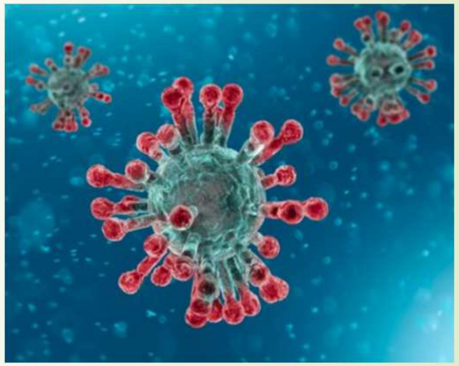
Speech Recognition and

Deep Neural Networks



Multispeech

Loria Inria, Nancy, France

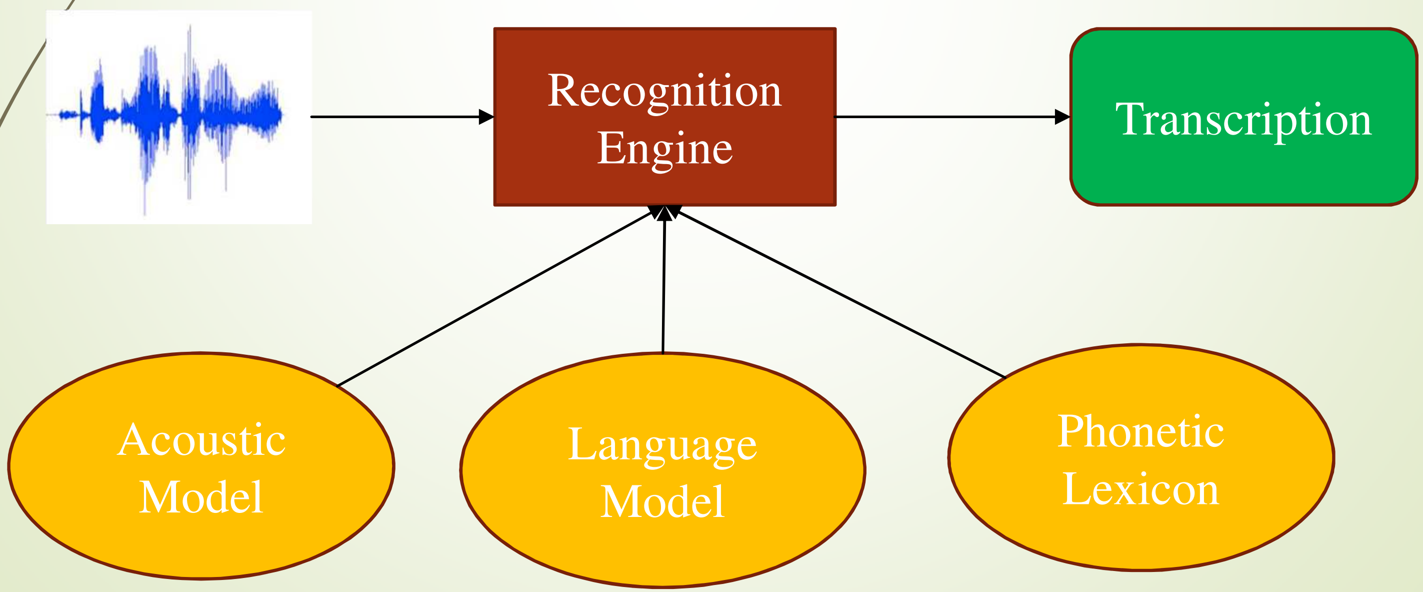
1

Information carried by speech

- Linguistic content (words)
  - Speech recognition
    - Recognition of all uttered words, or just some keywords
    - ➔ Vocal commands, speech transcription, vocal indexing, etc.
- Speaker (who speaks)
  - Speaker recognition
    - Speaker identification, or speaker authentication
    - ➔ Diarization (associating speech segments with speakers), etc.
- Language
  - Language recognition
    - Identification of the spoken language, or of the dialect, accent, etc.
- Paralinguistic information
  - Emotions
    - Neutral speech, joy, sadness, anger, etc.
  - Speaking style
    - Spontaneous vs. read speech, sport commentary, etc.

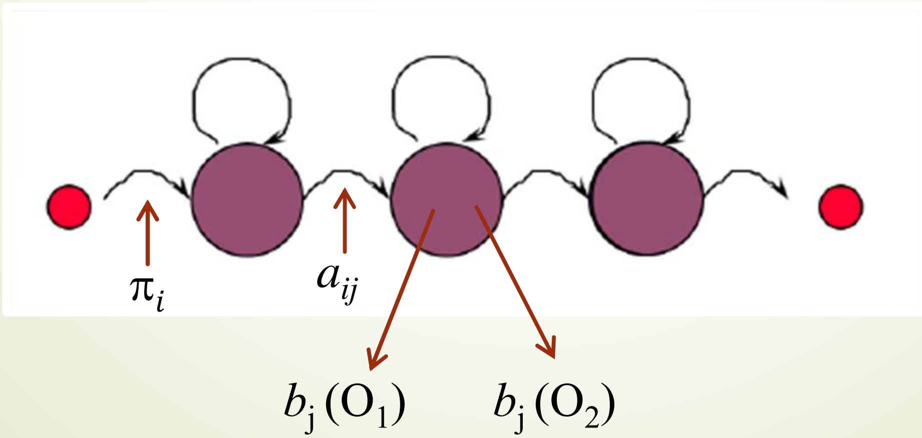
Automatic speech recognition system

- Input: audio file
- Output: transcription (text)



Acoustic models

- *Hidden Markov Models* (HMM)
  - Finite state automaton with N states, composed of three components: $\{A, B, \Pi\}$
    - ❖ $A[a_{ij}]$: matrix of transitions (NxN)
    - ❖ $\Pi[\pi_i]$: initial probabilities (N)
    - ❖ $B[b_j]$: observation probabilities

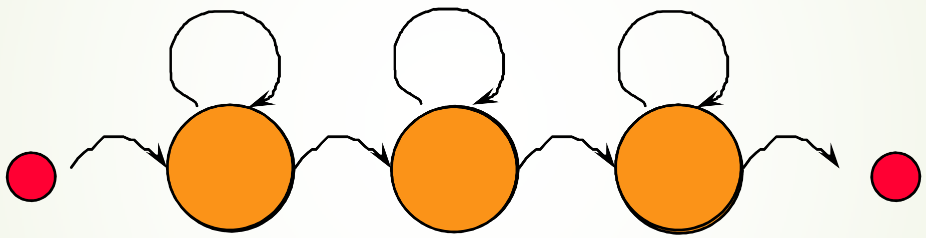


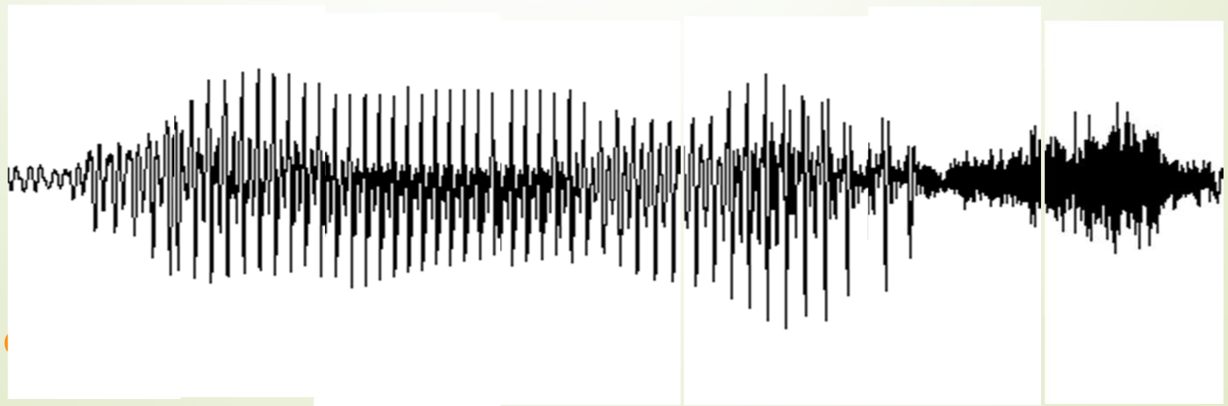
Why hidden?

- Because we see only observations
- We don’t know the state sequence that produced the observation sequence

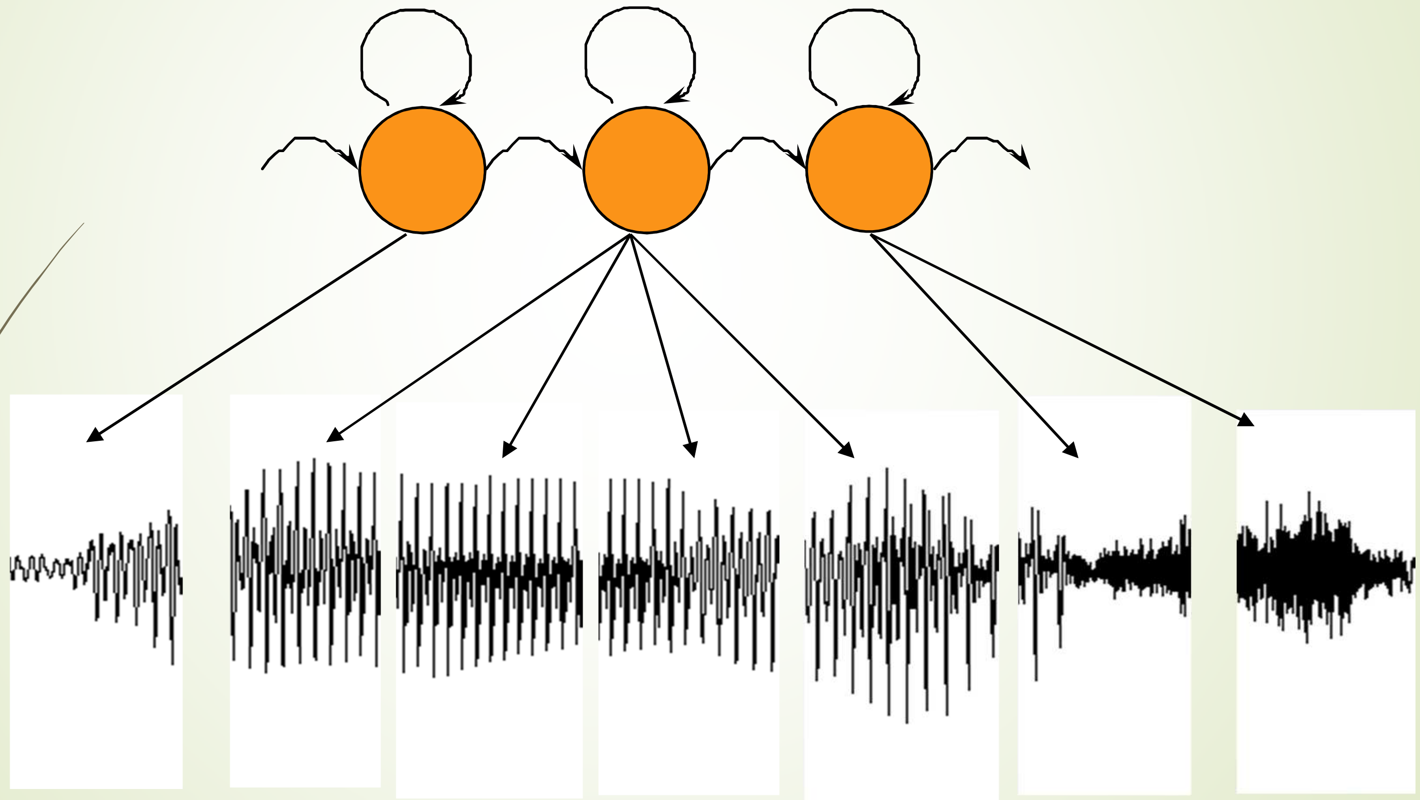
Speech modeled by HMM

- We assume that the speech is produced by a ***Markov system***





HMM



Observation probability

- Two possibilities:
  - GMM (Gaussian Mixture Model): Observation probability is modeled by a mixture of *M* Gaussians

$$b_j(x) = \sum_{m=1}^{M} c_{jm}\, \mathcal{N}(x; \mu_{jm}, \Sigma_{jm})$$

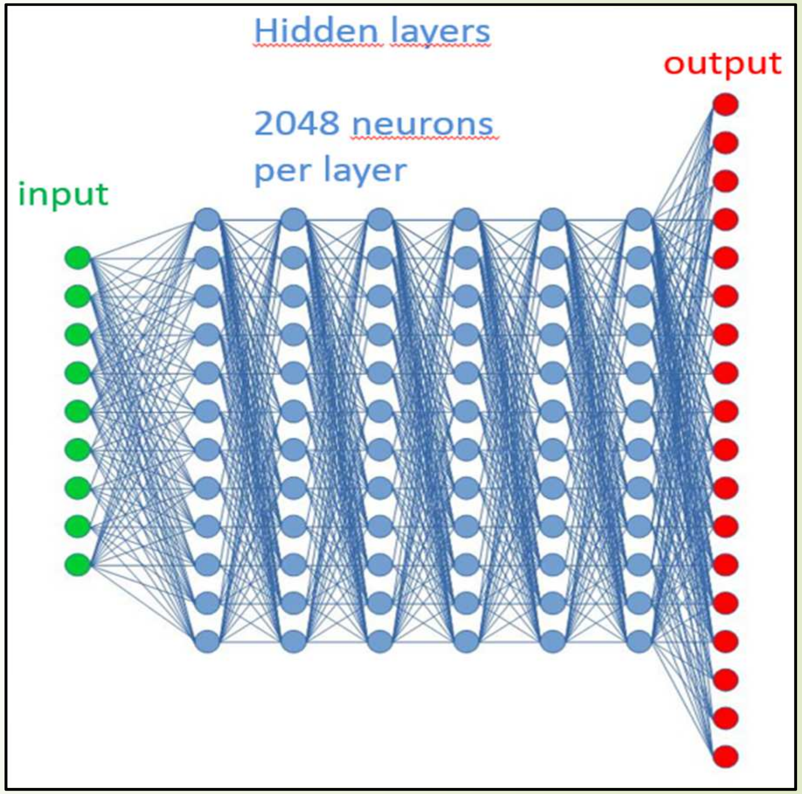
  - DNN (Deep Neural Network): Observation probability is modeled by a Deep Neural Network

Deep Neural Network (DNN)

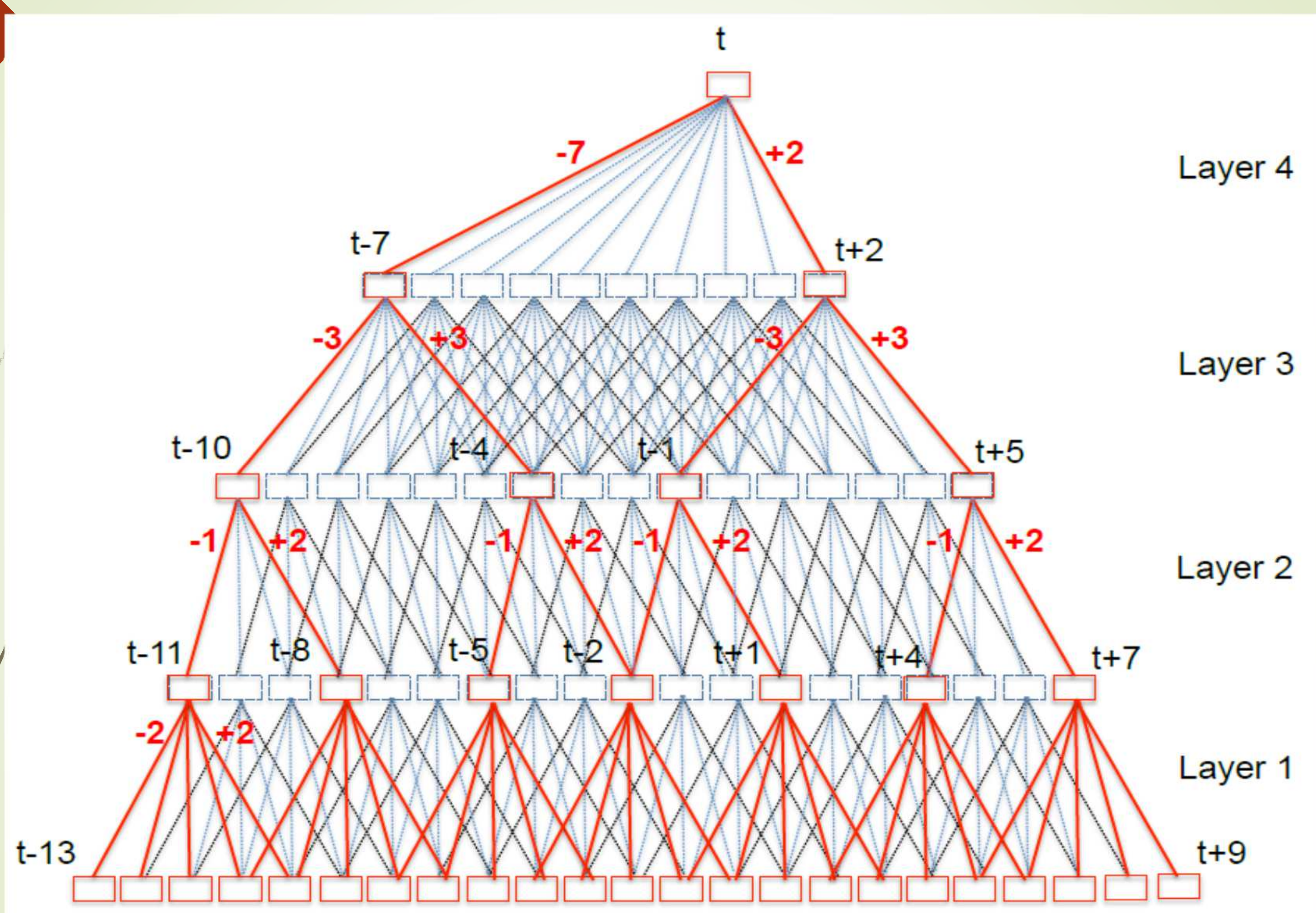
- A DNN is defined by three types of parameters:
  - The interconnection pattern between the different layers of neurons
  - The training process for updating the weights w_i of the interconnections
  - The activation function f that converts a neuron's weighted input to its output activation

Architecture example of DNN for acoustic model

- MLP (Multi Layer Perceptron)
- 6 hidden layers
- 2048 neurons for each hidden layer
- Input: size of the acoustic parameters (39)
- Output: number of HMM states (4048 context-dependent phone states)



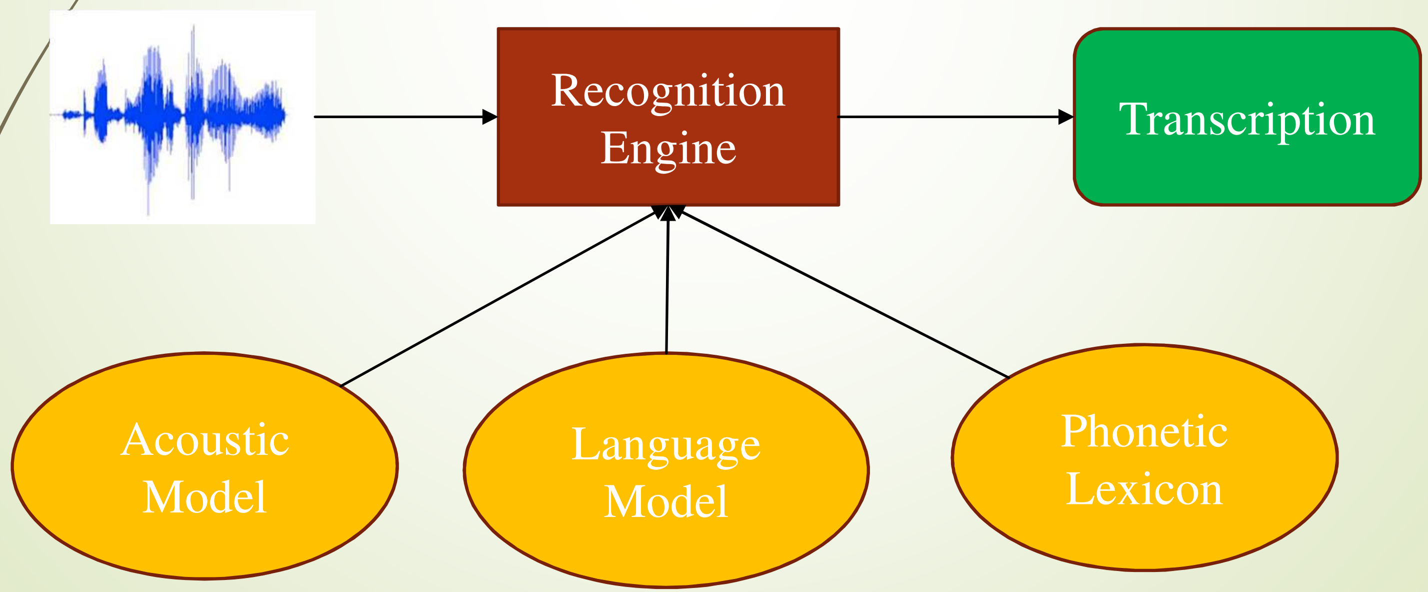
TDNN – Time Delay Neural Network [Peddinti et al. 2015]



V. Peddinti, G. Chen, V. Manohar, T.Ko, D. Povey, S. Khudanpur, JHU ASpIRE system: Robust LVCSR with TDNNs i-vector Adaptation and RNN-LM. *Proc. of the IEEE ASRU*, 2015.

Automatic speech recognition system

- Input: audio file
- Output: transcription (text)



Language model

- Compute the probability of a word knowing the previous words

- Two possibilities:
  - N-gram
  - *Recurrent Neural Networks (RNN)*

N-gram

- An *n*-gram model gives the probability of a word w_i given the *n-1* previous words:

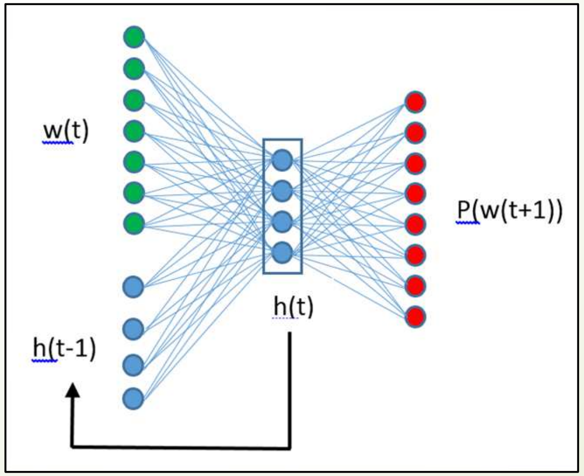
$$P(w_i|w_1,\ldots,w_{i-1}) = P(w_i|w_{i-(n-1)}, w_{i-(n-2)}, \ldots, w_{i-1})$$

- Advantages
  - Easy to compute
  - Rare events are taken into account
- Drawbacks
  - Only 3-grams or 4-grams can be evaluated (short term dependency)
  - No generalization
    - In the training corpus “*a blue car*” “*a red Ferrari*”

      The probability of “*a blue Ferrari*“ (never seen) will be badly estimated

Recurrent Neural Network Language Model (RNNLM)

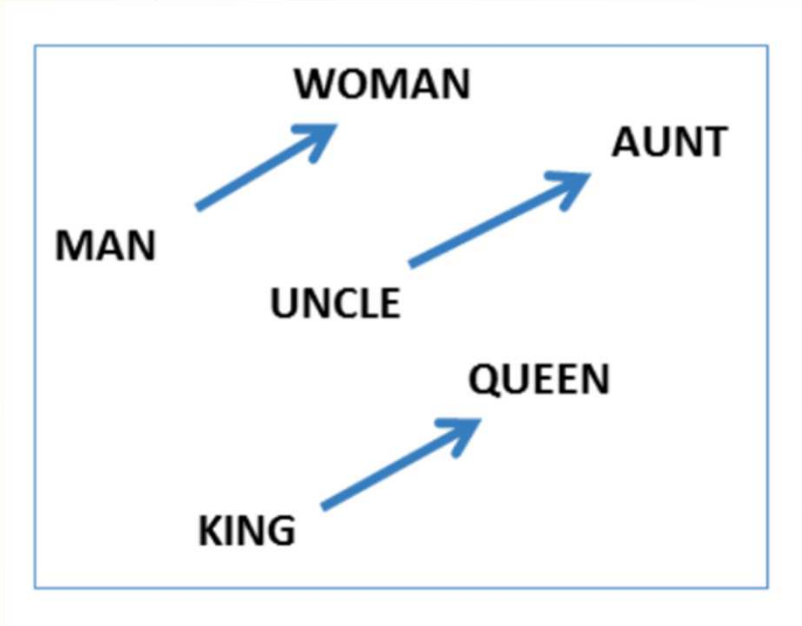
- Continuous space representation (*word embedding*)
  - Using NN for projecting words in a continuous space
- To take into account the temporal structure of language (word sequences)
  - Recurrent Neural Networks [Chen et al., 2015]

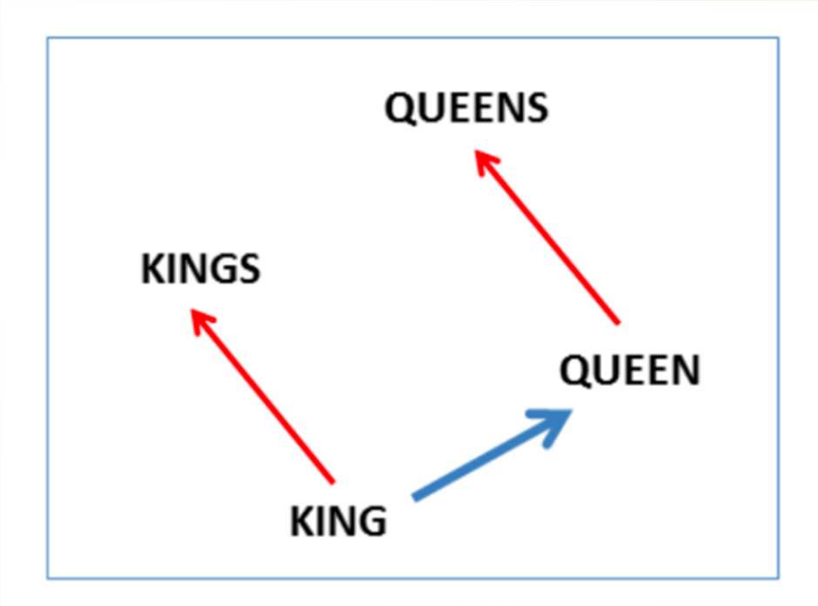


- [Chen et al, 2015] Xie Chen, Xunying Liu, Mark JF Gales, and Philip C Woodland, "Improving the training and evaluation efficiency of recurrent neural network language models," in Proc. ICASSP, 2015.

Some properties of *word embeddings*

- Vector relations between words





Long term dependency

- To take into account long term dependencies in a sentence
  - Ex: *les étudiantes inscrites à la conférence ACL sont arrivées*

Short term dependency: bigram

Long term dependency: 8-gram

➔ Add a memory mechanism
  - Long Short Term Memory (LSTM)

[Kumar et al. 2017] S. Kumar, Michael A. Nirschl, D. HoltmannRice, H. Liao, A. Theertha Suresh, and F. Yu, Lattice rescoring strategies for long short term memory language models in speech recognition, in ASRU Workshop, 2017.

[Li et al. 2020] K Li, Z Liu, T He, H Huang, F Peng, D Povey, S Khudanpur An Empirical Study of Transformer-Based Neural Language Model Adaptation, ICASSP 2020.

Language model for speech recognition

- Combination of LMs
  - Advantage of N-gram: rare events are taken into account
  - Advantage of RNNLM: generalization capacity

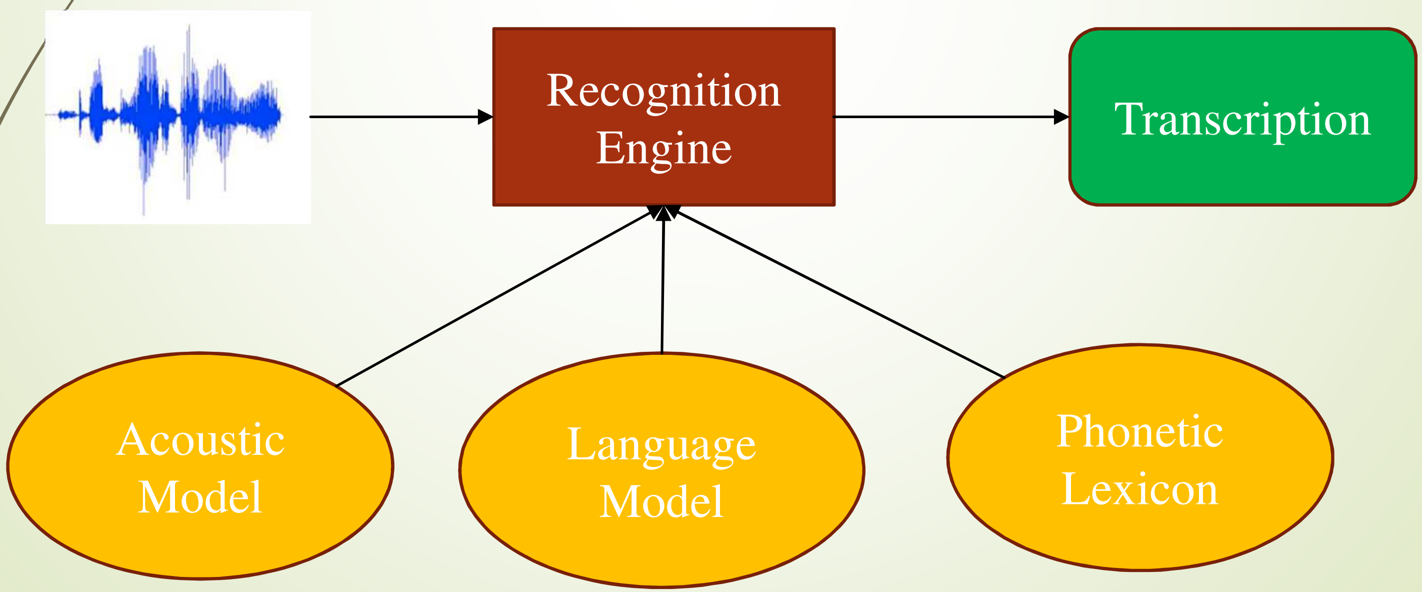
$$P(w|h) = \lambda\, P_{ngram}(w|h) + (1-\lambda)\, P_{RNNLM}(w|h)$$

Improvement of WER is about 20% relative [Sundermeyer &Ney 2015]

[Sundermeyer &Ney 2015] M. Sundermeyer, H. Ney, R. Schlüter (2015). From Feedforward to Recurrent LSTM Neural Networks for Language Modeling, IEEE Transactions on Audio, Speech and Language Processing, vol.. 23, no. 3, March 2015.

Automatic speech recognition system

- Input: audio file
- Output: transcription (text)



Phonetic lexicon

Examples in English (from cmudict)

soften	S AA F AH N	sorbet	S AO R B EY
soften(2)	S AO F AH N	sorbet(2)	S AO R B EH T

- The lexicon specifies the list of words known by the ASR system [Sheikh 2016]
  - An ASR system cannot recognized words that are not in the lexicon
  - It is impossible to have a lexicon covering all possible words (because of person names, company names, product names, etc.)
  - Diachronic evolution of vocabularies (due to new topics, new persons, etc.)
- The lexicon also specifies the possible pronunciations of the words
  - Must include the usual pronunciation variants
  - But one should not include too many useless variants as this increases possible confusions between vocabulary words

[Sheikh 2016] I. Sheikh. Exploiting Semantic and Topic Context to Improve Recognition of Proper Names in Diachronic Audio Documents. . Université de Lorraine, 2016.

Speech recognition errors

- **Insertion** / **Deletion / Substitution**
  - `Ref.   :      I want to go to Paris`
  - `Reco.  : well I want to go    Lannion`
- WER : Word Error Rate

$$WER = \frac{N_{sub} + N_{ins} + N_{del}}{N_{refwords}}$$

Experimental evaluation

- Kaldi-based Transcription System (KATS) [Povey et al., 2011]
  - Segmentation and diarization
    - Splits and classifies the audio signal into homogeneous segments
      - Non-speech segments (music and silence)
      - Telephone speech
      - Studio speech
  - Parametrization [MFCC]
    - 13 MFCC + 13 Δ and 13 $\Delta\,\Delta$
      - ➔ 39-dimension observation vector

[Povey et al., 2011] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlıcek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely (2011). The Kaldi Speech Recognition Toolkit, ASRU

[MFCC] https://en.wikipedia.org/wiki/Mel-frequency_cepstrum

Corpus

- The training and test data from the radio broadcast news corpus (ESTER project [Gravier et al., 2004])

- **Training:** 250 hours of manually transcribed shows for
  - France Inter
  - Radio France International
  - TVME Morocco

- **Evaluation:**
  - 4 hours of speech

[Gravier et al., 2004] Gravier, G. & Bonastre, Jean-François & Geoffrois, E. & Galliano, Sebastian & Tait, K. & Choukri, Khalid. (2004). The ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News.

Results (Word Error Rate)

Shows	# words	GMM-HMM	DNN-HMM
20070707_rfi (France)	5473	23.6	16.5
20070710_rfi (France)	3020	22.7	17.4
20070710_france_inter	3891	16.7	12.1
20070711_france_inter	3745	19.3	14.4
20070712_france_inter	3749	23.6	16.6
20070715_tvme (Morocco)	2663	32.5	26.5
20070716_france_inter	3757	20.7	17.0
20070716_tvme (Morocco)	2453	22.8	17.0
20070717_tvme (Morocco)	2646	25.1	20.1
20070718_tvme (Morocco)	2466	20.2	15.8
20070723_france_inter	8045	22.4	17.4
Average	41908	22.4	**17.1**

- ❖ DNN-based system outperforms the GMM-based system
- ❖ WER difference is **5.3**% absolute, and **24**% relative
- ❖ Improvement is **statistically significant** The confidence interval +/- 0.4 %

➔ DNN-based acoustic models achieves better classification and has better generalization ability

Human vs machine

Word Error Rates	Switchboard	Call Home
Professional transcribers	5,9%	11,3%
Automatic speech recognition (combination of many NN-based systems, trained on large data sets)	5,8%	11,0%

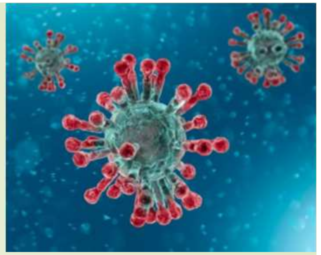
(2017 – Microsoft)

- The results obtained with a combination of many speech recognition systems **get similar** to those of professional transcribers

Conclusion

- From 2012, excellent results of DNN in many domains:
  - image recognition, speech recognition, language modelling, parsing, information retrieval, speech synthesis, translation, autonomous cars, gaming, etc.
- The DNN technology is now mature to be integrated into products.
- Nowadays, main commercial recognition systems (Microsoft Cortana, Apple Siri, Google Now and Amazon Alexa) are based on DNNs.

Conclusion



But performance still degrades in adverse conditions, such as

- High level noise
- Hands free distant microphones (reverberation problems)
- Accents (non-native speech)

Limited vocabulary (even if very large, there is still the problem of person names, location names, etc.)

Still far from an ***universal recognition system****, as powerful as a human listener in* ***all conditions***

But performance continue to improve…

Deep Neural Networks and speech recognition

Advantages

- Stunning performance
  - Revolution of the state of the art results
- Lot of applications
- No hypothesis on the input data
- Scalability with corpus size
- Generalization
  - Good performance for unseen data
- End-to-end systems [Hadian et al, 2018]
  - No need to define features
- Lot of toolkits easy to use
  - With many examples

Drawbacks

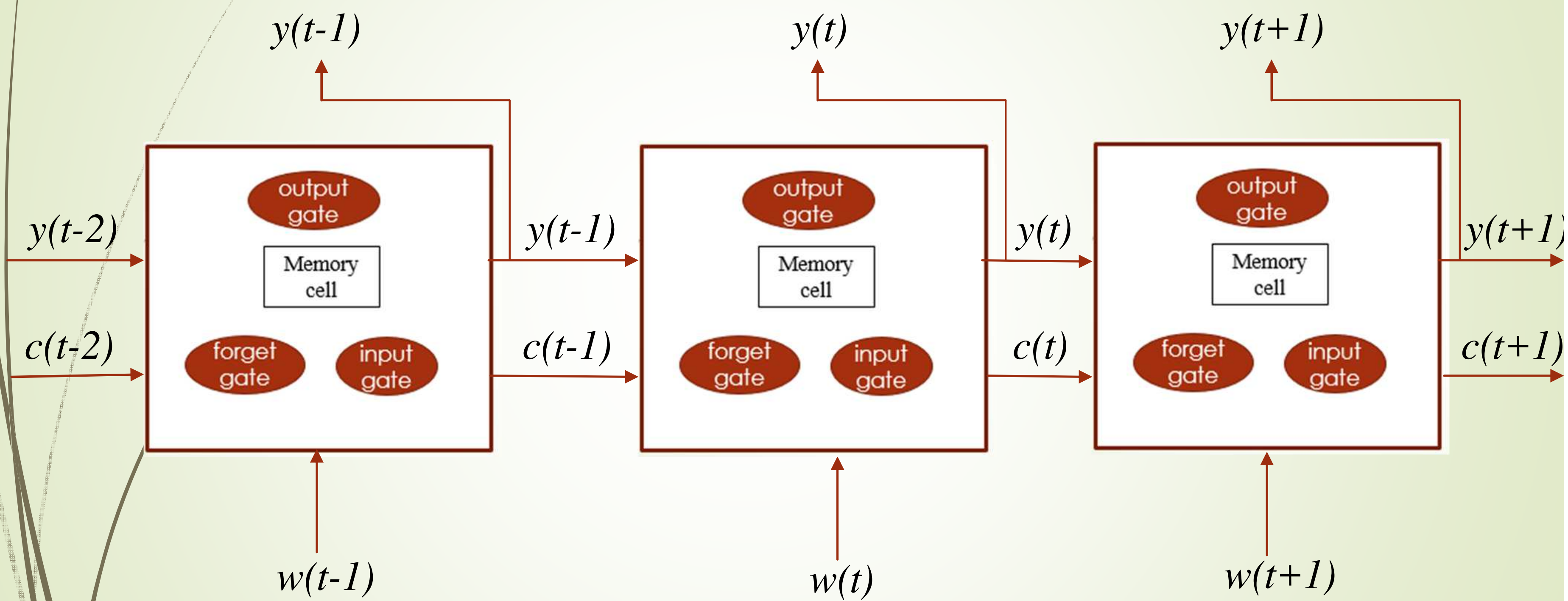
- Black box
- Hyper parameters tuning
- Huge training data needed
- Supervised training
  - Labelled training data needed
- Computationally intensive
  - Training requires GPUs or a cluster

H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "End-to-end speech recognition using lattice-free mmi," in Proc. Interspeech 2018, 2018, pp. 12–16

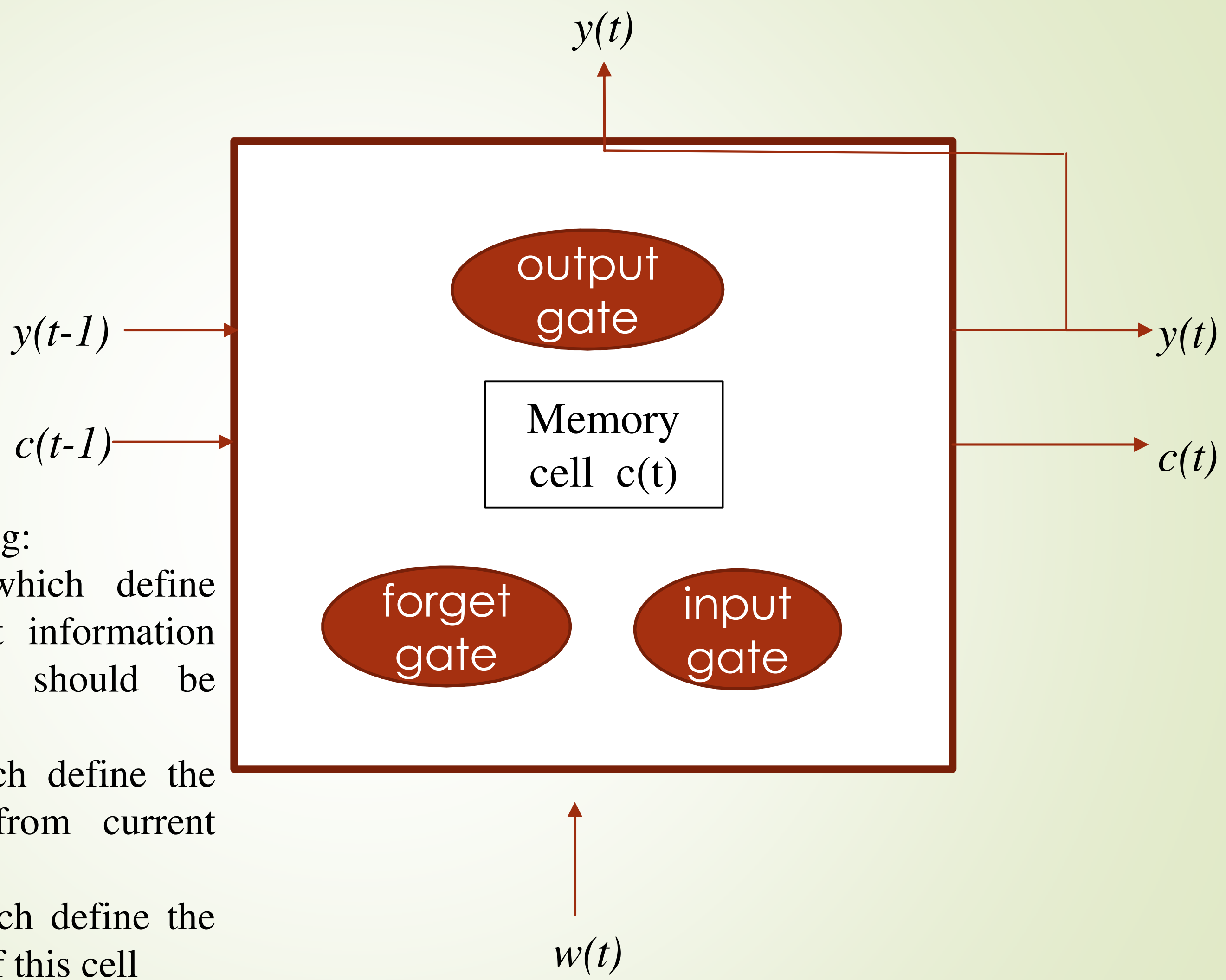
Continuous speech recognition

- Efficient algorithms and tools for building and optimizing models from data
  - Language ⇔ text corpora
  - Acoustic ⇔ speech corpora (with associated transcription)
- But many choices have to be done by the ASR system developer
  - Type of acoustic features, size of temporal windows, etc.
  - Acoustic model structure
    - Number of states / of densities / of Gaussian components per density
    - Or, type of neural network, number of layers, size of layers
- Some trade-off are necessary
  - Few parameters ⇔ rough modeling but reliable estimation
  - Many parameters ⇔ detailed modeling, but estimation may be unreliable
- Training from speech data leads to good recognition performance on similar speech data (but performance degrade in **different/new** conditions)

LSTM Long Short Term Memory



Long Short Term Memory (LSTM)



Complex structure including:

« ***forget gate*** » which define how much recurrent information (from past frame) should be kept

« ***input gate*** » which define the new contribution (from current time frame)

« ***output gate*** » which define the output contribution of this cell