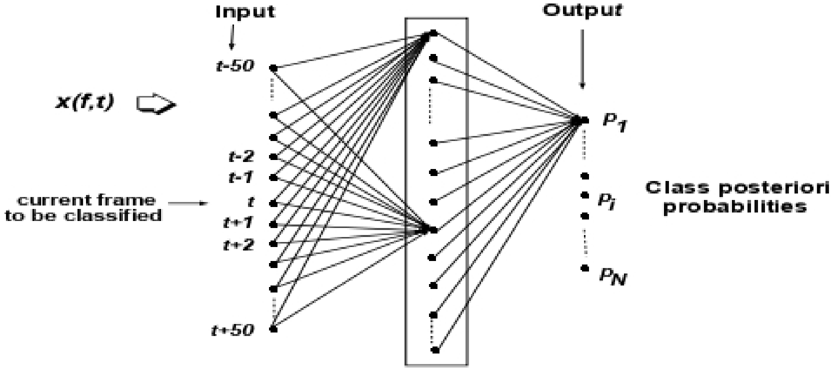


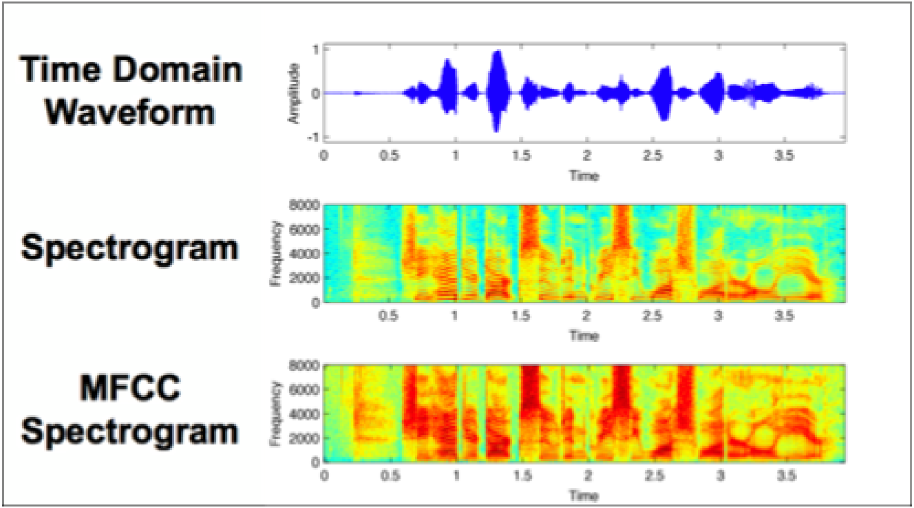
Figure 2: Neural TRAP

[taken from public domain]

Representation Learning: Features from Speech for DL

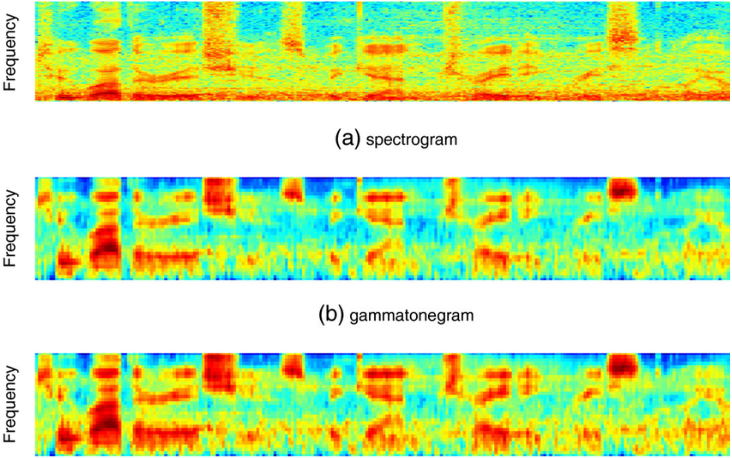
- In speech analysis tasks, deep models for representation learning can either be applied to speech features or directly on the raw waveform.
- Log-Mel spectrum is the most popular feature to train DL networks.
- Widely used spectrogram for CNNs due to their image like configuration.
- Log-Mel spectrogram is another speech representation that provides a compact representation.

Spectrogram vs MelSpectrogram



[taken from public domain]

Spectrogram vs Gammatone Spectrogram



(a) spectrogram

(b) gammatonegram

(c) gammatonegram (after non linearity)

[taken from public domain]

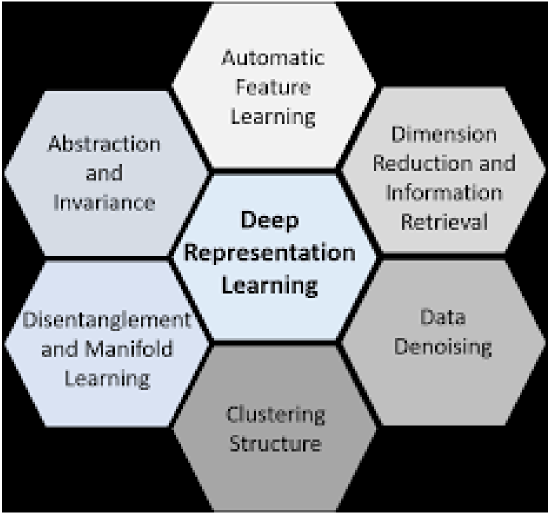
Representation Learning: Supervised Deep Learning

- DNNs outperformed GMM-HMM due to their ability to learn a hierarchy of representations from input data.
- Recurrent neural networks (RNNs) architectures including long-short term memory (LSTM) and gated recurrent units (GRUs) outperformed DNNs.
- Superior performance of RNN architectures was because of their ability to capture temporal contexts from speech.
- A cascade of CNNs, LSTM and DNNs layers were further shown to outperform LSTM only models.

Representation Learning: Unsupervised Deep Learning

- Lack of labelled data set the pace for the unsupervised representation learning research.
- For unsupervised representation from speech, AEs, RBMs, and DBNs were widely used.
- Significant interest in three classes of generative models including VAEs, GANs, and deep auto-regressive models.

Deep Representation Learning



[taken from public domain]

Representation Learning: Automatic Feature Learning

- Process of constructing explanatory variables or features that can be used for classification or prediction problems.
- Feature learning algorithms can be supervised or unsupervised.
- DL models are composed of multiple hidden layers and each layer provides a kind of representation of the given data.
- Automatically learnt feature representations are – given enough training data – usually more efficient and repeatable than hand-crafted features.
- Automatically learnt feature representation is more flexible and powerful.

Representation Learning: Dimension Reduction and Information Retrieval

- To eliminate data redundancy and irrelevancy.
- To make data more understandable and interpretable.
- Very difficult to analyse high dimensional data with a limited number of training samples.
- Information retrieval is finding information based on a user query.
- Finding a suitable representation of a query is a challenging task and DL based representation playing an important role.
- Representation learning models for information retrieval can learn features automatically with little or no prior knowledge.

Representation Learning: Data Denoising

- To deal with noisy conditions, one often performs data augmentation by adding artificially-noised examples to the training set.
- However, data augmentation may not help always, because the distribution of noise is not always known.
- In contrast, representation learning methods can be effectively utilised to learn noise robust features learning.
- They often provide better results compared to data augmentation.
- In addition, the speech can be denoised such as by DL based speech enhancement systems.

Representation Learning: Clustering Structure

- Clustering aims to categorise similar classes of data samples into one cluster using similarity measures (e. g., Euclidean distance).
- A large number of data clustering techniques have been proposed.
- Classical clustering methods usually have poor performance on high dimensional data, and suffer from high computational complexity on large-scale datasets.
- In contrast, DL based clustering methods can process large and high dimensional data with a reasonable time complexity and they have emerged as effective tools for clustering structures.

Representation Learning: Disentanglement Learning

- Method that disentangles or represents each feature into narrowly defined variables and encodes them as separate dimensions.
- Differs from feature extraction or dimensionality reduction techniques as it explicitly aims to learn such representations that aligns axes with the generative factors of the input data.
- Practically, data is generated from independent factors of variation.
- Disentangled representation learning aims to capture these factors by different independent variables in the representation.
- In this way, latent variables are interpretable, generalisable, and robust against adversarial attacks.

Representation Learning: Manifold Learning

- Manifold learning aims to describe data as low-dimensional manifolds embedded in high-dimensional spaces.
- Manifold learning can retain a meaningful structure in very low dimensions compared to linear dimension reduction methods.
- Manifold learning algorithms attempt to describe the high dimensional data as a non-linear function of fewer underlying parameters by preserving the intrinsic geometry.
- Such parameters have a widespread application in pattern recognition, speech analysis, and computer vision.

Representation Learning: Abstraction and Invariance

- Architecture of DNNs is inspired by hierarchical structure of brain.
- Thus deep architectures might capture abstract representations.
- Equivalent to discovering a universal model that can be across all tasks to facilitate generalisation and knowledge transfer.
- Abstract features are generally invariant to the local changes and are non-linear functions of the raw input.
- Abstraction representations capture high-level continuous-valued attributes that are robust .
- Learning invariant features has more predictive power which has always been required by the AI community.

Representation Learning Approaches

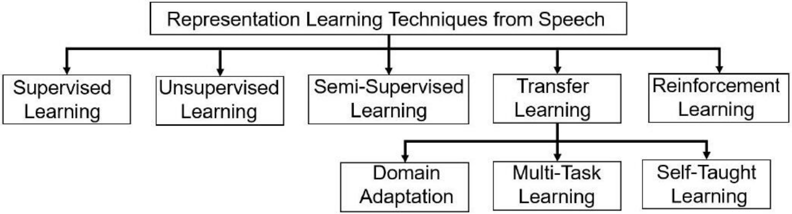


Fig. 3: Representation Learning Techniques.

[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 use for prediction or classification.

Way Forward for Representation Learning

- Demonstrated the significance of DL for representation learning.
- Training data, training complexity, optimization and tuning complexity.
- Field of non-linear signal processing with very good performance.
- Repeatibility and interpretabilty needs to be looked into.
- Dump data to DL model vs nonlinear speech processing model to interpret what is learnt.
- Cognitive system exploiting information from all levels for modelling
- Further based on selective attention

Summary

- Introduction to speech processing
- Human approach for speech processing
- Handcrafted features and representation learning for speech processing
- Different aspects of representation learning
- Way forward for feature extraction

Thank You