Laurent.Besacier@imag.fr
GETALP group leader

Grenoble Informatics
Laboratory (France)

Study Group for Machine
Translation and Automated
Processing of Languages and Speech

Paramètres Acoustiques, Reconnaissance et Alignement Automatiques

Tutorial on Automatic Speech Recognition

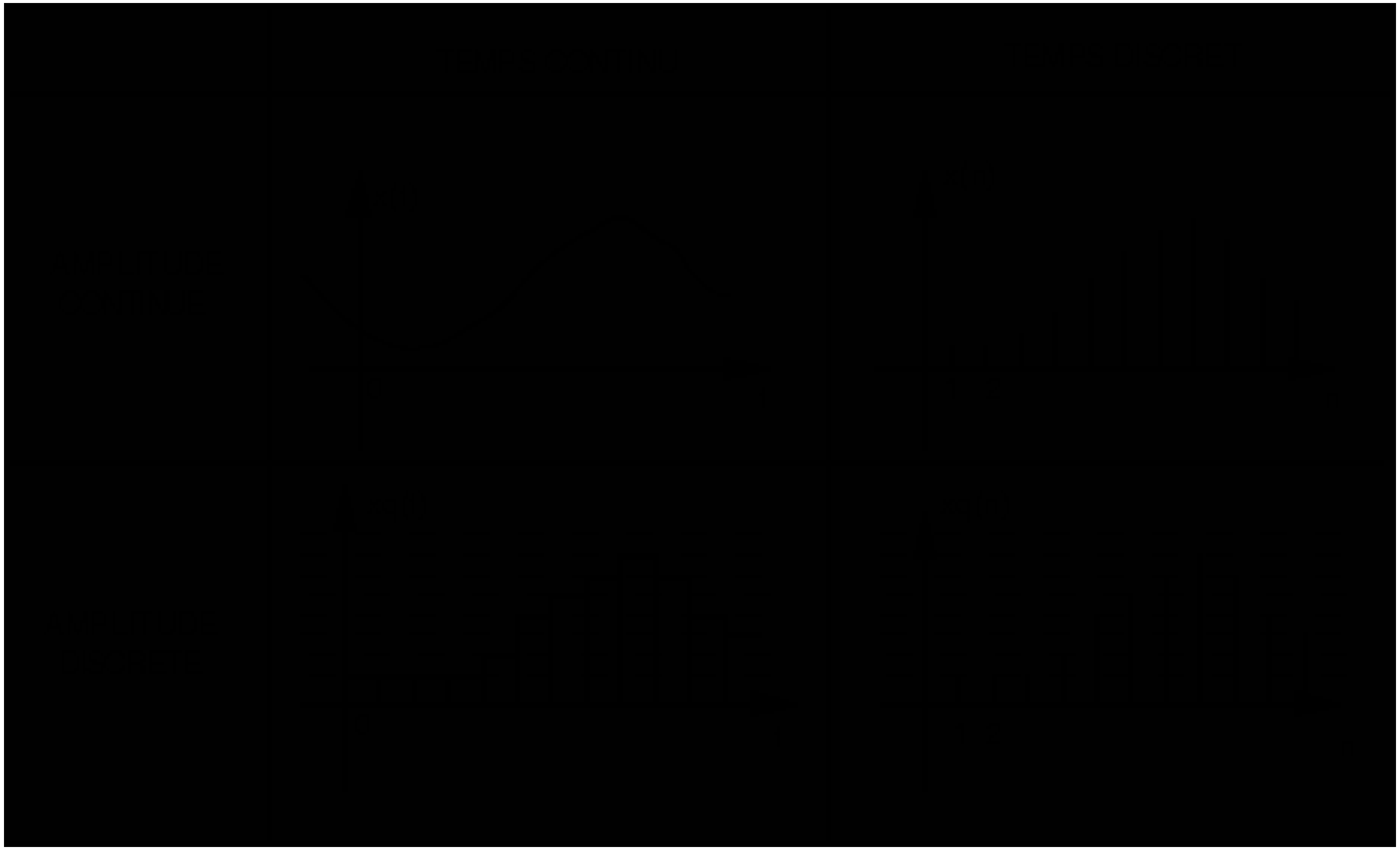
Outline

- Signal processing reminder
- The speech signal (not treated here)
- Automatic Speech Recognition (ASR)
  - Overview
  - Speech modelling (parameters, models)
  - Toolkits for ASR design

Outline

- **Signal processing reminder**
- The speech signal (not treated here)
- Automatic Speech Recognition (ASR)
  - Overview
  - Speech modelling (parameters, models)
  - Toolkits for ASR design

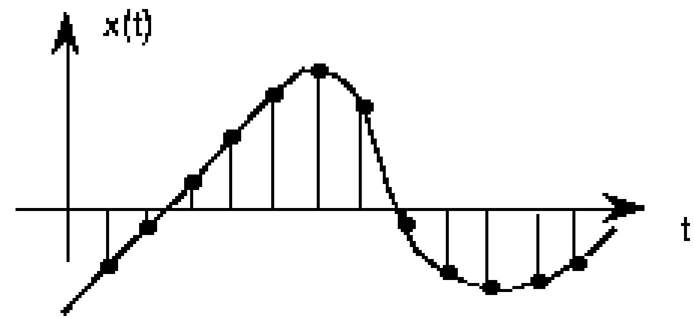
Digital / Analogic Signals



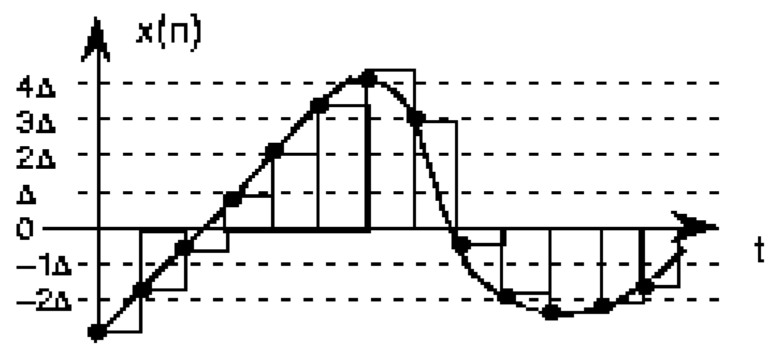
Analog-Digital Conversion

- **Sampling** and **Quantification**

sampling



quantification



SNR : Signal-to-Noise Ratio

- For $x(t)=s(t)+n(t)$

$$SNR=\frac{W_s}{W_n}$$

$$SNR_{dB}=10\log_{10}SNR$$

$$SNR_{dB}=20\log_{10}\frac{Amplitude_s}{Amplitude_n}$$

Fourier Transform - TF

- Spectral representation of signals
- Core mathematical tool in DSP

TF for continuous signals

- x(t) signal
- TF is a function of variable $\omega = 2\pi f$ defined by :

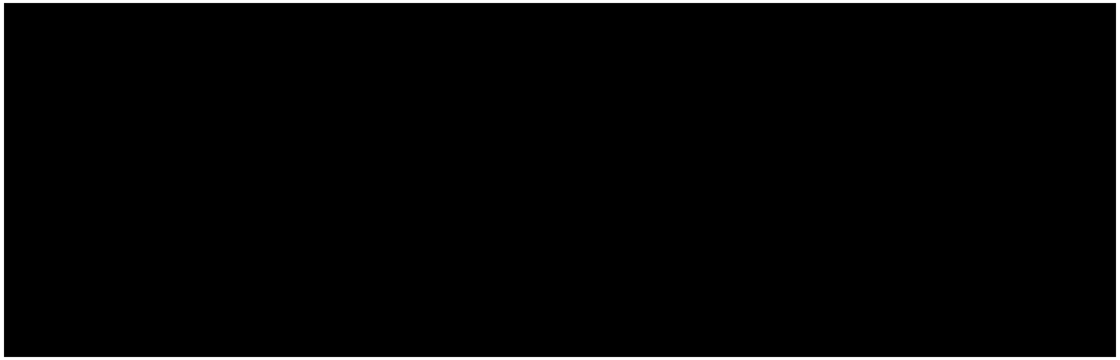
$$F\{x(t)\} = X(\omega) = \int_{-\infty}^{+\infty} x(t) e^{-j\omega t} dt$$

- Inverse transform

$$x(t) = F^{-1}\{X(\omega)\} = \int_{-\infty}^{+\infty} X(\omega) e^{+j\omega t} d\omega$$

TF of a periodic signal (cos)

$$\cos(w_0 t) \overset{TF}{\longleftrightarrow} \frac{1}{2}\left[\delta(w-w_0)+\delta(w+w_0)\right]$$



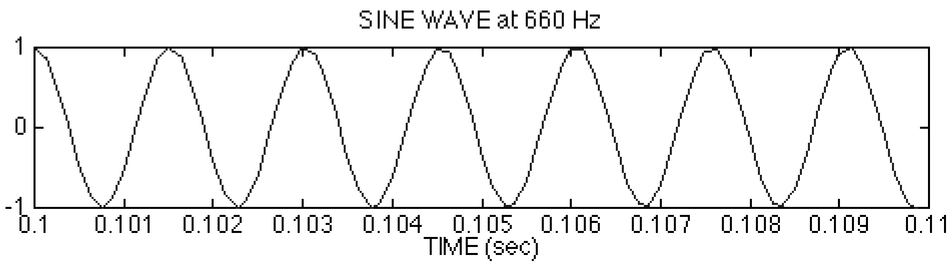
Time-frequency representation

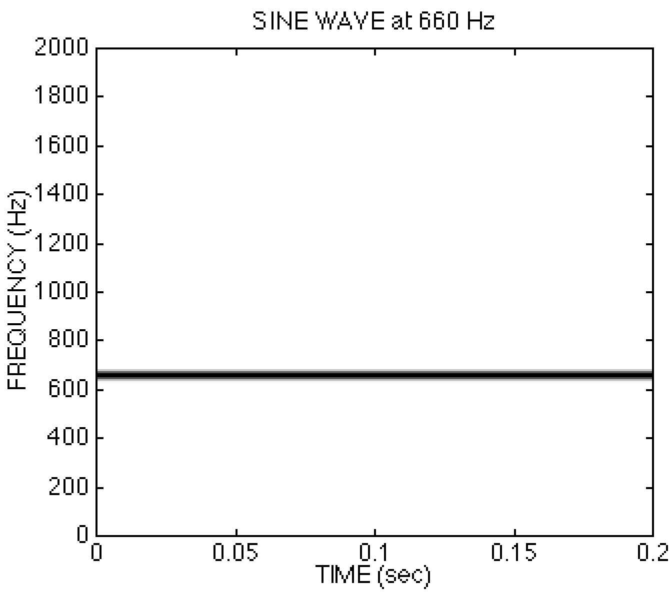
Time-frequency representation

Time-frequency representation Spectrogram

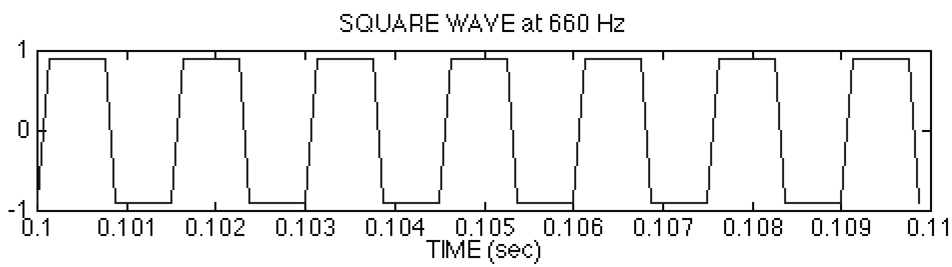
$$S_x(t,f) = \left| \int_{-\infty}^{+\infty} x(s)\, h(s-t)\, e^{-i2\pi fs}\, ds \right|^2$$

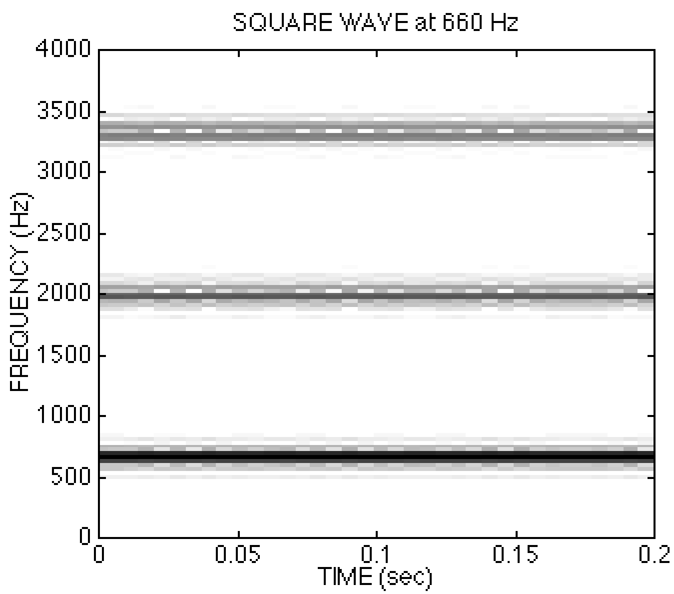
Case of a sinus

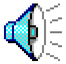




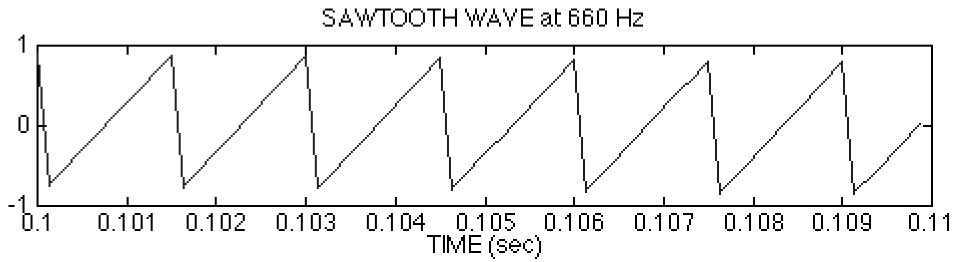
Case of square signal

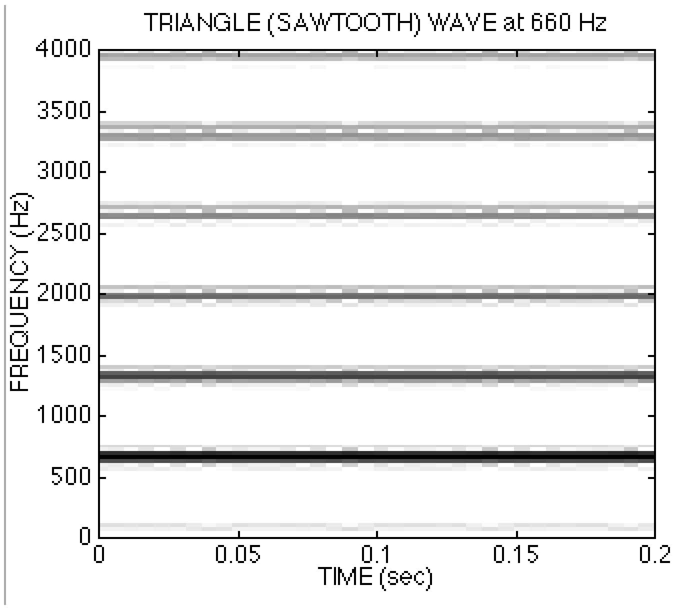




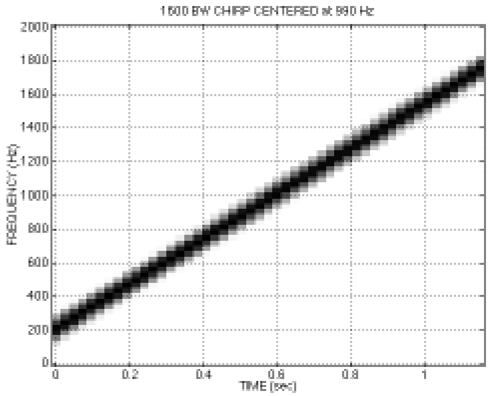


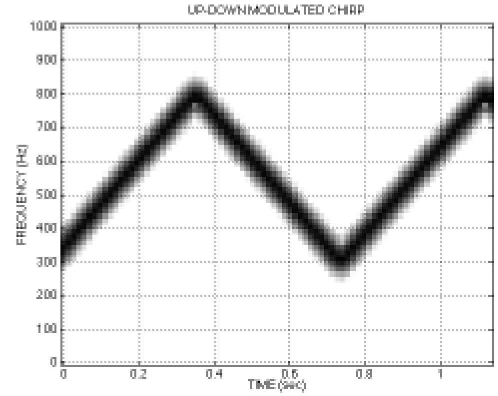
Sawtooth signal

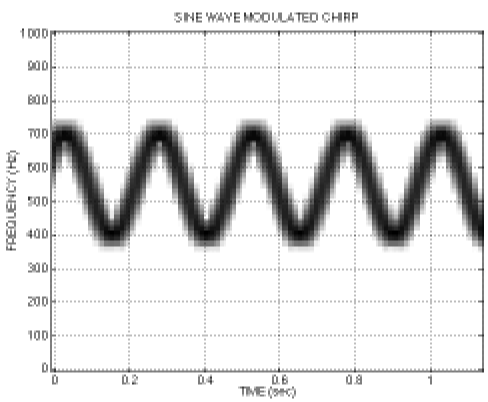


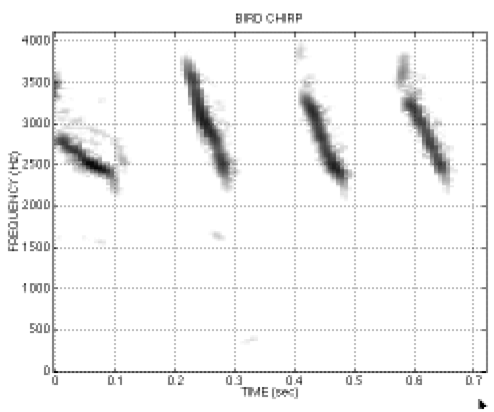


Chirps

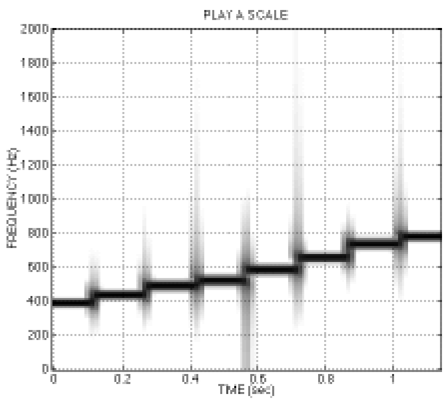

1600 BW CHIRP CENTERED at 990 Hz
FREQUENCY (Hz)
TIME (sec)

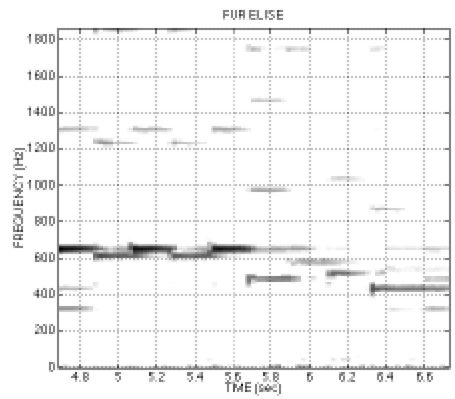

UP-DOWN MODULATED CHIRP
FREQUENCY (Hz)
TIME (sec)

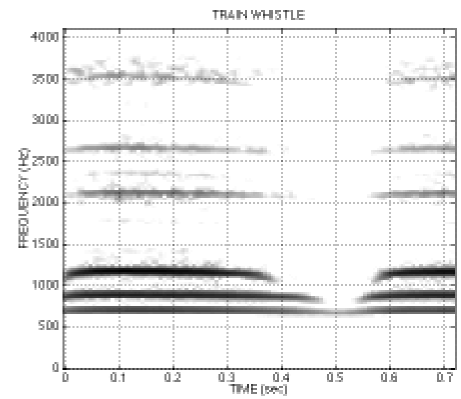

SINE WAVE MODULATED CHIRP
FREQUENCY (Hz)
TIME (sec)

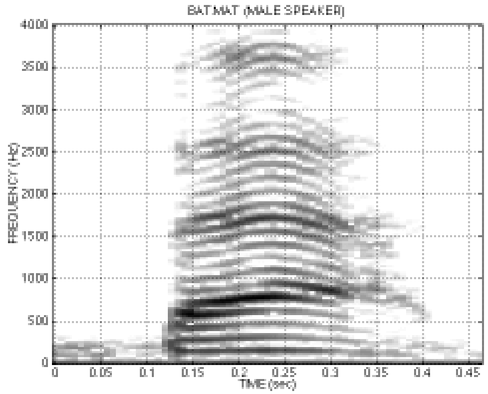

BIRD CHIRP
FREQUENCY (Hz)
TIME (sec)

Other examples









Outline

- Signal processing reminder
- **The speech signal (no treated here!)**
- Automatic Speech Recognition (ASR)
  - Overview
  - Speech modelling (parameters, models)
  - Toolkits for ASR design

Outline

- Signal processing reminder
- The speech signal
- **Automatic Speech Recognition (ASR)**
  - Overview
  - Speech modelling (parameters, models)
  - Toolkits for ASR design

Speech, a source of informations

Speech

Linguistic informations (what is uttered)

Extra-linguistic info. (speaker, language, speaker state)

Different levels of difficulty

- **Number of speakers** : systems mono-speakers …until multi-speakers
- **Vocabulary size**
- **Transmission channel** : «direct mic. », téléphone, mobile phone, VoIP

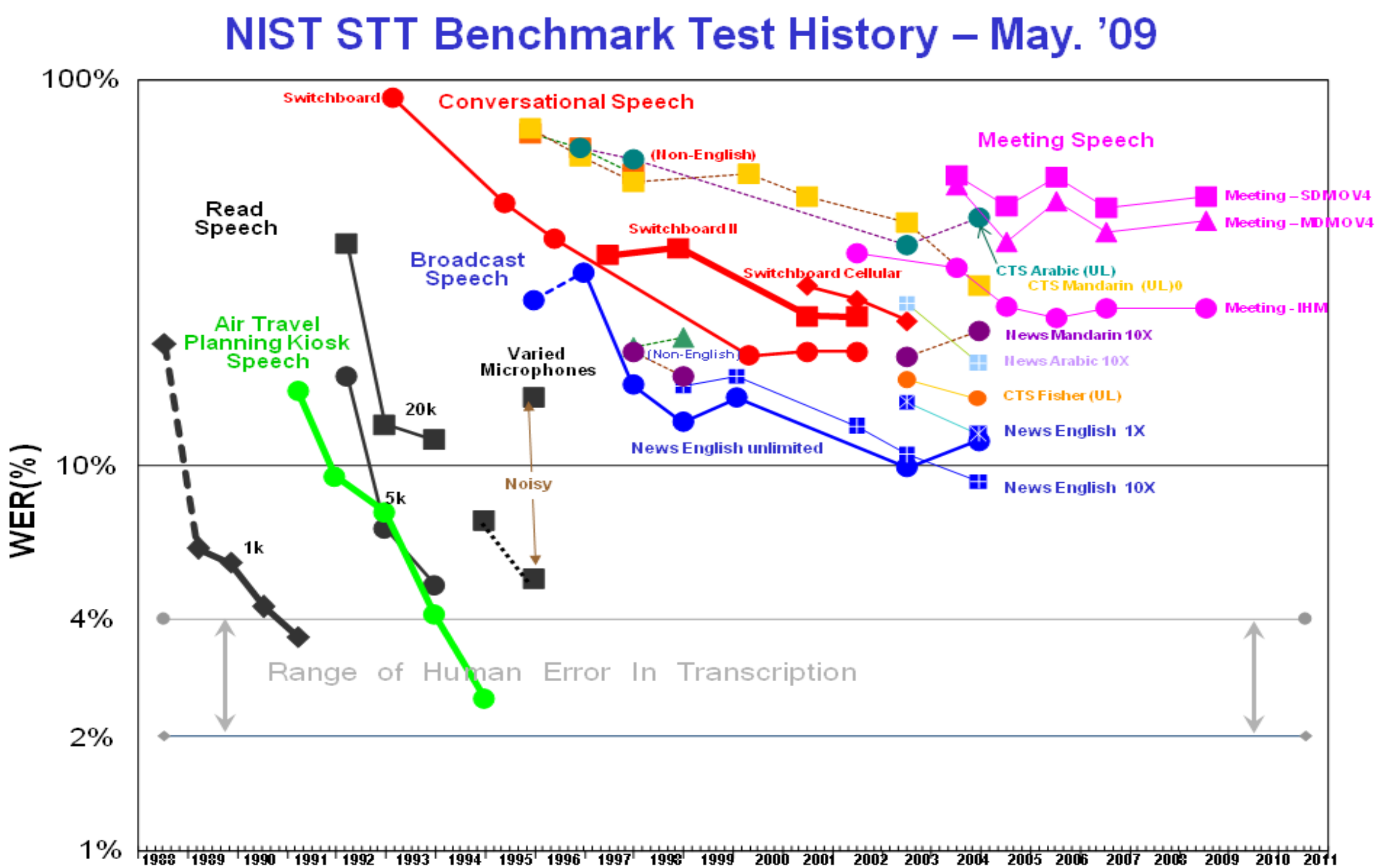
Different levels of difficulty

- **Acoustic Environment :** quiet**,** normal (officeroom), noisy (train station, street), extreme
- **Speaking style** : isolated words, read speech, spontaneous speech
- 1 person or conversation

Applications

- **Services** (vocal servers)
- **Vocal terminals** (on site)
- **Transportation** (vocal commands for navigation system)
- **Language learning**
- **Dictation**
- **Voice search**
- **Control / vocal commands**
- **<u>Personal Assistants</u> (*Siri, Cortana, Echo, Google Now*)**

Where we are…



+ further (big) progresses since 2010 (deep learning approaches)
See http://proceedings.mlr.press/v48/amodei16.pdf

Evolution of the ASR task…

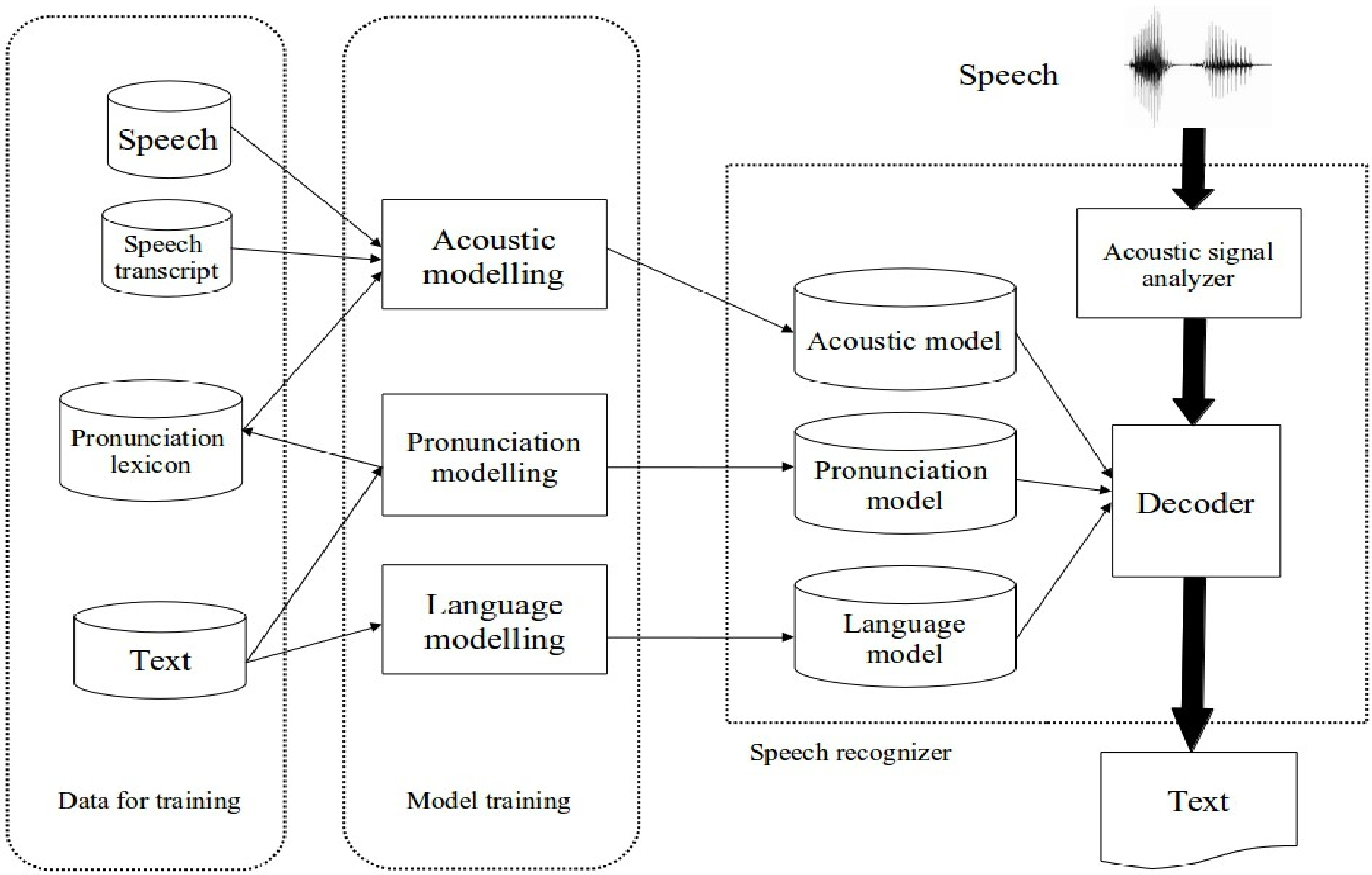
- Evolution of the domain
  - 'Simple' Transcription → Rich Transcription
  - Controlled Audio Stream → Continuous Audio Stream
  - One sensor → Multiple sensors
  - Monolingual → Multilingual
  - Audio only → Multimodal
  - Transcription → Understanding / Dialog
- Increasing difficulty of the tasks

1990	2000	2010	2016
Dictation	***Broadcast news Transcription***	***Meetings Smart rooms***	***Personal Assistants***

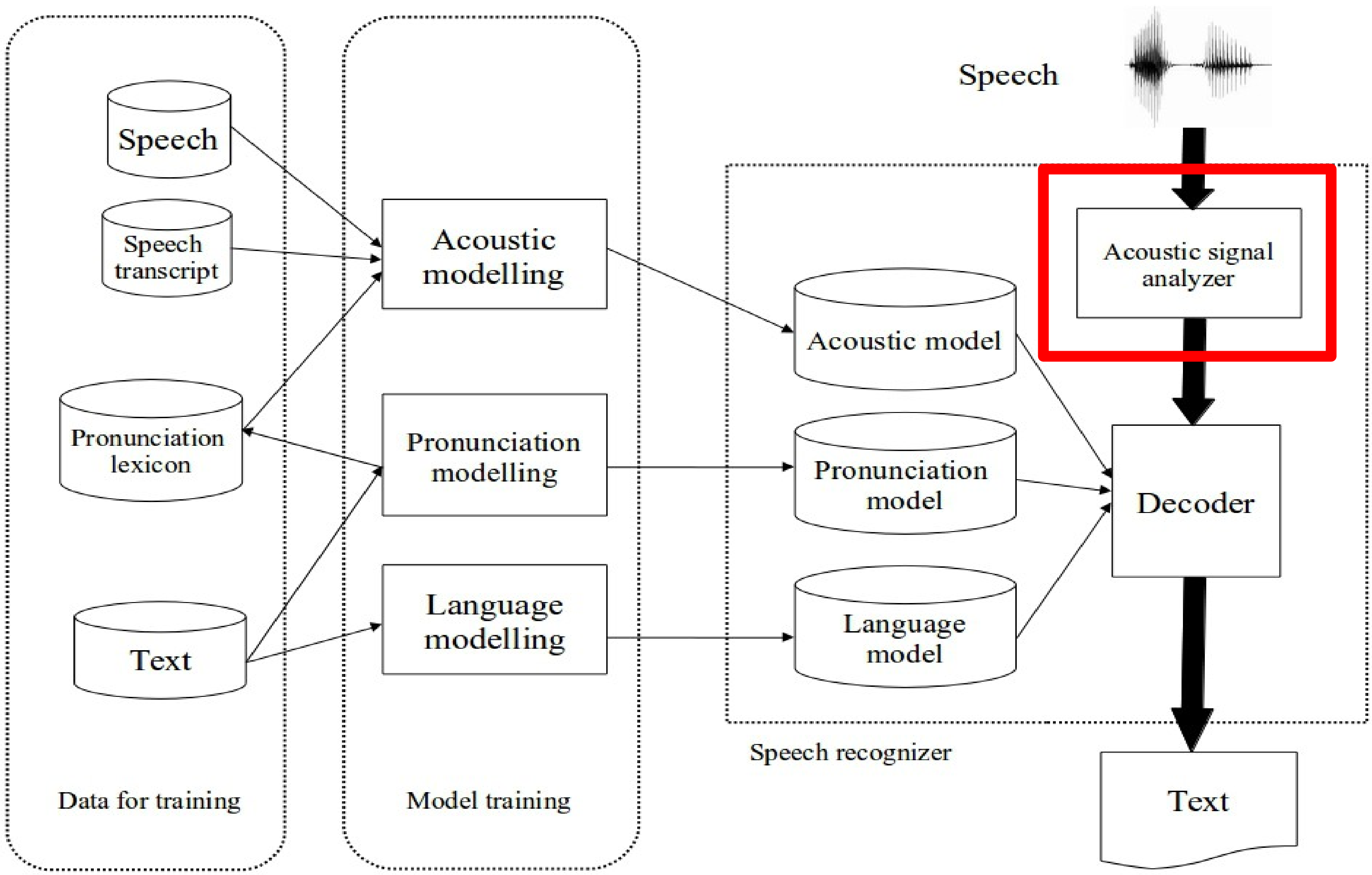
Outline

- Signal processing reminder
- The speech signal
- **Automatic Speech Recognition (ASR)**
  - Overview
  - Speech modelling (parameters, models)
  - Toolkits for ASR design

ASR Systems Overview



ASR Systems Overview

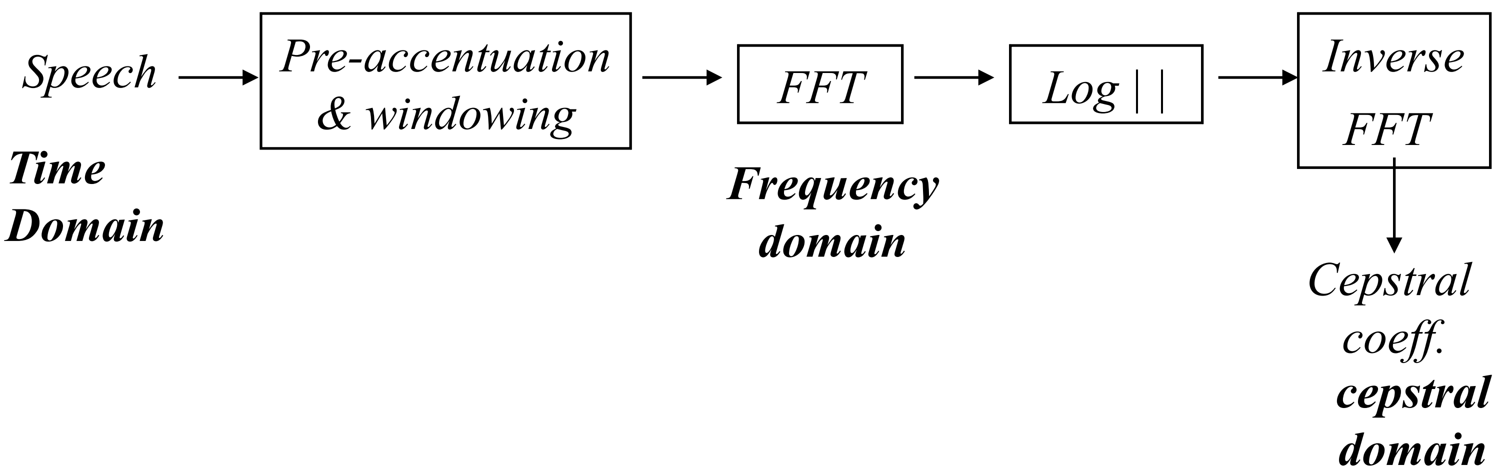


Speech parameters

- Mostly for automatic speech recognition and speech compression
  - Spectral analysis
  - Cepstral analysis
  - Linear prediction
  - Raw signal (new) deep learning approaches
    - Image of the spectrogram (new) deep learning approaches
- Also used
  - Prosodic information (fundamental frequency, energy features, duration)

Acoustic parameters

- Filterbank coefficients : signal energy in different frequency bands
- Cepstral coefficients



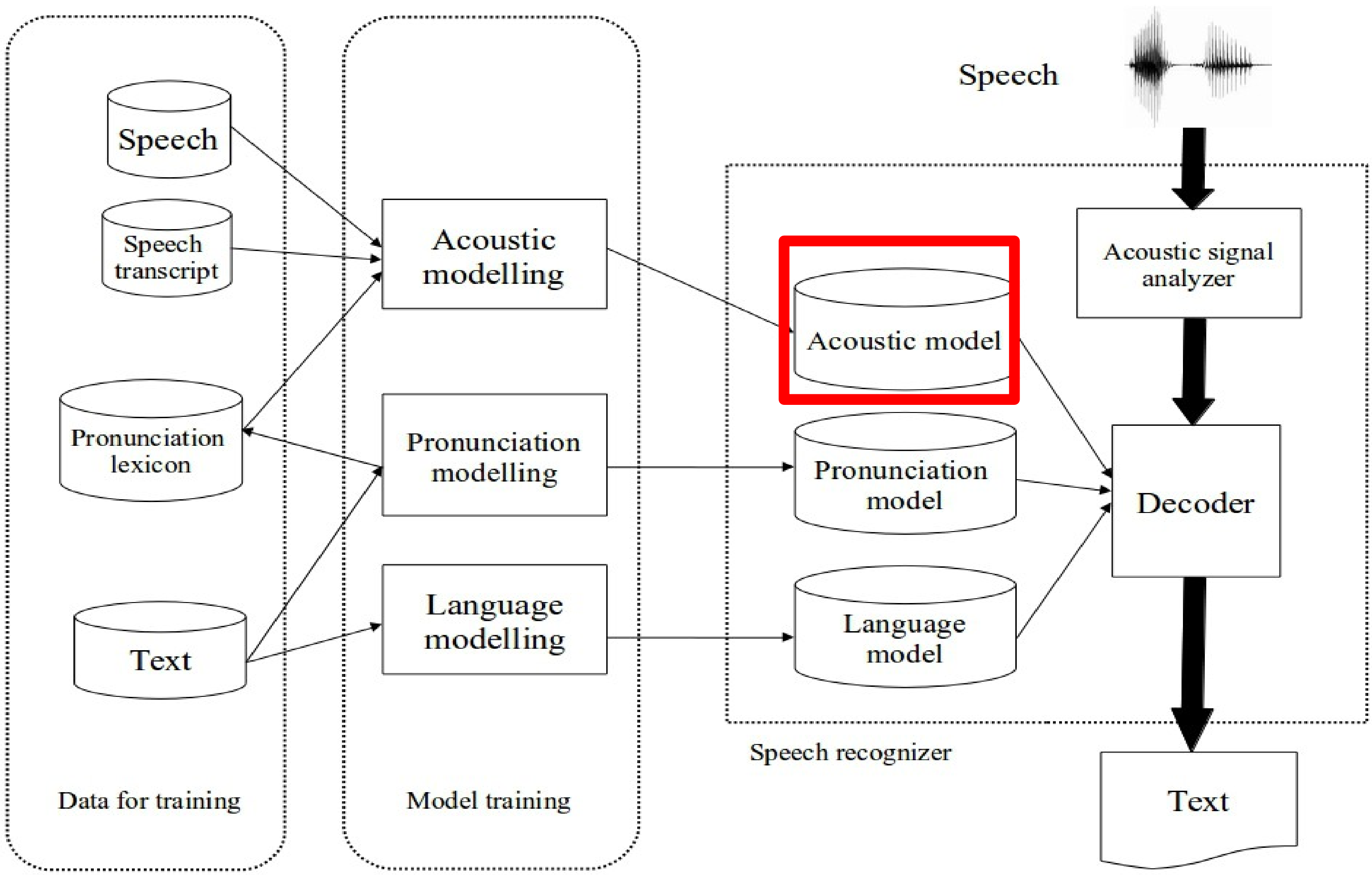
Acoustic parameters

- LPC (Linear Predictive Coding)
  - A sample is predicted as a weighted sum of preceding samples

$$\hat{s}_n = \sum_{i=1}^{p} a_i s_{n-i}$$

  - *p* is the model order
  - a_i = linear prediction coefficients
  - different methods to predict this coeff. (levinson-durbin algo.)

ASR Systems Overview



Statistical modelling

$$\hat{P}(Y|X)$$

Sequence of acoustic observations

- *Signal frames*
- *Filterbank coefficients*
- *Cepstral coefficients*
- *Time-frequency principal components*
- *…*

Sound object (or class) hypothesis

- *Sound type (speech / music / …*
- *speaker / language / channel*
- ***phone / syllable / word***
- *Sound event (jingle)*
- *Past or future of a break (ex: speaker change)*
- *…*

→ Generic Approach

Bayes

- x : observation (signal)
- ci : class to be recognized

$$c_i = \underset{i}{\text{argmax}}\ p(c_i / x) = \underset{i}{\text{argmax}} \frac{p(x/c_i).P(c_i)}{p(x)} \approx \underset{i}{\text{argmax}}\ p(x/c_i).P(c_i)$$

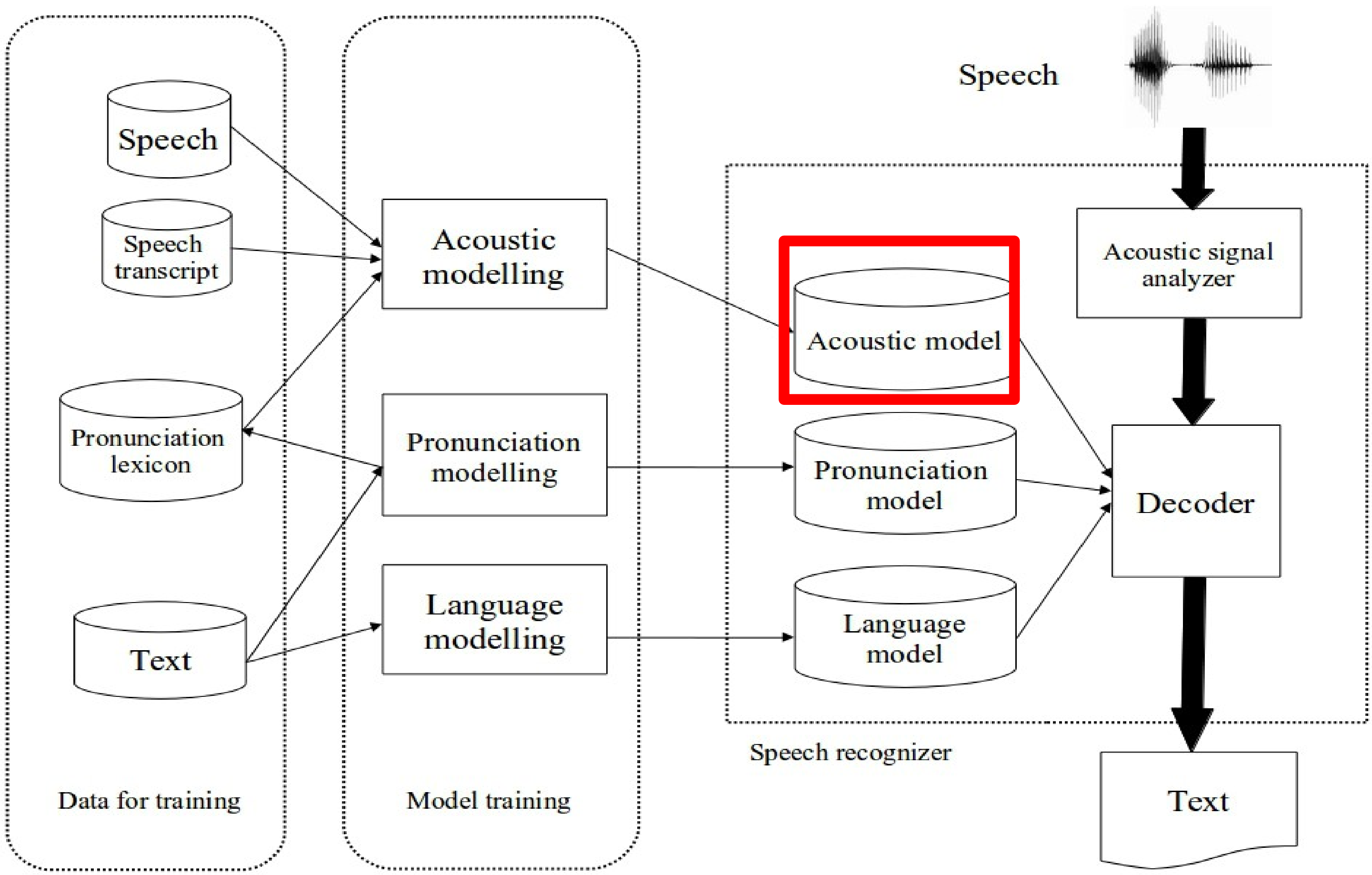
- Automatic Speech Recognition (ASR)

Acoustic model

$$w_i = \underset{i}{\text{argmax}} \frac{p(x/w_i).P(w_i)}{p(x)} = \underset{i}{\text{argmax}}\ p(x/w_i).P(w_i)$$

Language model

ASR Systems Overview



Phone (Acoustic) Models

- Generally, the acoustic units modeled are phonemes rather than words
  - Exemple : ~40 phone models for french

- To calculate p(x/w_i) an acoustic model, as well as a pronunciation dictionary are needed

Context Dependent vs. Context Independent Models

- Independent : each unit is modeled independently of the others
- Dependent : different models for a same phone unit according to the left-right context
- triphones : only nearest left and right phonemes are considered

=>due to coarticulation

=>problem : corpora should be big enough to estimate robust models

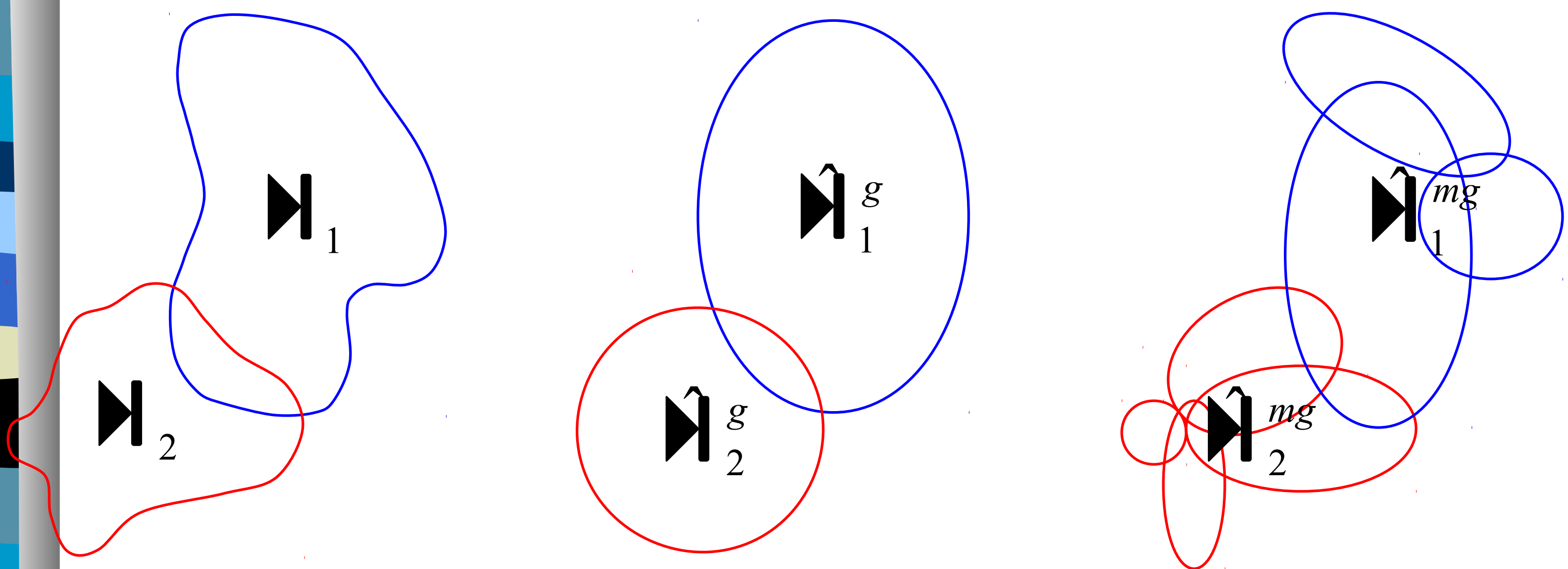
What are those models ?

Many possibilities but we'll talk only of
… stochastic models (HMM/GMMs)
and deep neural nets (DNNs)...

What are those models ?

… Hidden Markov Models with
Gaussian Distributions

Gaussians



Real distribution

Gaussian model

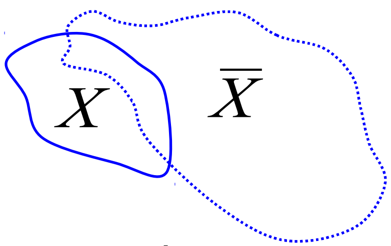
Gaussian mixture model (GMM)

Automata

- For sequence processing
- Complex sequential patterns decomposed into piecewise stationary segments
- Each segment : deterministic or stochastic function
- Can describe grammar, lexicon, phone models…
- Example : Hidden Markov Models (HMMs)
  - 2 concurrent stochastic processes :
    - Sequence of HMM states (sequential structure of the data)
    - State output processes (local characteristics of the data)
    - Example : left-right HMM phone model with gaussian mixture output distributions

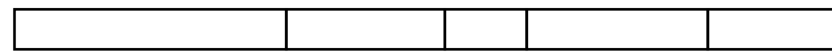
You can solve different problems with that …

Detection



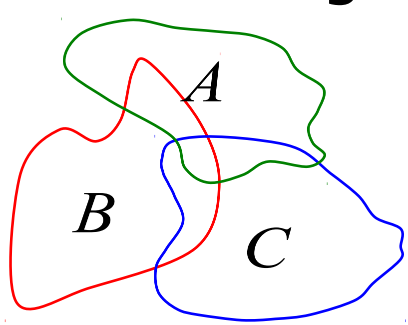
→ Binary decision tests

Segmentation



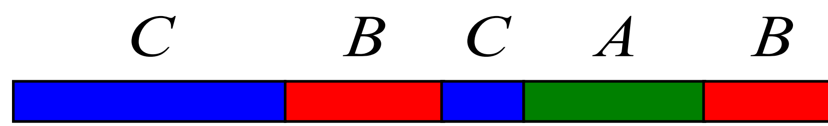
→ Change point detection

Clustering



→ Maximum A Posteriori

Decoding



→ State sequence search

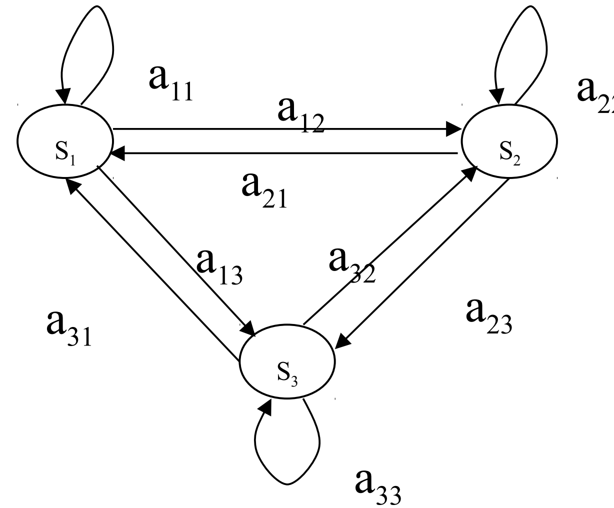
Hidden Markov Models (HMMs)

•A HMM is defined by :

- N, number of states in the model, $S=\{S_1,S_2,\ldots S_N\}$
- M, number of output (emission) symbols per state, $V=\{v_1, v_2\ldots v_M\}$
- Propability distribution are defined
  - Transition probabilities $A=\{a_{ij}\}$.
  - Emission probabilitiy of symbol k in state j : b_{jk}
  - Initial state probabilites

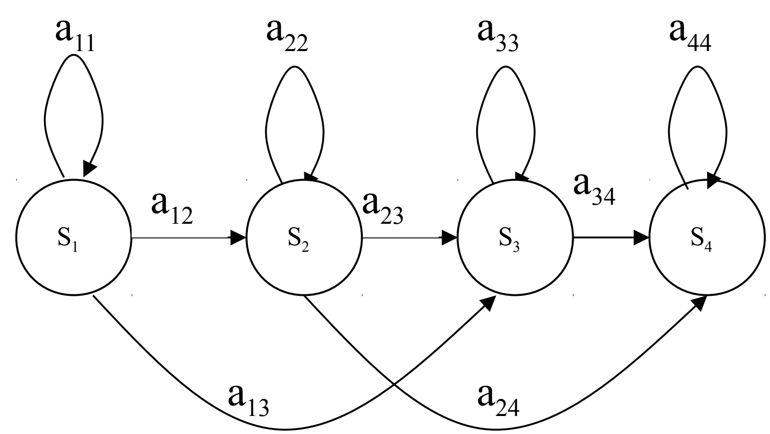
•If the set of emission symbols V is finite, the HMM is called **discrete** (if V is infinite, then the HMM is **continuous**).

HMM for speech recognition



Ergodic HMM

- Temporal aspect of speech
  - Use of left-right HMMs (Bakis model).
- Left-right HMM properties
  - aij = 0 when j < i
  - aNN = 1



Left-right HMM

Three fundamental problems of HMMs

•Given observations O and HMM λ

How to calculate $P(O|\lambda)$?

- The solution to this problem called **evaluation** is the algorithm ***Forward-Pass***

•Given observations O and HMM λ

How to choose the most probable state sequence Q that maximizes $P(Q|O, \lambda)$?

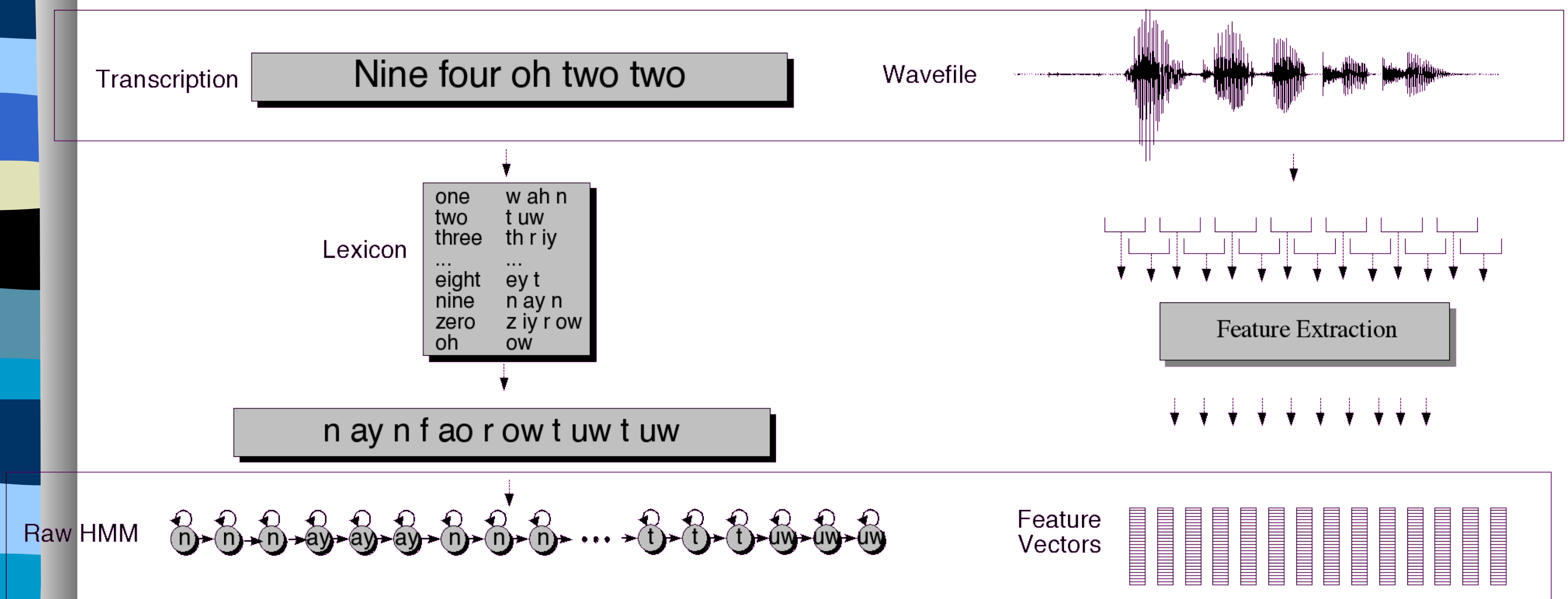
- The solution to this problem called decoding is the algorithm ***Viterbi***

•Given observations O and HMM λ

How to adjust (train) the parameters of the model to maximize $P(O|\lambda)$?
This is the **training** of the model parameters.

- Algorithm Baum-Welch, algorithm **EM** (expectation-maximization)

Forced alignment



Forced Alignment

- Computing the “Viterbi path” over the training data is called “forced alignment”
- Because we know which word string to assign to each observation sequence
- We just don’t know the state sequence
- So we constrain the path to go through the correct words
- Result: state sequence (so alignment between signal and phonemes)

What are those models ?

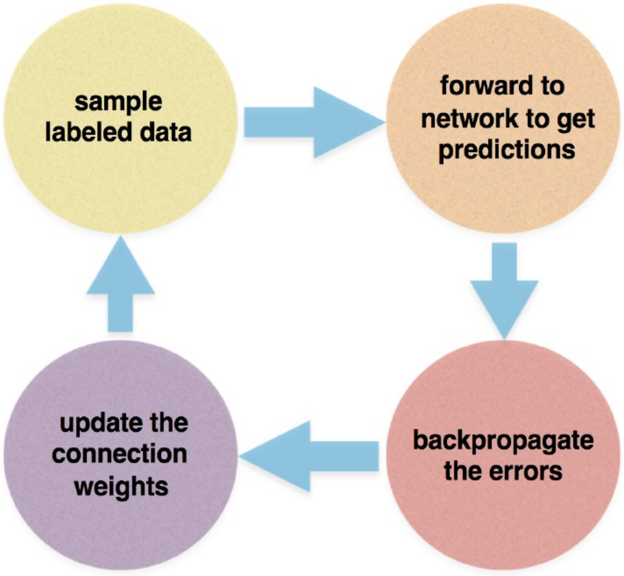
… Deep neural networks

What is deep learning ?

- Part of the ML field of learning representations of data
- Learning algorithms derive meaning out of data by using a hierarchy of multiple layers of units (neurons)
- Each unit computes a weighted sum of its inputs and the weighted sum is passed through a non linear function
- Each layer transforms input data in more and more abstract representations
- Learning = find optimal weights from data
  - ex: deep automatic speech transcription or neural machine translation systems have 10-20M of parameters

Supervised learning process

- Learning by generating error signal that measures the differences between network predictions and true values
- Error signal used to update the network parameters so that predictions get more accurate



Brief History

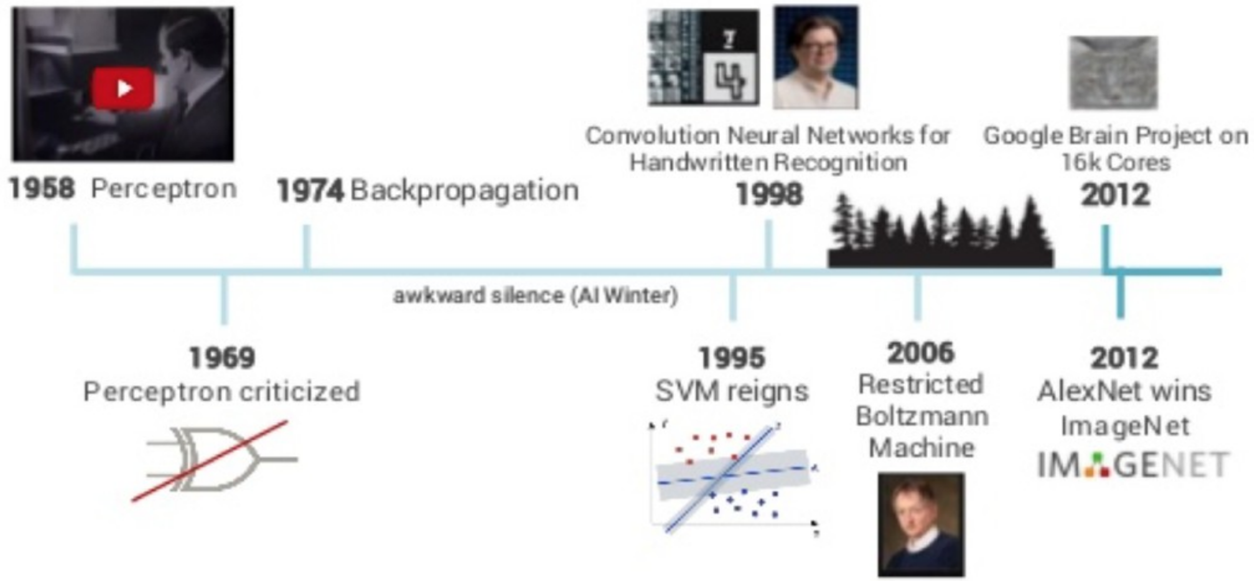


Figure from https://www.slideshare.net/LuMa921/deep-learning-a-visual-introduction

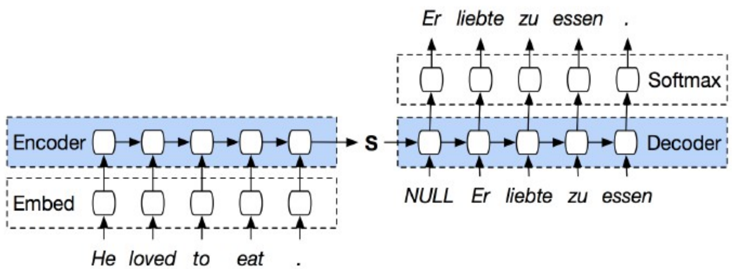
- 2012 breakthrough due to
  - Data (ex: ImageNet)
  - Computation (ex: GPU)
  - Algorithmic progresses (ex: SGD)

Success stories of deep learning in recent years

Figure from [He et al., 2017]

- Convolutional neural networks (CNNs)
  - For stationary signals such as audio, images, and video
  - Applications: object detection, image retrieval, pose estimation, etc.

Success stories of deep learning in recent years



Input sentence:	*Translation (PBMT):*	*Translation (GNMT):*	*Translation (human):*
李克強此行將啟動中加總理年度對話機制，與加拿大總理杜魯多舉行兩國總理首次年度對話。	Li Keqiang premier added this line to start the annual dialogue mechanism with the Canadian Prime Minister Trudeau two prime ministers held its first annual session.	Li Keqiang will start the annual dialogue mechanism with Prime Minister Trudeau of Canada and hold the first annual dialogue between the two premiers.	Li Keqiang will initiate the annual dialogue mechanism between premiers of China and Canada during this visit, and hold the first annual dialogue with Premier Trudeau of Canada.

Images from: https://smerity.com/media/images/articles/2016/ and http://www.zdnet.com/article/google-announces-neural-machine-translation-to-improve-google-translate/

- Recurrent neural networks (RNNs)
  - For variable length sequence data, e.g. in natural language
  - Applications: sequence to sequence prediction (machine translation, speech recognition) . . .

It's all about the features . . .

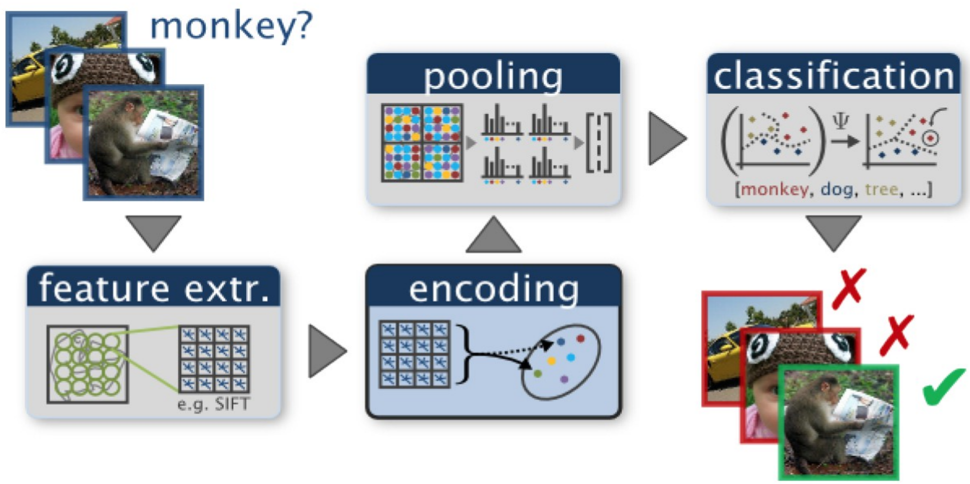
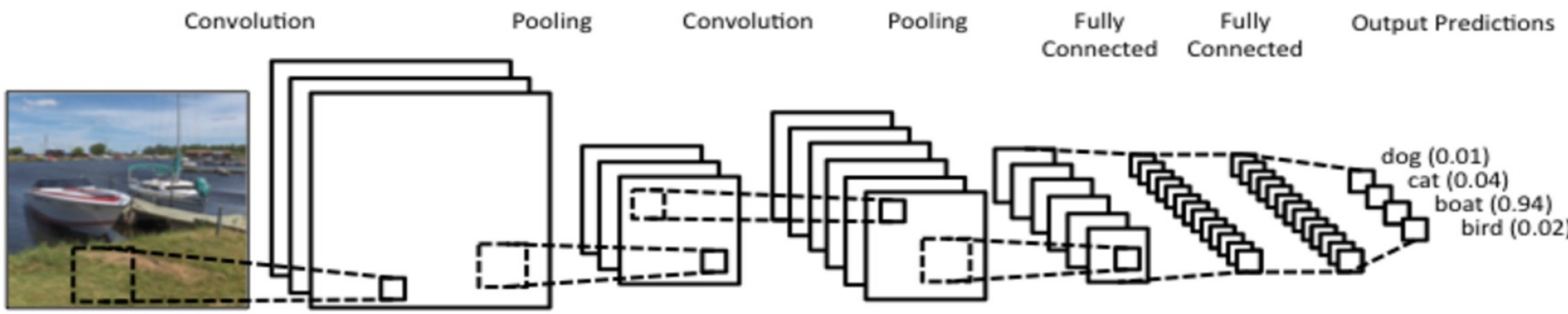


Image from [Chatfield et al., 2011]

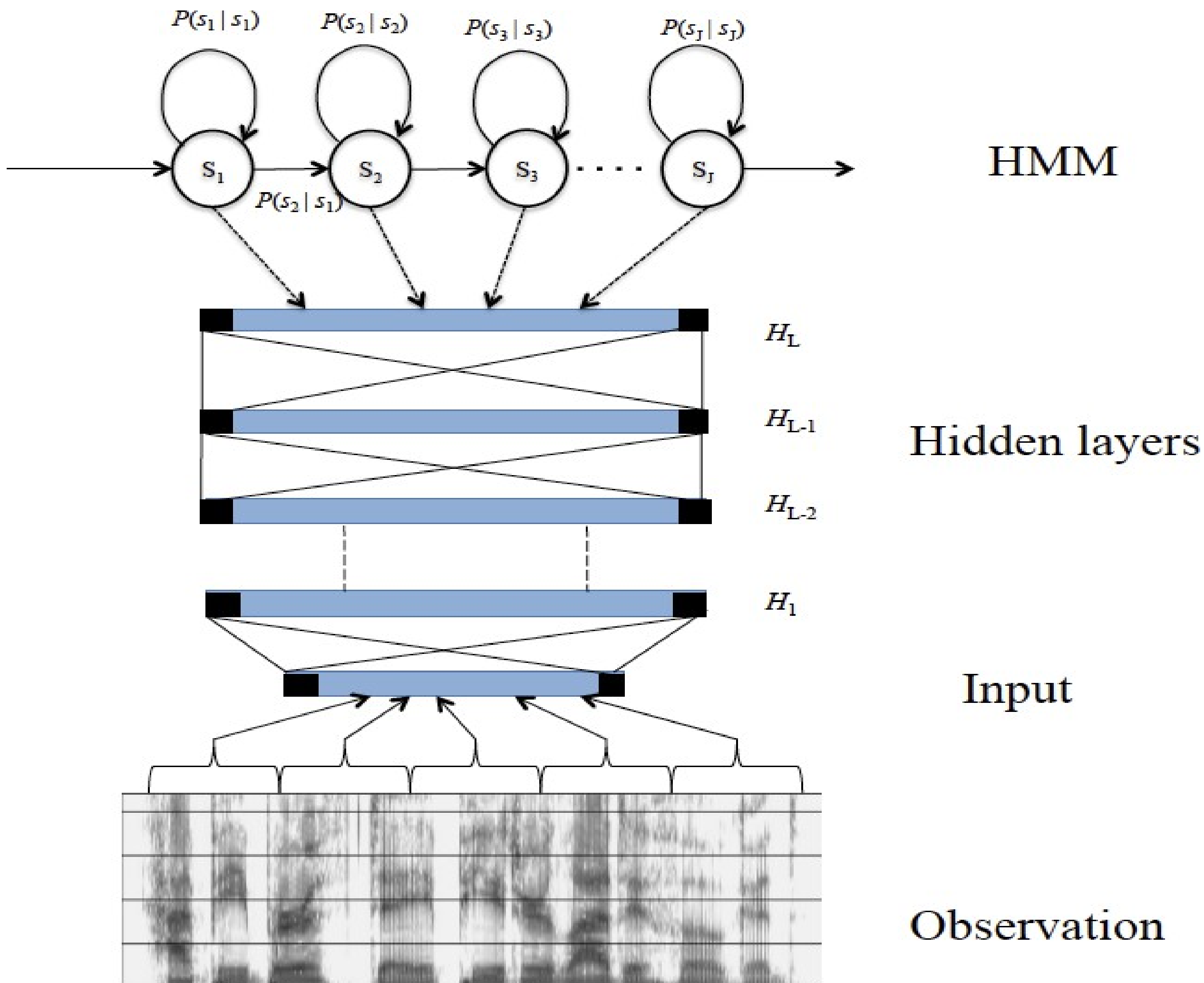
- With the right features anything is easy . . .
- Former vision / audio processing approach
  - Feature extraction (engineered) : SIFT, MFCC, . . .
  - Feature aggregation (unsupervised): bag-of-words, Fisher vec.,
  - Recognition model (supervised): linear/kernel classifier, . . .

It's all about the features . . .



- Deep learning blurs boundary feature / classifer
  - Stack of simple non-linear transformations
  - Each one transforms signal to more abstract representation
  - Starting from raw input signal upwards, e.g. image pixels
- Unified training of all layers to minimize a task-specific loss
- Supervised learning from lots of labeled data

Hybrid HMM/DNNs (2012)



NN trained end-2-end (2016)

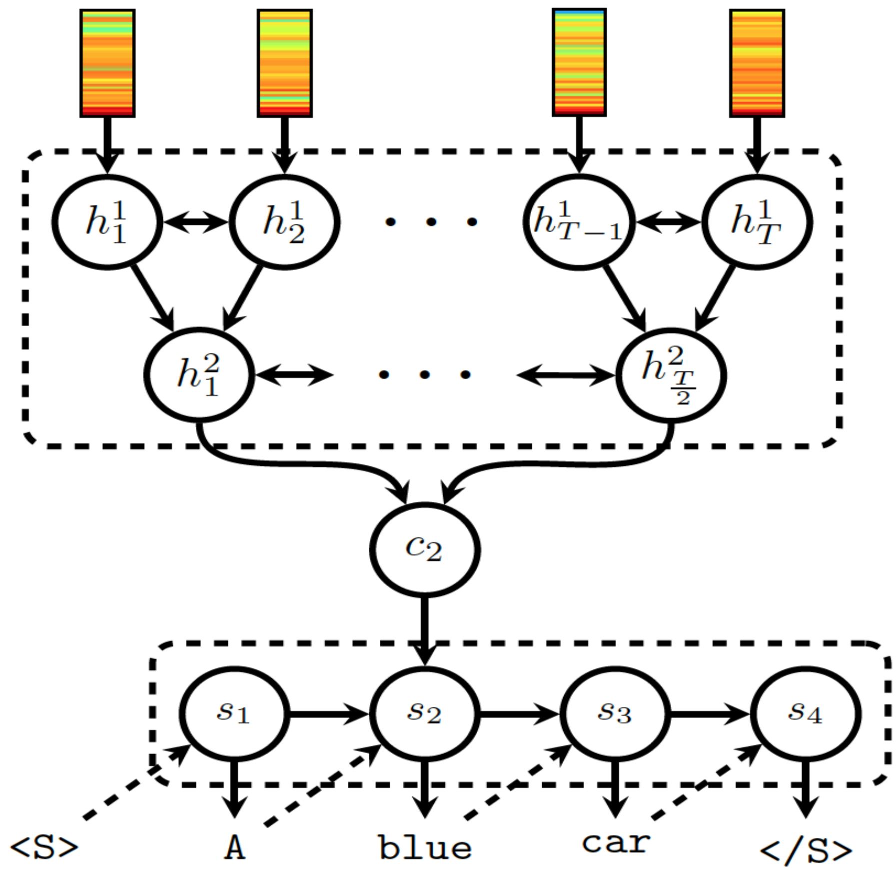


Image from Alexandre Berard's thesis

DNN-HMM vs. GMM-HMM
(2012)

- **Table:** TIMIT Phone recognition (3 hours of training)

Features	Setup	Error Rates
GMM	Incl. Trajectory Model	24.8%
DNN	5 layers x 2048	23.0%

~10% relative improvement

- **Table:** Voice Search SER (24-48 hours of training)

Features	Setup	Error Rates
GMM	MPE (760 24-mix)	36.2%
DNN	5 layers x 2048	30.1%

~20% relative improvement

- **Table:** Switch Board WER (309 hours training)

Features	Setup	Error Rates
GMM	BMMI (9K 40-mix)	23.6%
DNN	7 layers x 2048	15.8%

~30% relative improvement

- **Table:** Switch Board WER (2000 hours training)

Features	Setup	Error Rates
GMM	BMMI (18K 72-mix)	21.7%
DNN	7 layers x 2048	14.6%

DL take home messages

- Core idea of deep learning
  - Many processing layers from raw input to output
  - Joint learning of all layers for single objective
- A strategy that is effective across different disciplines
  - Computer vision, speech recognition, natural language processing, game playing, etc.
- Widely adopted in large-scale applications in industry
  - Face tagging on FaceBook over 109 images per day
  - Speech recognition on iPhone
  - Machine translation at Google, Systran, DeepL, etc.
- Open source development frameworks available (pytorch, tensor flow and the like)
- Limitations: compute and data hungry
  - Parallel computation using GPUs
  - Re-purposing networks trained on large labeled data sets

Some directions of ongoing research (1/2)

- Optimal architectures and hyper-parameters
  - Possibly under constraints on compute and memory
  - Hyper-parameters of optimization: learning to learn (meta learning)
- Irregular structures in input and/or output
  - (molecular) graphs, 3D meshes, (social) networks, circuits, trees, etc.
- Reduce reliance on supervised data
  - Un-, semi-, self-, weakly- supervised, etc.
  - Data augmentation and synthesis (e.g. rendered images)
  - Pre-training, multi-task learning
- Uncertainty and structure in output space
  - For text generation tasks (ASR, MT, NLG): many different plausible outputs (see our ACL paper)

Some directions of ongoing research (2/2)

- Analyzing learned representations
  - Better understanding of black boxes
  - Explainable AI
  - Neural networks to approximate/verify long standing models and theories (link with cognitive sciences)
- Robustness to adversarial examples that fool systems
- Introducing prior knowledge in the model
- Biases issues (GenderShades and the like)
- Common sense reasoning
- etc.

Outline

- Signal processing reminder
- The speech signal
- **Automatic Speech Recognition (ASR)**
  - Overview
  - Speech modelling (parameters, models)
  - Toolkits for ASR design

ASR Toolkits (1)

- HTK (Cambridge)
  - *htk.eng.cam.ac.uk*
- SPHINX (CMU)
  - http://cmusphinx.sourceforge.net
- JULIUS (Japon)
  - http://julius.sourceforge.jp/
- RWTH (Aachen, Allemagne)
  - http://www-i6.informatik.rwth-aachen.de/rwth-asr/
- KALDI (JHU, USA)
  - http://kaldi.sourceforge.net/

ASR Toolkits (2)

- *HTK* et *SPHINX* broadly used and documented
  - HTK Bible (book)
    - http://htk.eng.cam.ac.uk/docs/docs.shtml
  - Sphinx workshops
    - http://www.cs.cmu.edu/~sphinx/Sphinx2010/index.html
- *Julius* allows to use grammars instead of n-grams
- See also http://persephone.readthedocs.io

ASR Toolkits (3)

- Tools for extracting parameters, acoustic modelling and decoding
- Pre-trained acoustic models for some languages
  - Toy examples
  - http://www.speech.cs.cmu.edu/sphinx/models/

See also http://kaldi.sourceforge.net /

- **Practical example with KALDI**
  - **https://github.com/besacier/ALFFA_PUBLIC/tree/master/ASR**

FIN