CNRS - INP - UT3 - UT1 - UT2J

Institut de Recherche en Informatique de Toulouse

L'aventure de l'Intelligence Artificielle au service de la reconnaissance automatique de la parole : de 1897 à nos jours

ou

« Quelle est la place des connaissances dans l'élaboration des systèmes de reconnaissance automatique de parole (SRAP) »

Régine André-Obrecht - professeure émérite Université Toulouse III - Paul Sabatier

Equipe SAMoVA « *Structuration, Analyse, Modélisation de documents Vidéo et Audio* »

- **Un avant-propos essentiel : la reConnaissance**
  - ✓ L'intelligence artificielle de 1956 aux années 1980
  - ✓ Les connaissances en parole et les défis en traitement de la parole
  - ✓ La définition de la reconnaissance automatique de la parole
- **Les deux approches des années 70-80 « analytique/globale »**
  - ✓ Les systèmes analytiques type systèmes experts
  - ✓ La reconnaissance globale
- **L'intégration de la variabilité par approche probabiliste**
  - ✓ La modélisation acoustique et le modèle de langage
  - ✓ La reconnaissance automatique de parole continue « large vocabulaire »
- **L'approche Deep Learning**
  - ✓ Les systèmes hybrides (approche probabiliste + approche neuronale)
  - ✓ Les systèmes End-to-End (E2E)
- **Jusqu'où ? Quelle reConnaissance ?**

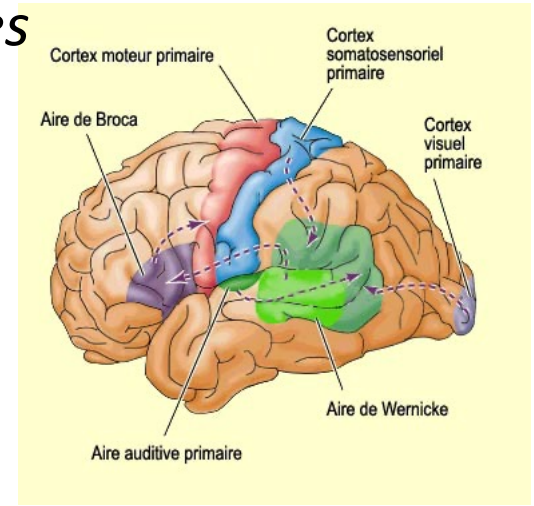
- **Un avant-propos essentiel : la reConnaissance**

  ✓L'intelligence artificielle de 1956 aux années 1980
  ✓Les connaissances en parole et les défis en traitement de la parole
  ✓La définition de la reconnaissance automatique de la parole

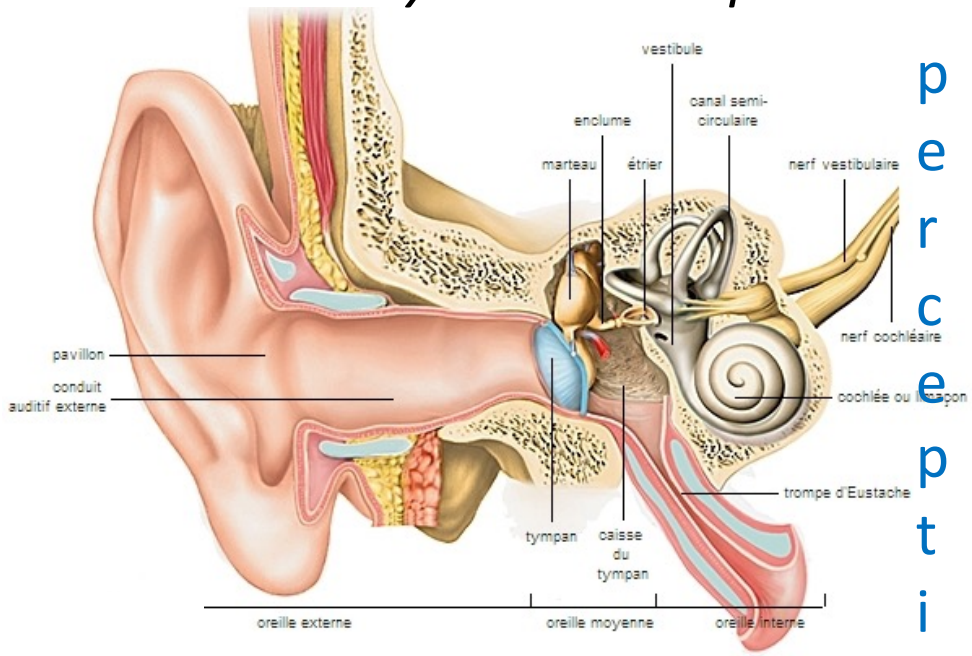
- **Les deux approches des années 70-80 « analytique/globale »**
- **L'intégration de la variabilité par approche probabiliste**
- **L'approche Deep Learning**
- **Jusqu'où ? Quelle reConnaissance ?**

Les connaissances en parole

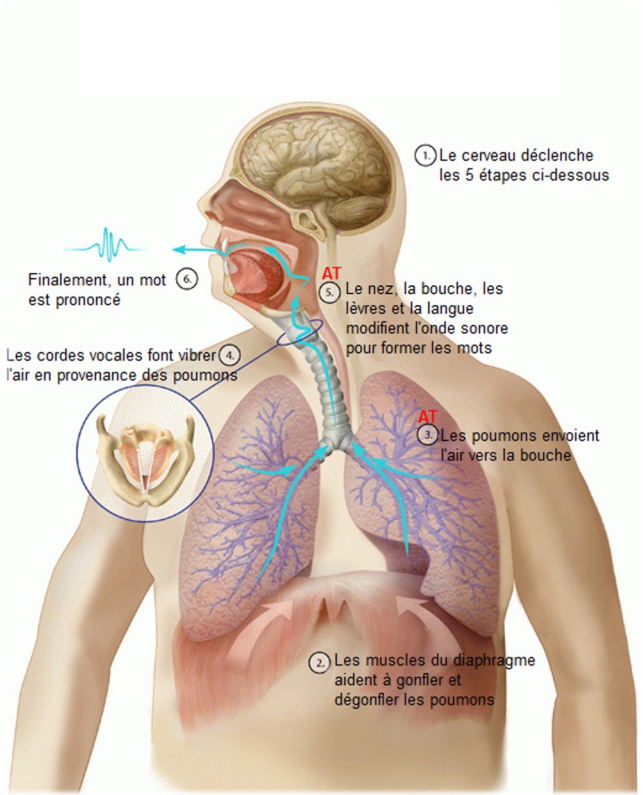
Neurosciences



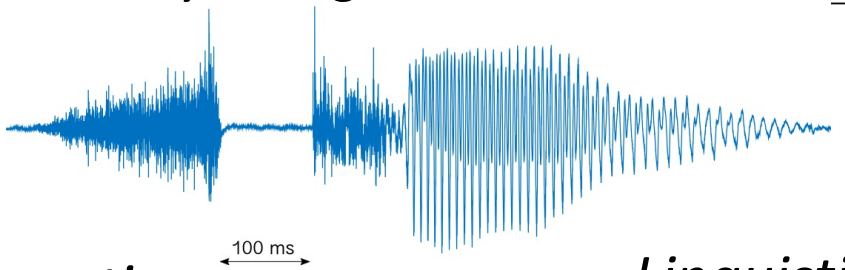
Psychoacoustique



p r o d u c t i o n



Physiologie



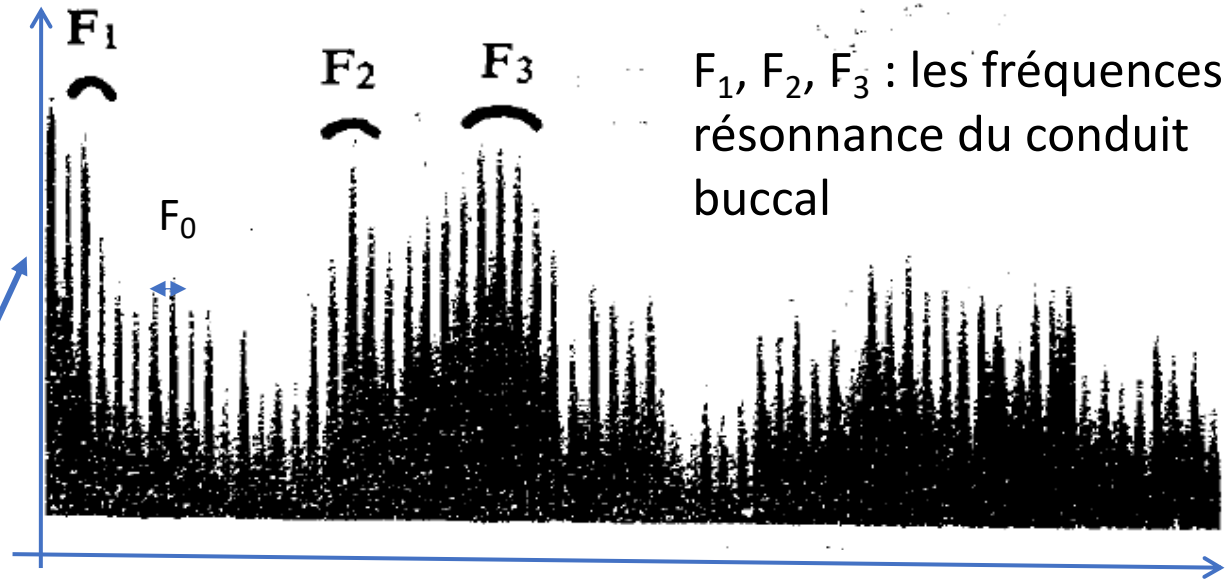
p e r c e p t i o n

Mécanique - Acoustique

Linguistique/Phonologie/Phonétique

Parole = sujet d'études pluridisciplinaires

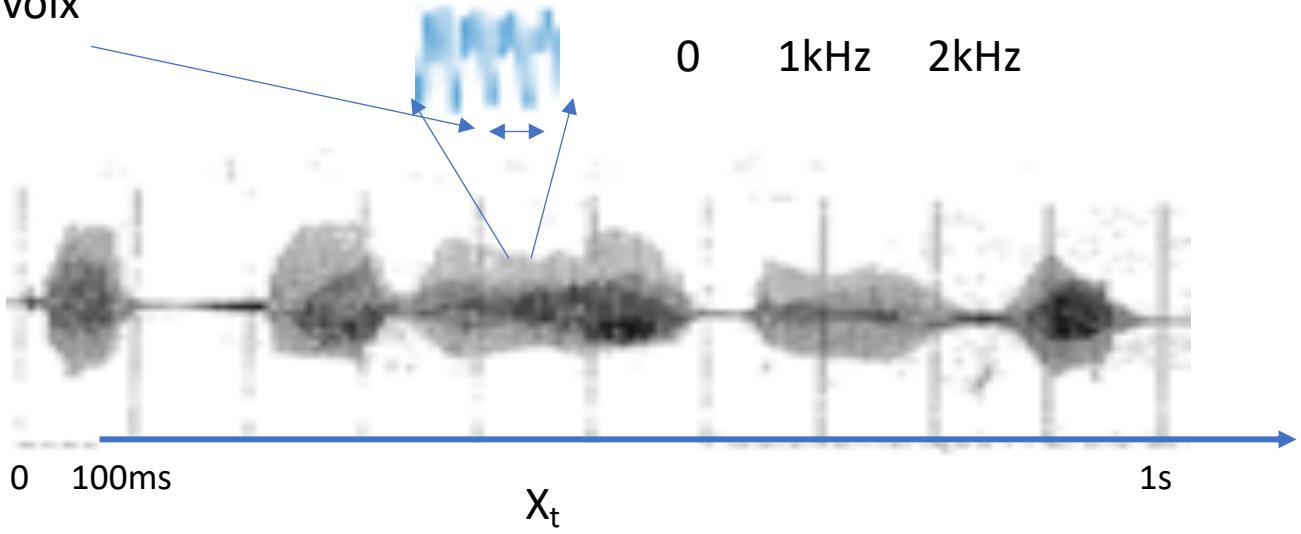
a priori : l'information parole est de nature spectrale …



F_1, F_2, F_3 : les fréquences de résonnance du conduit buccal

Transformée de Fourier sur **10ms**

F_0 : fréquence fondamentale intonation de la voix



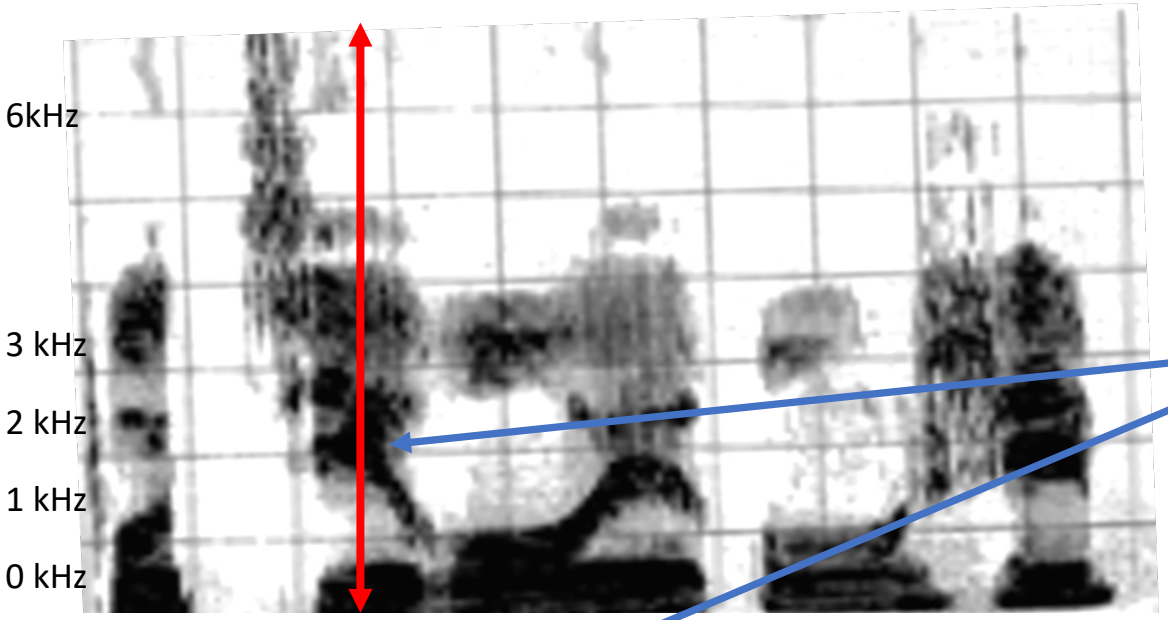
1 s de signal de parole
16 000 valeurs/s
(Fe = 16 kHz)

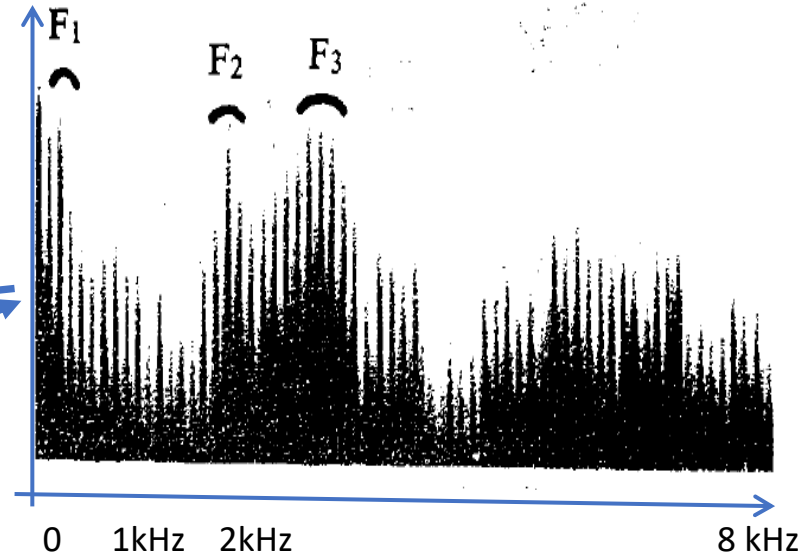
X_t

X_t vecteur de nature spectrale

a priori : l'information parole est de nature spectrale … + avec une dimension temporelle

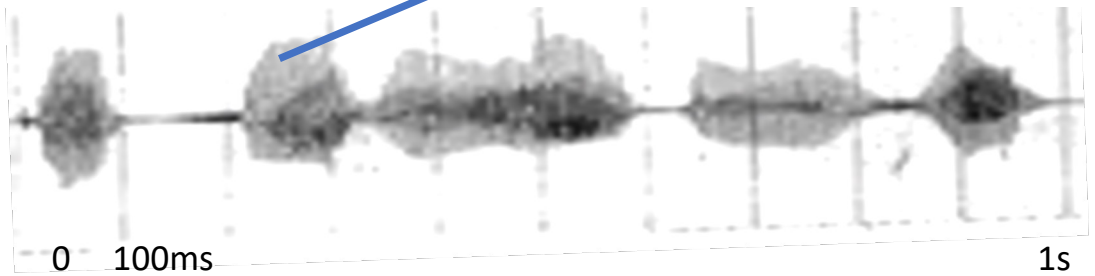
Spectrogramme





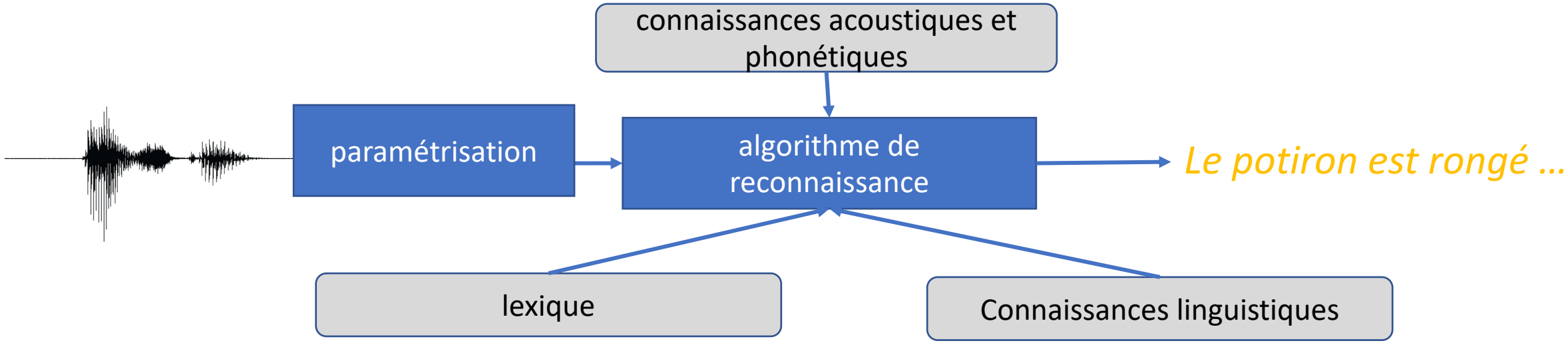
… *p o t i r on est r on g é* …

Signal de parole



$X_1, X_2, \ldots$ $X_t, \ldots$ X_T **T de l'ordre de 320, par seconde**

→ un système de reconnaissance automatique de parole



avec un signal de parole difficilement apprivoisable …

… les défis du traitement automatique de la parole

Signal de parole : combinaison « unique » d'informations très diverses

- Non reproductibilité de la production
- Variabilité temporelle (vitesse d'élocution, instabilité temporelle)
- Variabilité extrême intra locuteur et inter locuteur
- Dépendance extrême à l'environnement

→ **Prise en compte de la difficulté via des tâches contrôlées**

Un domaine d'études en marge de celui de l'IA jusqu'aux années 2010 alors que toutes les techniques utilisées relèvent de l'IA

L'intelligence artificielle des années « 50 »

Définition de M. Minsky et John Mc. Carthy (1956) :

« La construction de programmes informatiques capables d'accomplir des tâches qui sont, pour l'instant, accomplies de façon plus satisfaisantes par des êtres humains. »

Colloque scientifique – campus de l'université de Darmouth –NH –

20 chercheurs réunis pendant 2 mois dans une démarche pluridisciplinaire.

Marvin Minsky (MIT, 1927-2016), père des 1ères machines neuronales (circuits électroniques)

John Mac Carthy (1927-2011), père de la logique symbolique (LISP en 1958)

Claude Shannon (Bells, 1916, 2001), père de la théorie de l'information et de la cryptographie

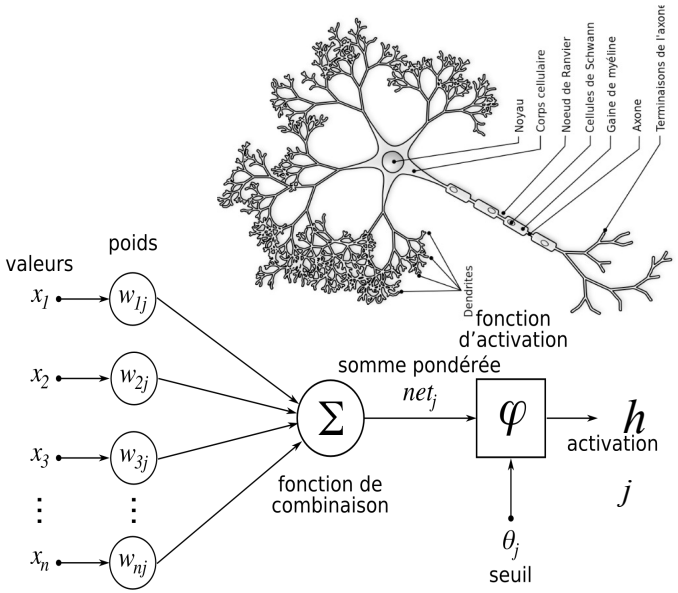
Franck Rosenblatt (1928, 1971), pionnier des réseaux neuronaux

Objectif central (1950-1970) fabriquer des systèmes qui remplacent l'homme pour des raisonnements complexes avec une saine compétition entre **deux écoles entre 1950 et 1960**

Approche connexionniste ←→ Approche symbolique

Le froid polaire de l'intelligence artificielle « 70 »

Approche connexionniste (bottom-up) ← → Approche symbolique (top-down)



Des Faits
Des Règles → Des Faits

L'homme est mortel
Socrates est un homme → Socrates est mortel

mais le froid s'installe en IA

dans les années 60

La non résolution de la fonction XOR par une machine neuronale

et dans les années 70

Le désastre en termes de performances des machines de traduction automatique

- Un avant-propos essentiel : la reConnaissance

- **Les deux approches des années 70-80 « analytique/globale »**

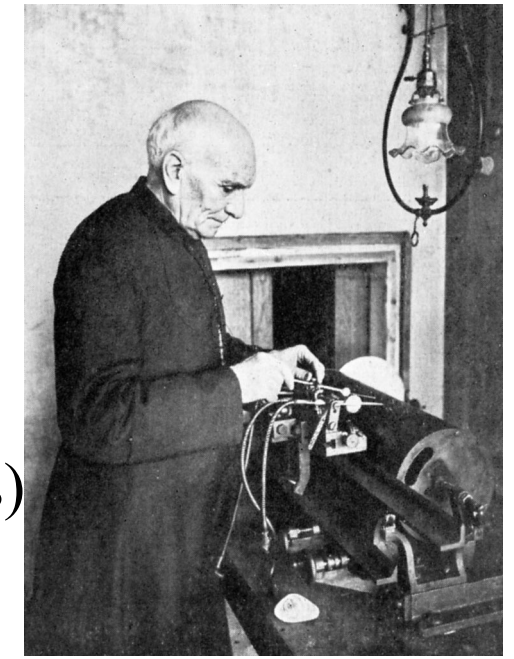
  ✓ Les systèmes analytiques type systèmes experts
  ✓ La reconnaissance globale

- L'intégration de la variabilité par approche probabiliste

- L'approche Deep Learning

- Jusqu'où ? Quelle reConnaissance ?

Les balbutiements de la reconnaissance automatique de parole



Connaissances anciennes

Phonétique expérimentale des sons d'une langue

(1897 – Abbé Jean Pierre Rousselot)

Phonétique articulatoire et acoustique

Analyse acoustique et l'électroacoustique (bancs de filtres analogiques)

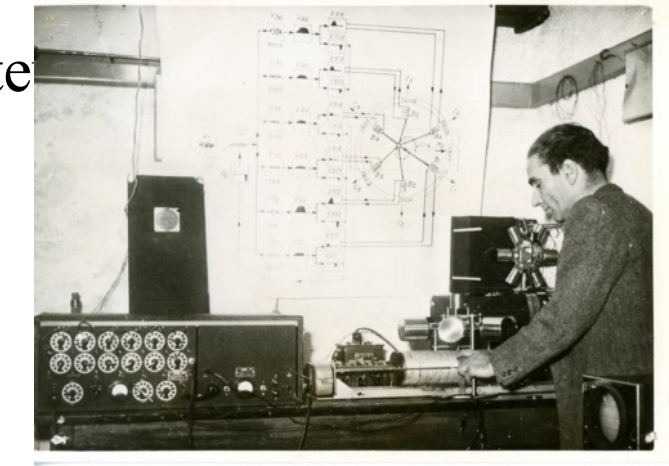
Machines

- 1952 : machine cablée pour reconnaitre 10 chiffres, monolocute

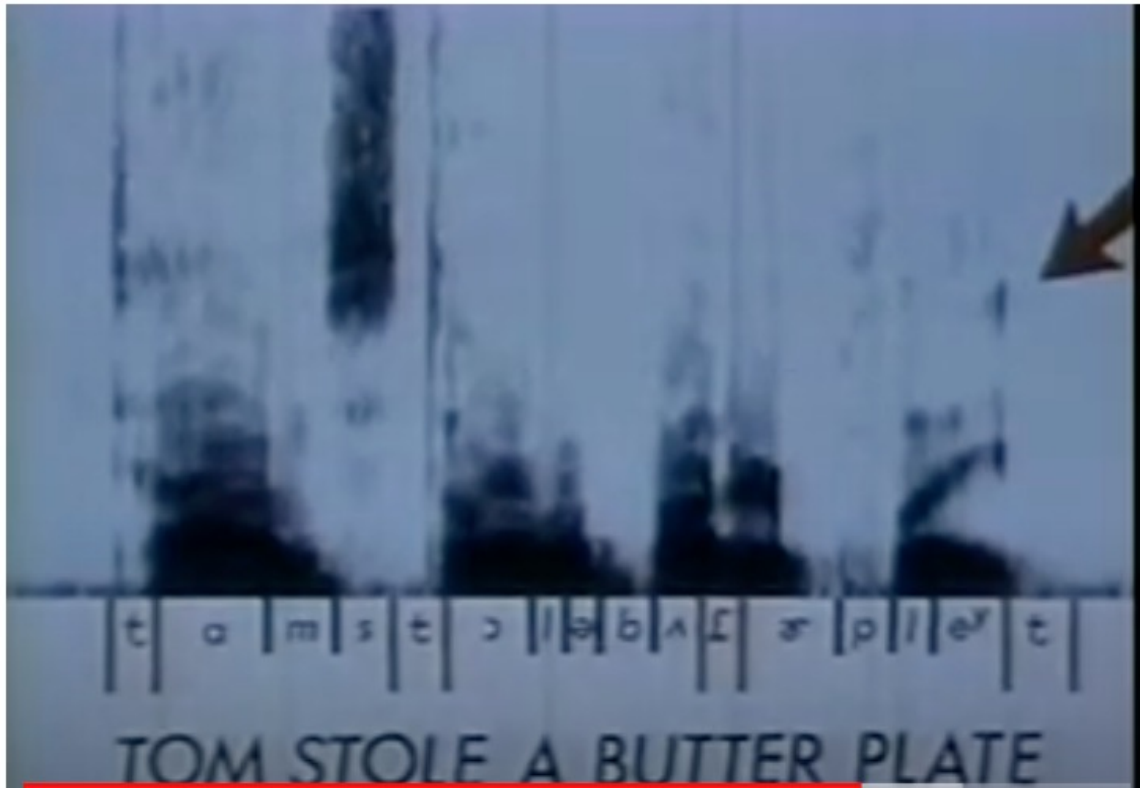
K.H. Davis, R. Biddulph et S. Baleshek

- 1961 : phonétographe pour identifier des phonèmes

J. Dreyfus-Graff



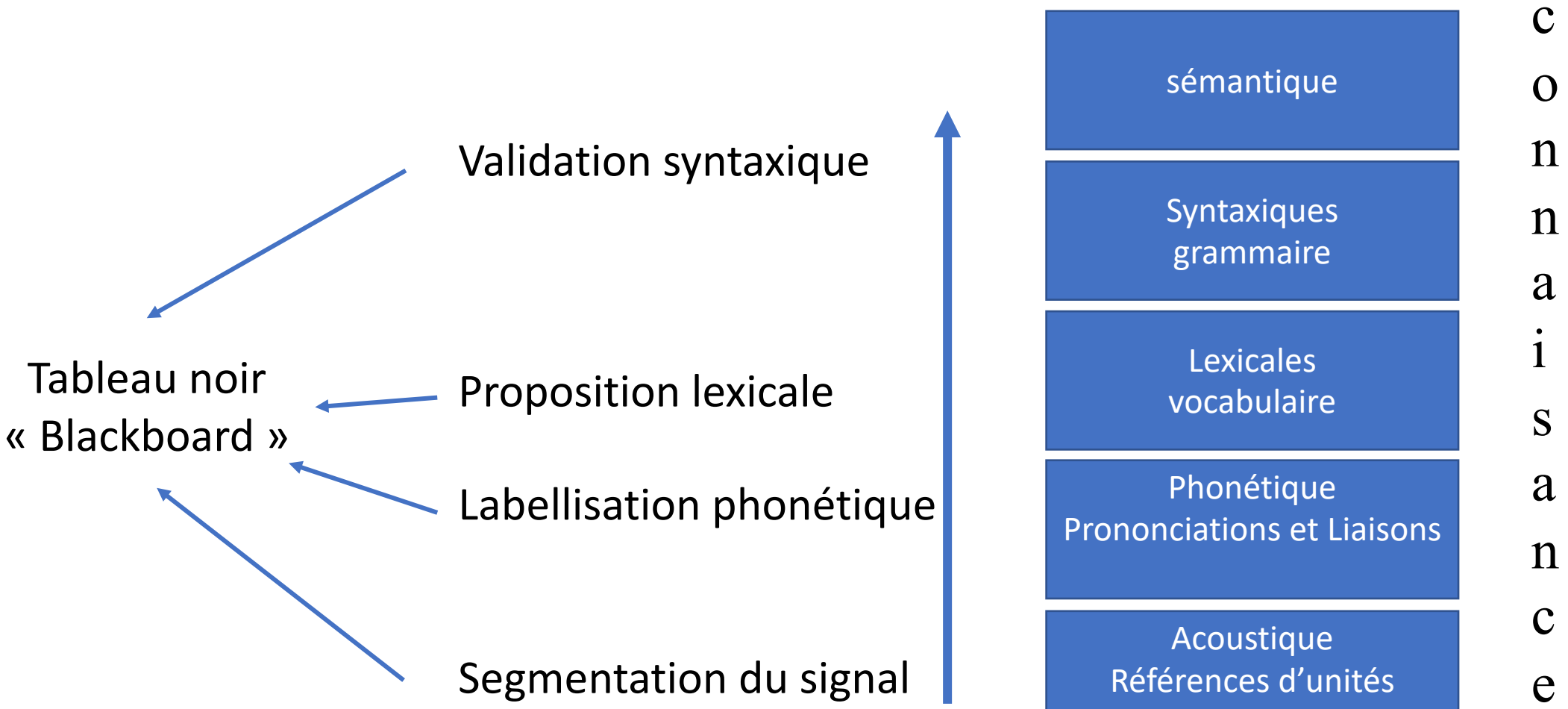
L'expertise humaine croissante dans les années 70-80



1979 – Victor Zue – enregistrement à CMU
Lecture d'un spectrogramme « Speech as eyes see it »

- Présence d'un phonéticien par projet de RAP
- Cours de lecture de spectrogramme

Approche analytique en RAP : « système « expert »



Quelques systèmes émergents aux EU et en France

1971 Projet Speech Understanding Research du DARPA (Defense Advanced Research Projects Agency)

→ HEARSAY II - CMU 1980 – Système monolocuteur, 1000 mots, parole connectée (10) → 80%

- MYRTILLE – CRIN/LORIA – Nancy 1978
- KEAL – CNET – Lannion - 1977
- ESOPE – LIMSI – Orsay - 1978

Mais aucun système expert ou à base de connaissances ne dépasse les performances du phonéticien

Limites / connaissances intégrées :

- Taille du vocabulaire et complexité syntaxique
- Compromis entre la complexité linguistique et le nombre de locuteurs
- Prise en compte du contexte (laboratoire, environnement réel…)

Une alternative performante, l'approche globale : comparaison de formes et alignement temporel

Apprentissage : recueil d'un ensemble de N prononciations dites « références »

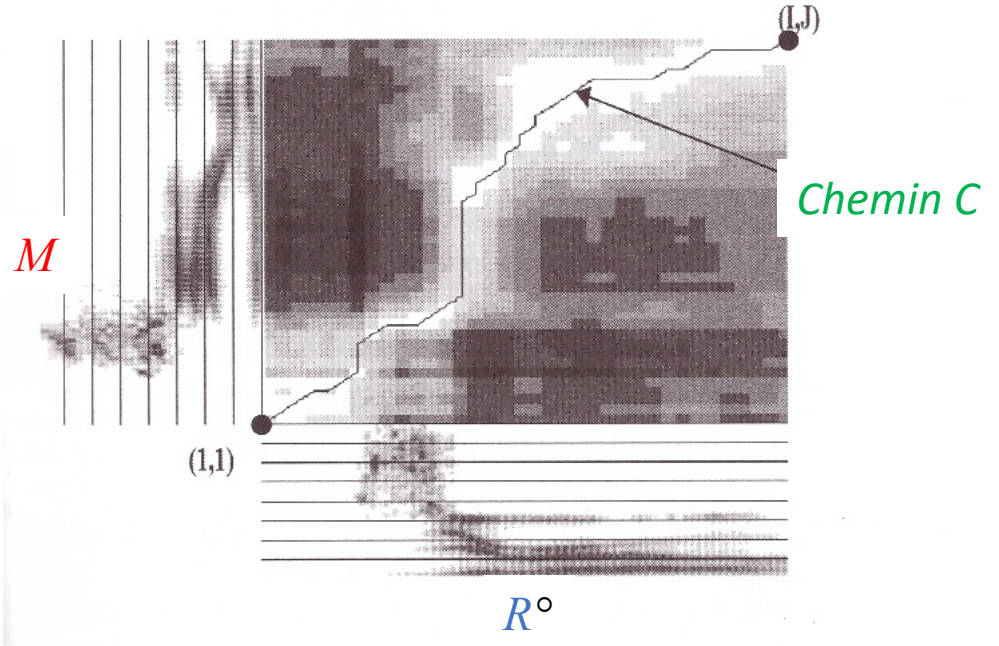
$R = \{R^1, R^2, \ldots, R^N\}$

$R^i = (r_1^i, r_2^i, r_{ni}^i)$ suite de Ni vecteurs spectraux

Reconnaissance : calcul de distance

$$D(M, R^\circ, C) = \sum_k w_C(k) D(m_{i(k)}, r^\circ_{j(k)})$$

$$D(M, R^\circ) = \min_C \left(\frac{D(M, R^\circ, C)}{\sum_k w_C(k)} \right)$$



+ algorithme de programmation dynamique (Bellman, Vintsjuk 1968, Sakoe, Chiba 1978)

Sakoe, H., and S. Chiba. "Dynamic Programming Algorithm Optimization for Spoken Word Recognition." IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 26, no. 1, Feb. 1978, pp. 43–49. Crossref, https://doi.org/10.1109/tassp.1978.1163055.

De nombreux systèmes avec d'excellentes performances…

Développement en laboratoire

Taux de reconnaissance en mots

- 1968 : Kiev - *Vintsjuk* – 1er système – 12 mots monolocuteur – 100%
- 1978 : *Sakoe, Chiba* (Nec) Suite de 1 à 4 mots - 10 chiffres – 5 locuteurs - 99%
- 1980 : CNET – *C. Gagnoulet** – Séraphine – 12 mots - indépendant du locuteur – labo : 96% - cabine téléph. 84%
- 1982 : LIMSI – *JL Gauvain*** – Mozart – 10 chiffres – 20 locuteurs – 84%

Commercialisation

1972 : VIP100 (Threshold Technology Inc), 32 mots isolés, monolocuteur, 20 000$

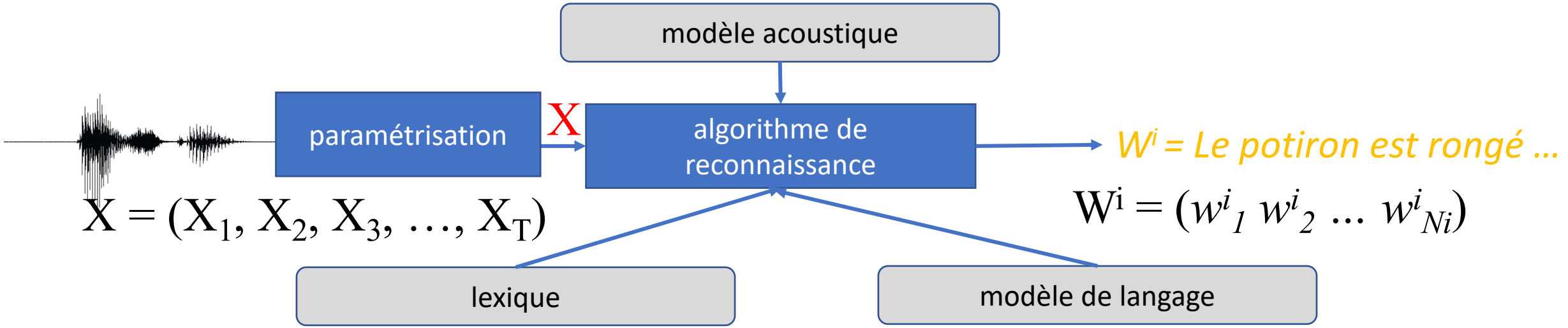
1978 : VRM carte de circuit imprimés (Interstate), 100 mots isolés, monolocuteur, 1 000$

*Gagnoulet, C., Couvrat, M., Jouvet, D. (1982). Seraphine: a connected word recognition system. In: Haton, JP. (eds) Automatic Speech Analysis and Recognition. NATO Advanced Study Institutes Series, vol 88. Springer, Dordrecht. https://doi.org/10.1007/978-94-009-7879-9_12. - 1982

**Gauvain JL, Mariani Joseph : Mozart: un système de reconnaissance globale de parole continue. 11e ICA, Paris 1982

- Un avant-propos essentiel : la reConnaissance
- Les deux approches des années 70-80 « analytique/globale »
- **L'intégration de la variabilité par approche probabiliste**
  - ✓ La modélisation acoustique et le modèle de langage
  - ✓ La reconnaissance automatique de parole continue « large vocabulaire »
- L'approche Deep Learning
- Jusqu'où ? Quelle reConnaissance ?

Modèles acoustiques et modèle de langage



Recherche de la meilleure suite de mots sachant l'écoute X,

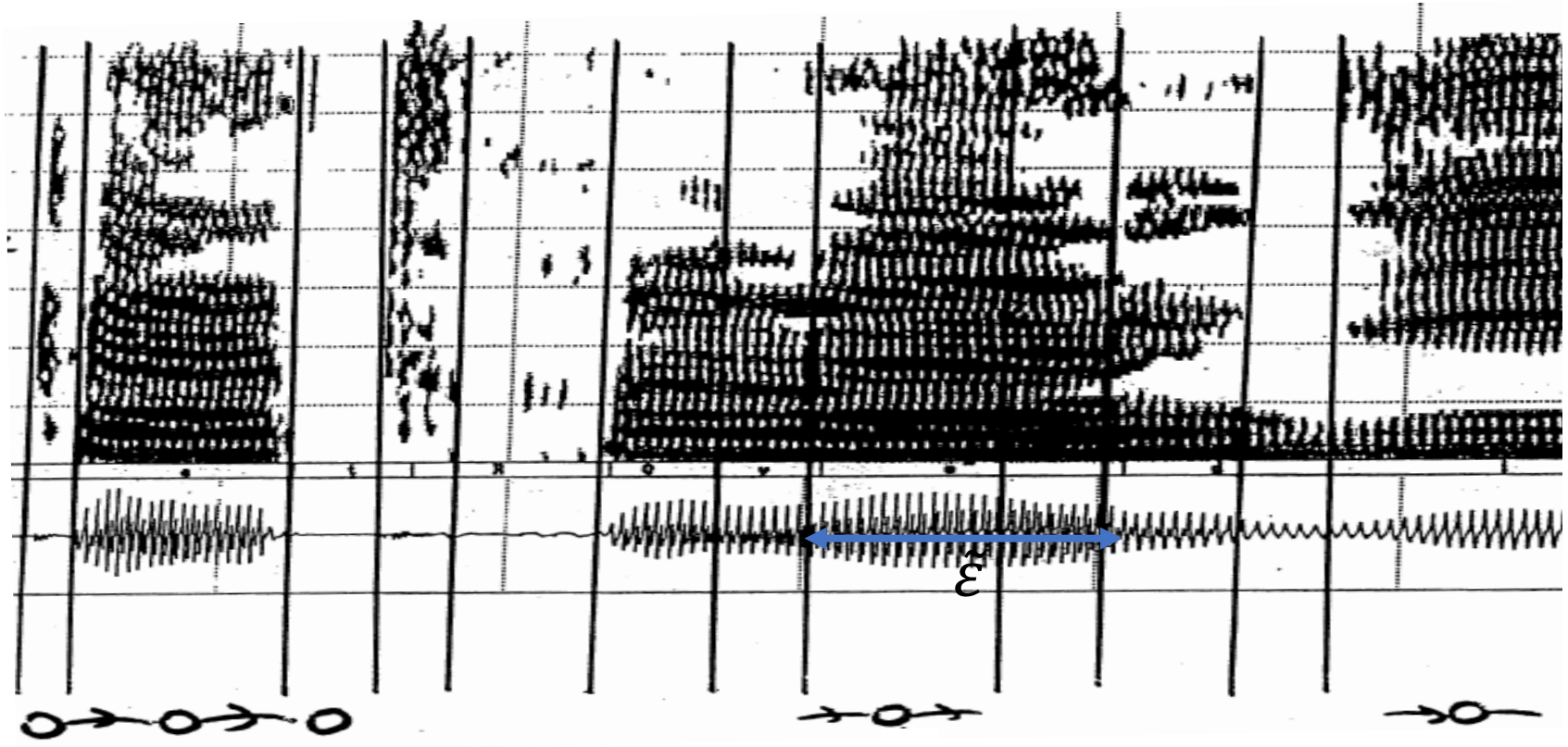
$$\widehat{W} = \boldsymbol{arg}\max_{j} \boldsymbol{P}(\boldsymbol{W^j}/\boldsymbol{X})$$

Formule de Bayes $\widehat{W} = \arg\max_{j} \boldsymbol{P}(\boldsymbol{X}/\boldsymbol{W^j})\boldsymbol{P}(\boldsymbol{W^j})/P(X)$

Modèle acoustique $P(X/ W^i)$: Modèles de Markov Cachés*

W^i = « 90 » = « quatre-vingt-dix »

X_1 …. …. $X_{t.}$ … X_T



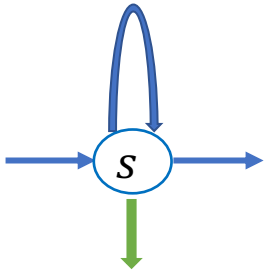
- Suite de mots : « 4 20 10 »

- Suite de phonèmes :
  / k a t r ∂v ε̃ d i s /

- Suite de sons élémentaires, chacun étant associé à une distribution probabiliste
  $P(X_t / \text{son})$

*Hidden Markov Models = HMM

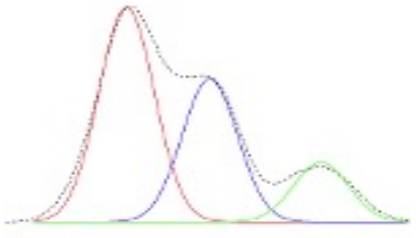
Modèle acoustique P(X/ W^i) : Modèles de Markov Cachés



Baker – CMU 1973. - Jelinek – IBM 1969-1975

1 modèle M_s par son élémentaire s

$$P(X_t/s) = \sum^{I} \lambda_s^i \, N(X_t, m_s^i, \Sigma_s^i)$$



- 1 modèle par mot du lexique par « concaténation » de modèles M_s
- Intégration des variantes de prononciation (parisien, sud ouest !)
- Inventaire de toutes les liaisons « inter-mot »

=> graphe probabiliste pour chaque phrase **W^i** probable

Mélange de lois gaussiennes (GMM)

Nombre de GMMs → jusqu'à 12 000
Dimension de X_t → de l'ordre de 50

} **pour capter la variabilité**

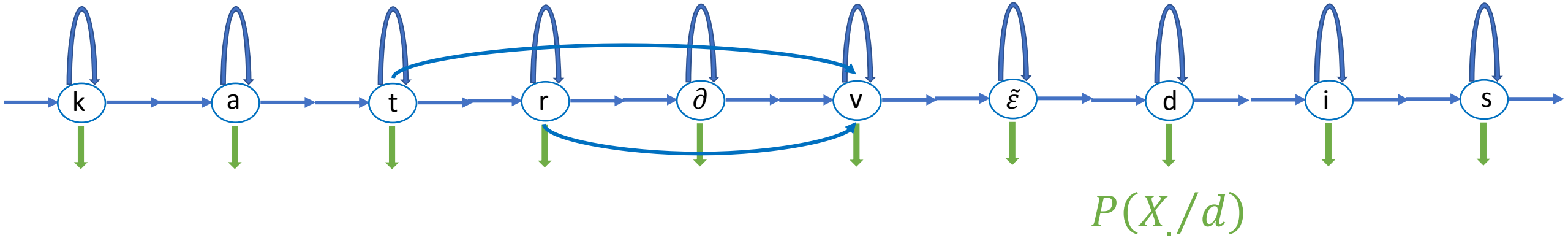
Rabiner, Lawrence R.. "A tutorial on hidden Markov models and selected applications in speech recognition." Proc. IEEE 77 (1989): 257-286.

Modèle acoustique P(X/ W^i) : Modèles de Markov Cachés

$$W = w_1 - w_2 - w_3$$

$$90 = 4 - 20 - 10$$

$$= \quad /k/ \quad /a/ \quad /t/ \quad /r/ \quad /\partial/ \quad /v/ \quad /\tilde{\varepsilon}/ \quad /d/ \quad /i/ \quad /s/$$



Avancées algorithmiques (1970 - 1980 - 1990)

- estimation des paramètres par l'algorithme de Baum-Welch
-Recherche du meilleur chemin dans un graphe par l'algorithme de Viterbi

Modèle de langage P (W) : le modèle n-gram

Un lexique ou dictionnaire de formes $\{ w^j, j = 1, J\}$

Un modèle probabiliste sur l'ensemble des phrases (!!!)

$$P(w_1\ w_2 \ldots\ w_M) = P(w_1)\ P(w_2/w_1) \ldots P(w_i/\ w_1\ w_2 \ldots w_{i-1}) \ldots P(w_M/\ w_1\ w_2 \ldots w_{M-1})$$

$$= P(w_1)\ P(w_2/w_1) \ldots P(w_i/\ w_{i-n+1}\ w_{i-n+2} \ldots w_{i-1}) \ldots P(w_M/\ w_{M-n+1} \ldots w_{M-1})$$

modèle probabiliste n-gram = passé limité à (n-1) avec $2 < n < 5$

Exple : W = « Les enfants en récréation joue » 2-gram → P(W) ≠ 0 (erreur linguistique et non acoustique
5-gram → P(W) = 0

Un système « phare » de reconnaissance « très grand vocabulaire »

Un exemple français : LIMSI* (**2005**) transcription des « **Broadcasts News** »

Modèles acoustiques pour 35 phonèmes

→ 65k mélanges de lois gaussiennes appris à partir de 90h audio

Modèle de langage : 200k mots, 4-gram, appris à partir de textes (500M mots)

Taux erreur mot : 10,7% sur 10h test avec des performances hétérogènes !!

Mais est-ce le début de la **perte des connaissances phonétiques et linguistiques** ?

Jelinek, responsable IBM, statisticien (dans les années 90) : *« Every time we fire a phonetician or a linguist, the performance of our system goes up »*

* J.L. Gauvain, G. Adda, M. Adda-Decker, A. Allauzen, V. Gendner, L. Lamel, H. Schwenk. Where AreWe In Transcribing French Broadcast News? Interspeech 2005

- Un avant-propos essentiel : la reConnaissance
- Les deux approches des années 70-80 « analytique/globale »
- L'intégration de la variabilité par approche probabiliste
- **L'approche Deep Learning**
  - ✓ Les systèmes hybrides (approche probabiliste + approche neuronale)
  - ✓ Les systèmes End-to-End (E2E)
- Jusqu'où ? Quelle reConnaissance ?

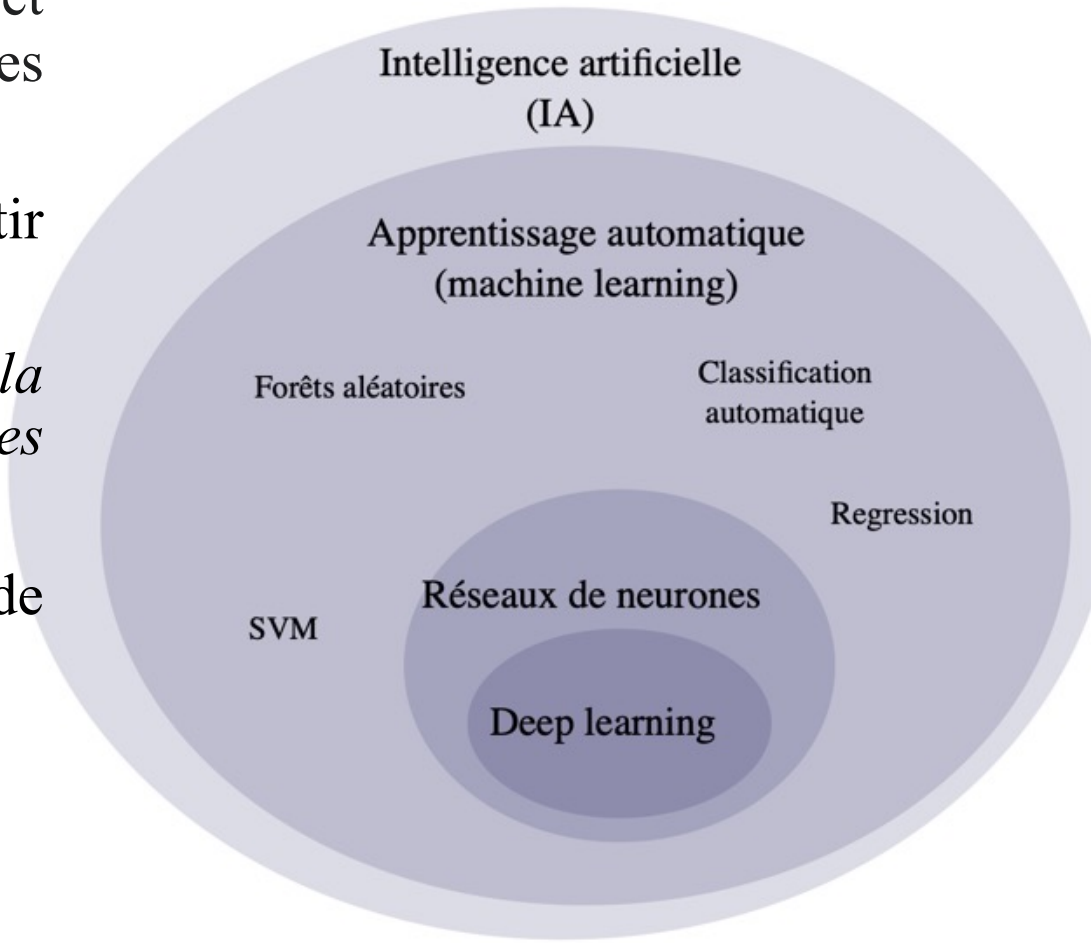
Réveil de l'intelligence artificielle (1990)

- **Intelligence artificielle** : Ensemble de théories et techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine
- **Machine Learning (1990)** : Apprentissage à partir d'exemples

Yann le Cun (2015) : [L'IA] « est inséparable de la capacité à apprendre, telle qu'on l'observe chez les êtres vivants ».*

- **Deep Learning** : Apprentissage basé sur des réseaux de neurones

Michael Jordan (2018), « c'est l'agenda intellectuel de Wiener qui domine aujourd'hui sous la bannière de la terminologie de McCarthy »

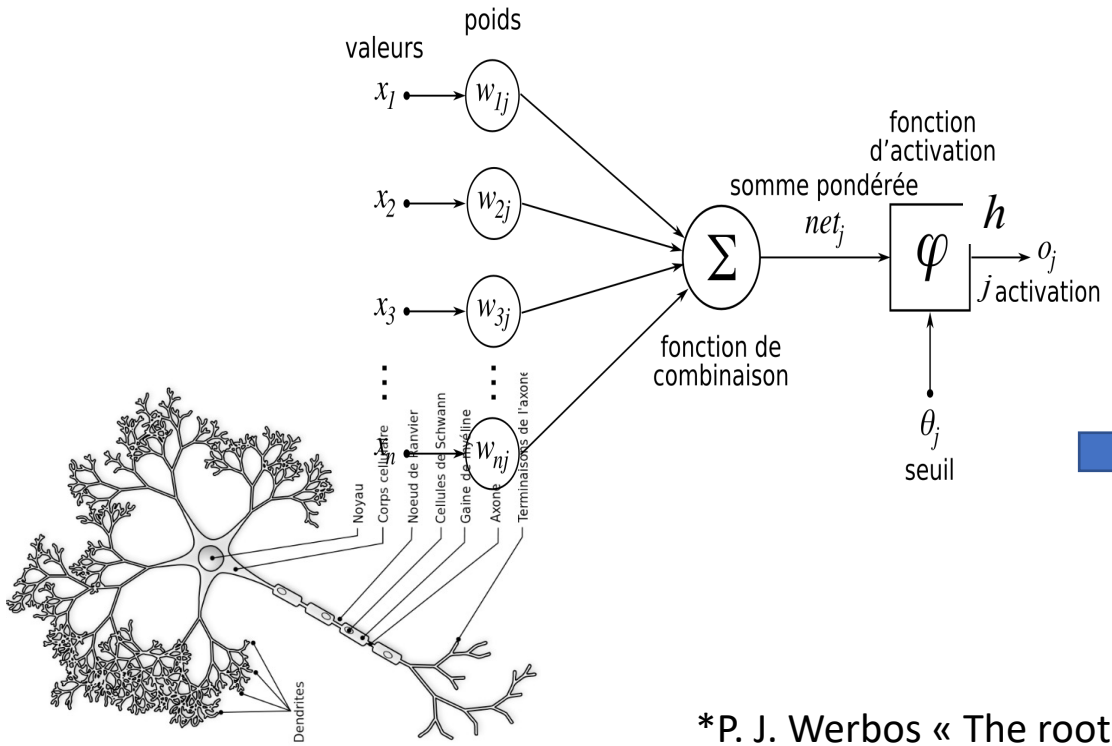


*LeCun, Yann, et al. "Deep Learning." Nature, vol. 521, no. 7553, May 2015, pp. 436–44. Crossref, https://doi.org/10.1038/nature14539.

Historiquement : le neurone formel et le perceptron multicouche

Neurone formel McCulloch et Pitts (1943)

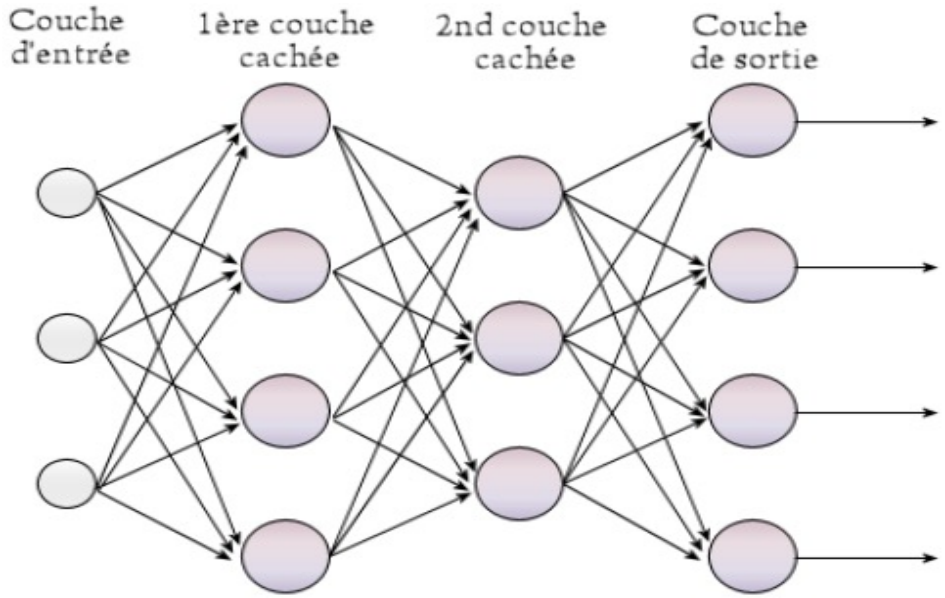
Algorithme : Perceptron de F. Rosenblatt (1957)



Perceptron multicouche « Multi-Layer Perceptron » (MLP)

Algorithme de rétropropagation du gradient

(PJ Werbos*-1974 – D Rumelhart** 1986)

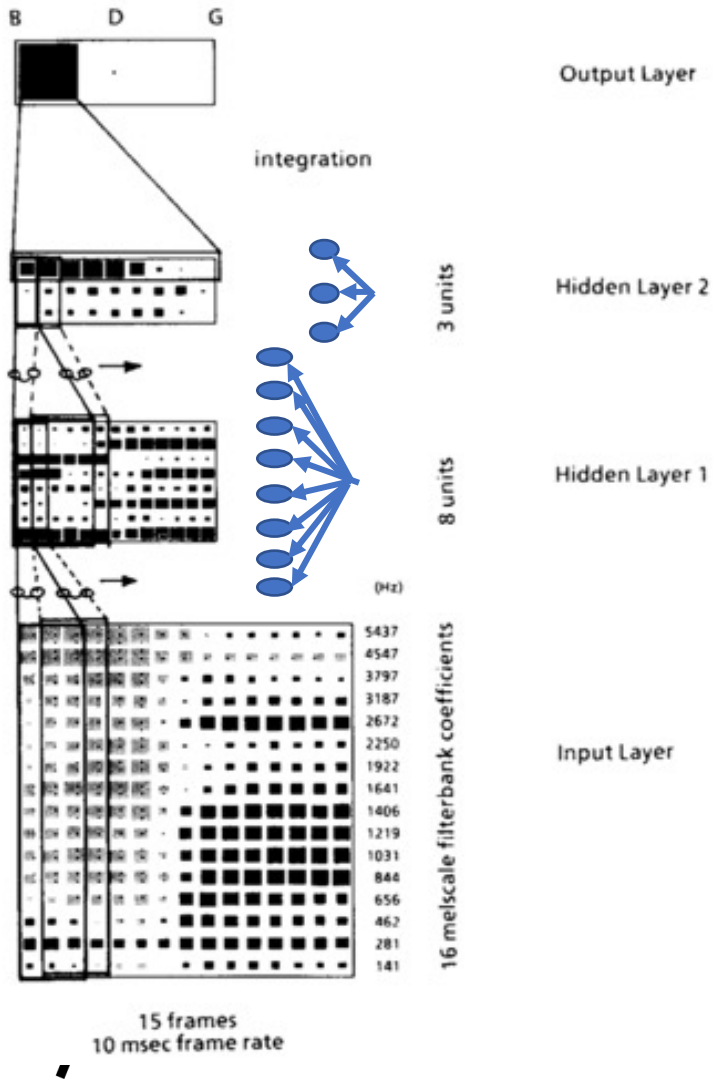


X_t

classifieur

*P. J. Werbos « The roots of backpropagation : from ordered derivatives to neural networksand political forecasting » New York: John Wiley 1 Sons, ISBN 0-471-59897-6

Rumelhart, David E.; Hinton, Geoffrey E.; Williams, Ronald J. (1986-10-09). "Learning representations by back-propagating errors". *Nature*. **323 (6088): 533–536



La reconnaissance du pouvoir discriminant d'un réseau de neurones, en parole dès 1989 (A. Waibel*)

1ère expérience : reconnaissance des trois occlusives voisées /b/, /d/, /g/ avec le Time Delay Neural Network (TDNN) à 2 couches cachées

1ère couche cachée de 8 unités
avec, à l'instant t, pour entrées (X_{t-1}, X_t, X_{t+1})
→ X_t de dimension 15 → 45 * 8 = 440 pondérations

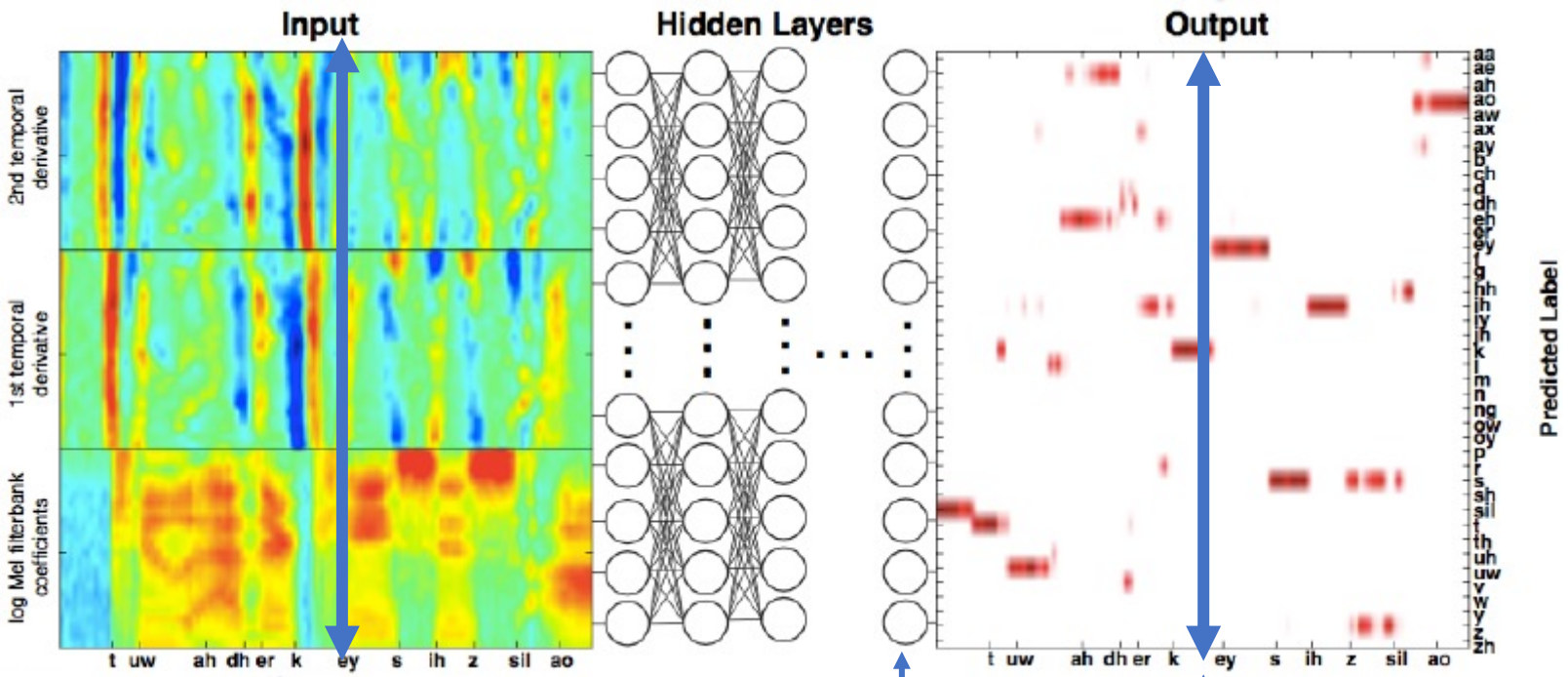
2e couche cachée de 3 unités, 1 unité par phonème,
avec pour entrées, les sorties de la 1ère couche à t-2, t-1, t, t+1, t+2
→ 40 * 3 = 120 pondérations

TDNN 98% > HMM (3 états) 93%

mais au prix d'une segmentation manuelle et de 560 poids

*Alexander Waibel : Phoneme Recognition using TDNN , ICASSP march 1989

Un modèle phonétique Deep Neural Network (DNN)- 2015



$X_t \in R^{(3*24)}$

1 neurone de sortie par phonème

$P(X_t/\varphi_i)$

DNN phonétique

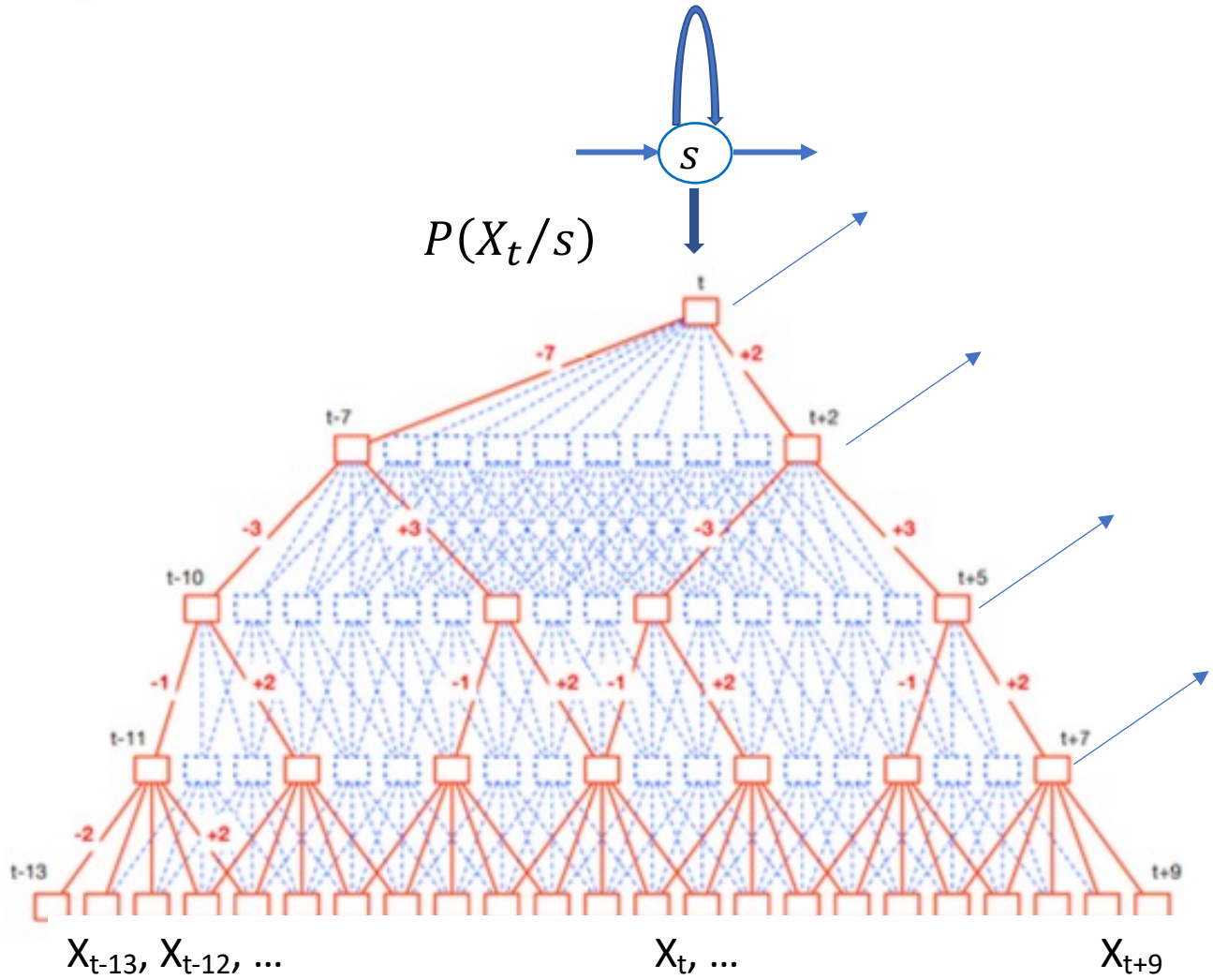
(Columbia University, Microsoft USA)

792 entrées (11 vecteurs *24*3)
5 couches cachées
256 unités par couche
41 sorties (40 phonèmes + silence)

Expérience (2015)
1000 phrases (TIMIT)
10 phrases par locuteur
51 hommes, 49 femmes

Tasha Nagamine, Michael L. Seltzer, Nima Mesgarani,Department of Electrical Engeering, Columbia University and Microsoft Research, Redmond, Exploring How Deep Neural Networks Form Phonemic Categories, Interspeech 2015

Une version étendue du DNN-HMM : le TDNN-HMM



Bourlard - 1990

Povey - 2018.
Plus de 20 Millions de paramètres
Modèle de langage 4-gram

Bases de données
- *Switchboard parole téléphonique (>2000h)*

9% < Taux erreur mot < 10%

Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohamadi, Sanjeev Khudanpur, « Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks », Interspeech 2019.

Depuis 2015, place à l'approche End-to-End (E2E)

Changement radical de paradigme

Relier la séquence audio à la transcription en « unités » sans modélisation intermédiaire

→ un seul apprentissage

$X = (X_1, X_2, ..., X_T)$. $\rightarrow Y = (Y_1, Y_2, ..., Y_N)$

Une évolution historique en trois étapes :

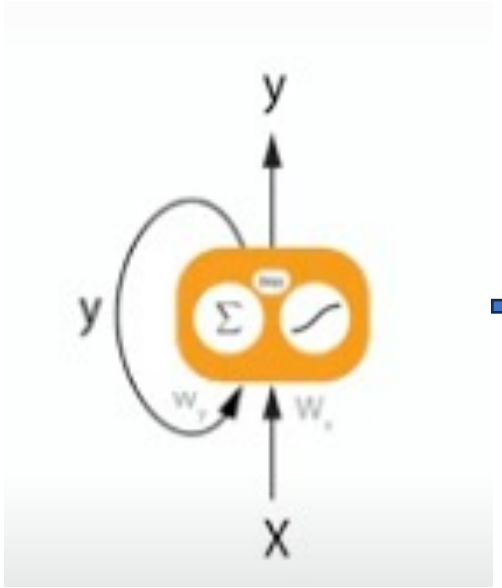
- L'approche « Connectionist Temporal Classification » CTC (2006), la version de base
- L'approche RNN-Transducer (Recurrent Neural Network) (2012).
- L'approche Transformer (encodeur-décodeur) avec attention (2015)

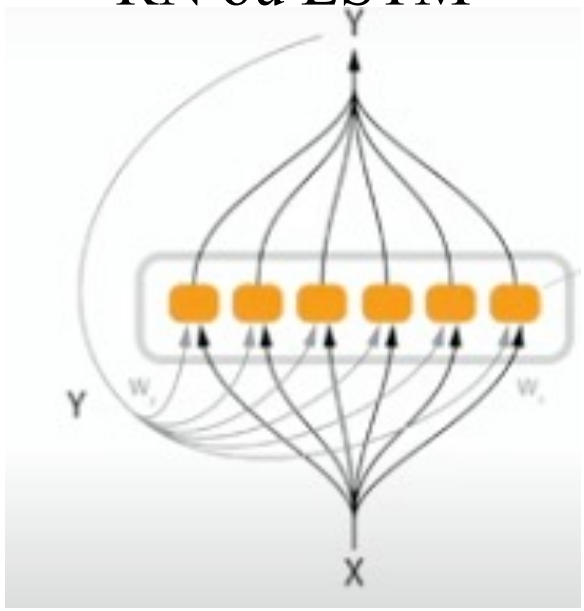
avec des réseaux de neurones qui se complexifient

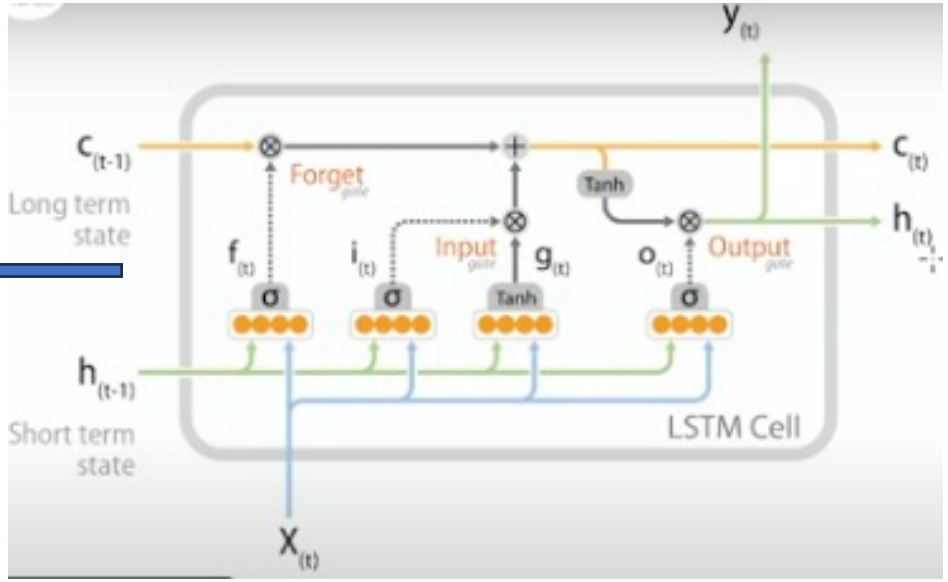
Avec une complexification des réseaux de neurones

Prendre en compte la dimension temporelle = Mémoriser le passé avec +/- d'efficacité

Neurone récurrent | Recurrent Neural Network RN ou LSTM | Long Short-Term Memory



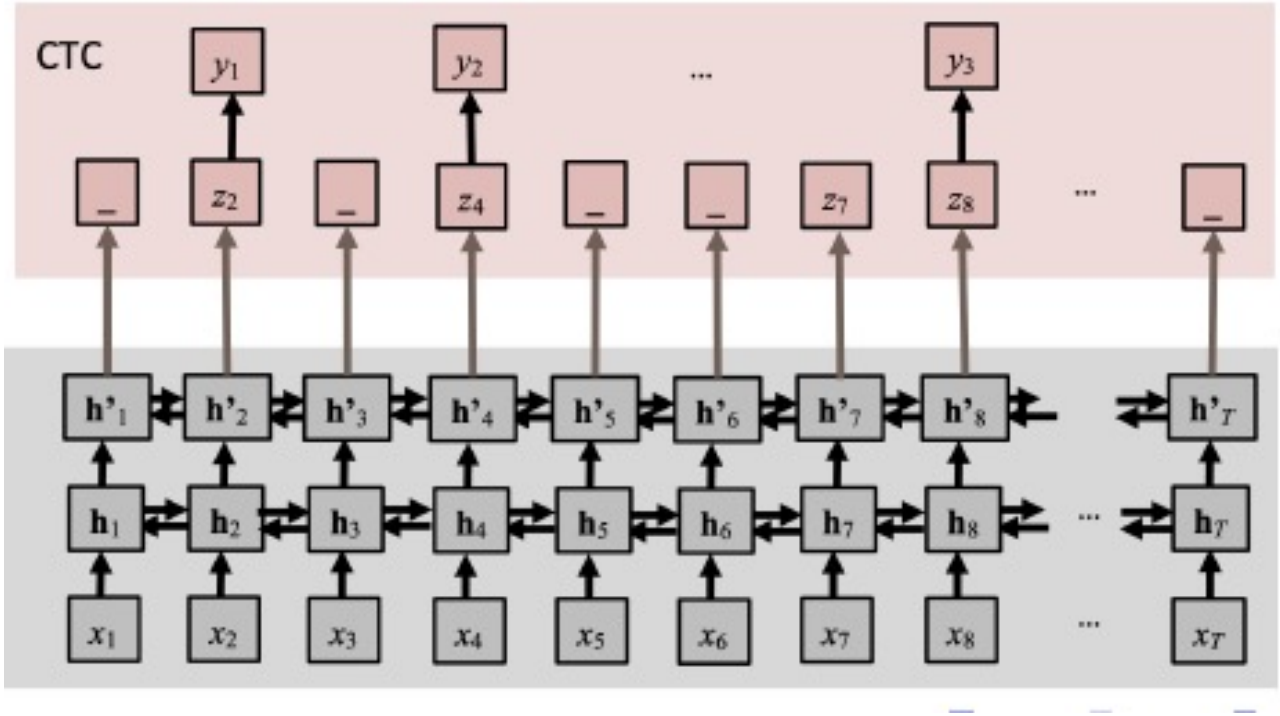




X_t vecteur d'entrée => **h_t vecteur de sortie intermédiaire du réseau**

L'approche E2E : le Connectionist Temporal Classification

Sorties : caractères, sous-mots …



Mauvaise propriété en parole
Indépendance des sorties

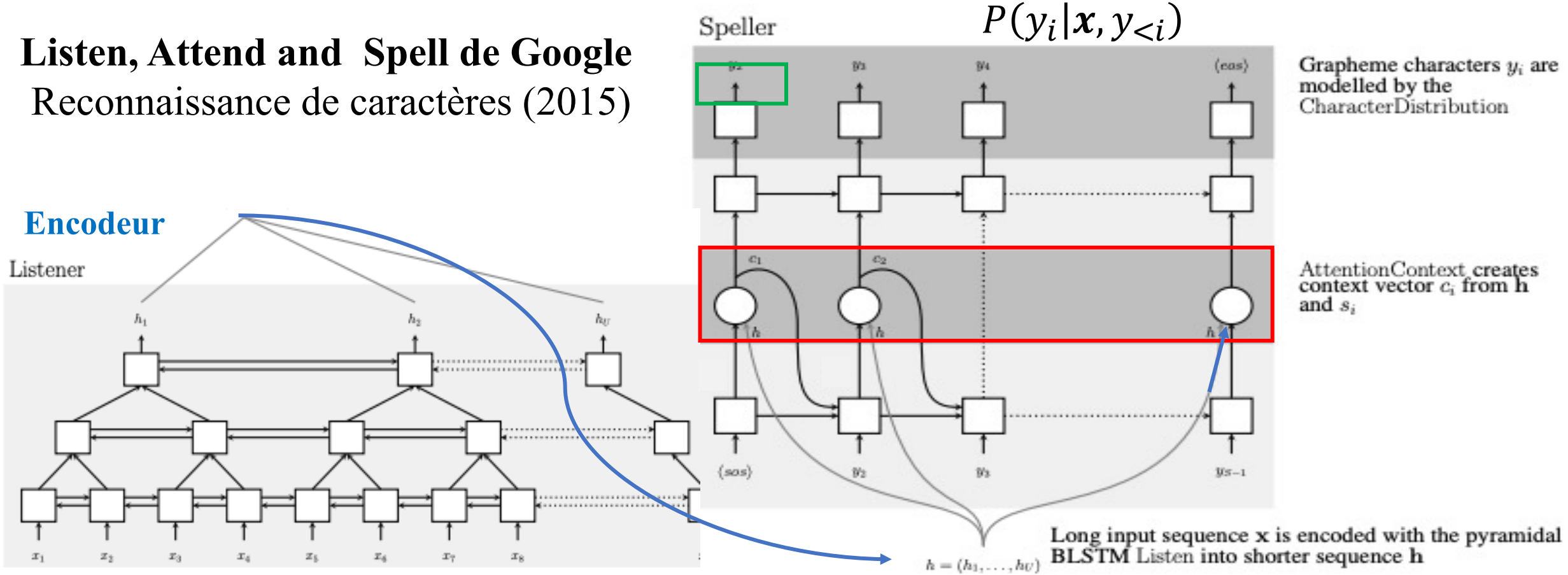
Complexité augmentée avec une birection « forward-backward »

L'approche E2E : l'encodeur-décodeur avec Attention

Listen, Attend and Spell de Google

Reconnaissance de caractères (2015)

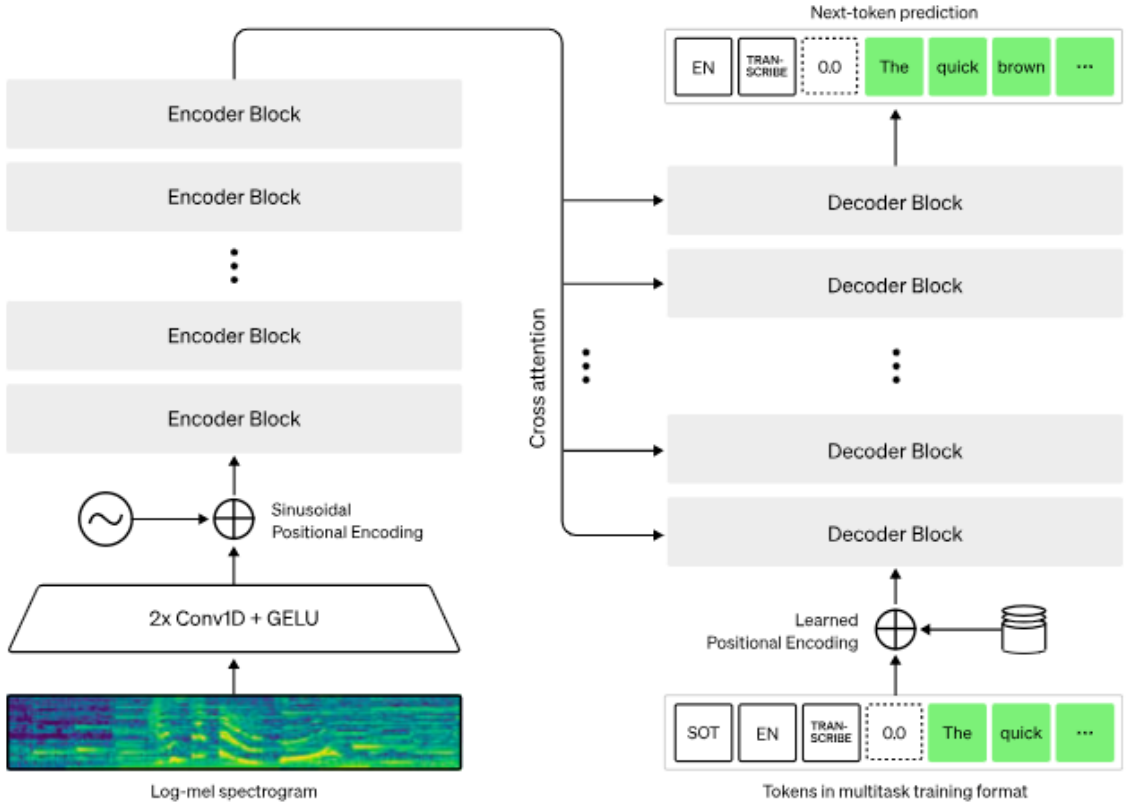
$P(y_i|\boldsymbol{x}, y_{<i})$



Taux erreur mot = 10,3%

William Chan (CMU), Navdeep Jaitly, Quoc V. Le, Oriol Vinyals (Google Brain), « Listen, Attend ans Spell », ICASSP 2016

L'approche E2E : le système Open AI (Whisper)*



Entrée : 30s parole

Nombre de paramètres de 39 M à 769M

Données : 680k heures internet transcrites

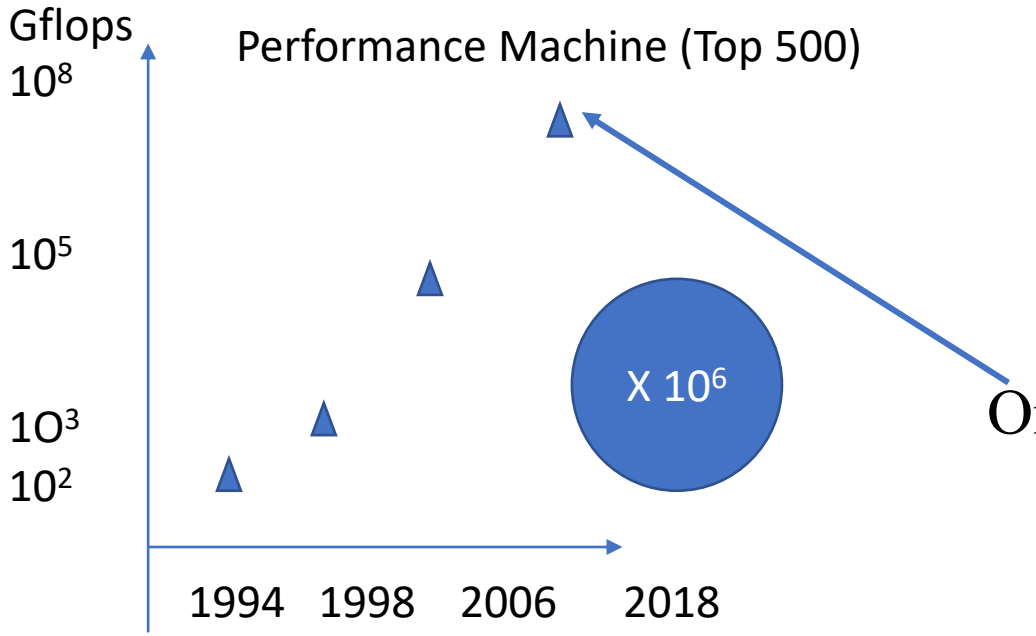
« taux erreur mot » toutes conditions : 12,8 % (2022)

- https://openai.com/research/whisper

Avec quels moyens ?

Des moyens calcul

- ✓ 1977, Cray1: 160 Mflops,
  5 tonnes, 8 Millions $
- ✓ 2020, carte graphique d'un PC/jeux:
  10 TFlops (*60 000), 300€.



Des bases de données

CommonVoice : 9 283h
7 335h validées dans 60 langues

LibriSpeech. : 1 000h d'audiobooks
https://paperswithcode.com/dataset/librispeech
Avec plus de 2000 locuteurs
+ modèles n-gram (977 k mots différents)

Jusqu'à 680k heures de données transcrites

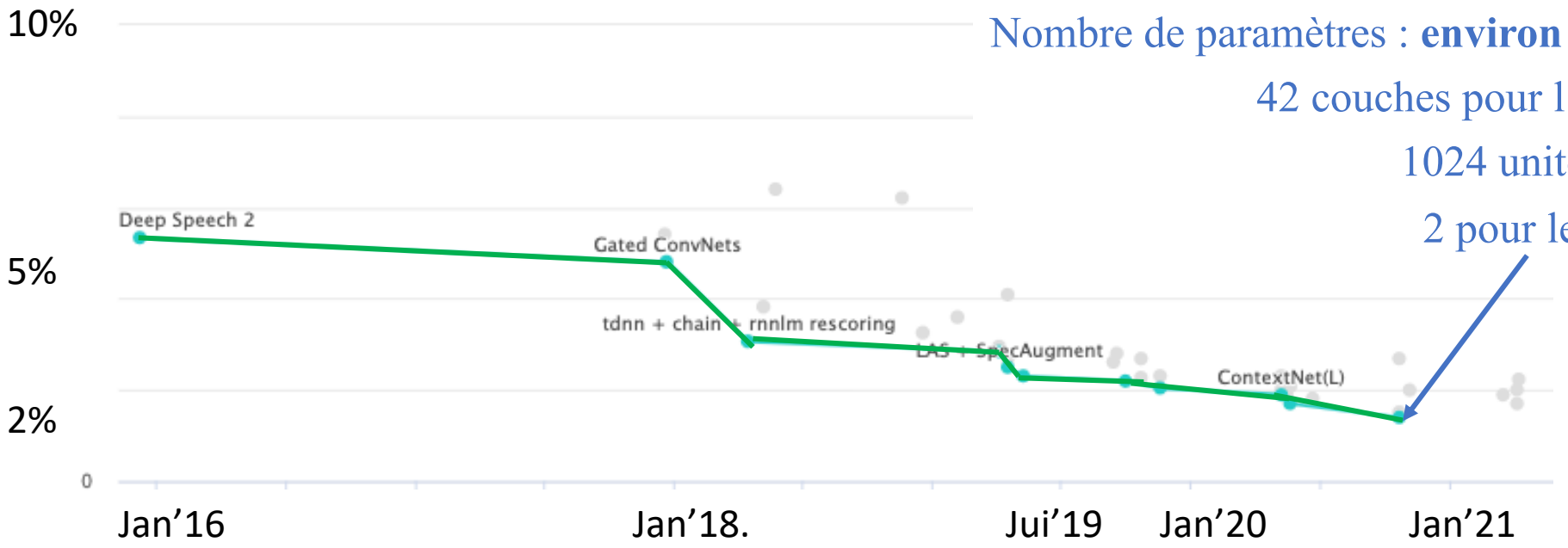
Ordinateur Jean Zay CNRS = 28 pétaFlops (10^{15})

… pour quelle performance ?

Evolution des performances

(Benchmark https://paperswithcode.com/sota)

Taux erreur mot sur Librispeech clean (audiobook 5h)



Google juillet 2020

Nombre de paramètres : **environ 1 billion**

42 couches pour l'encodeur,

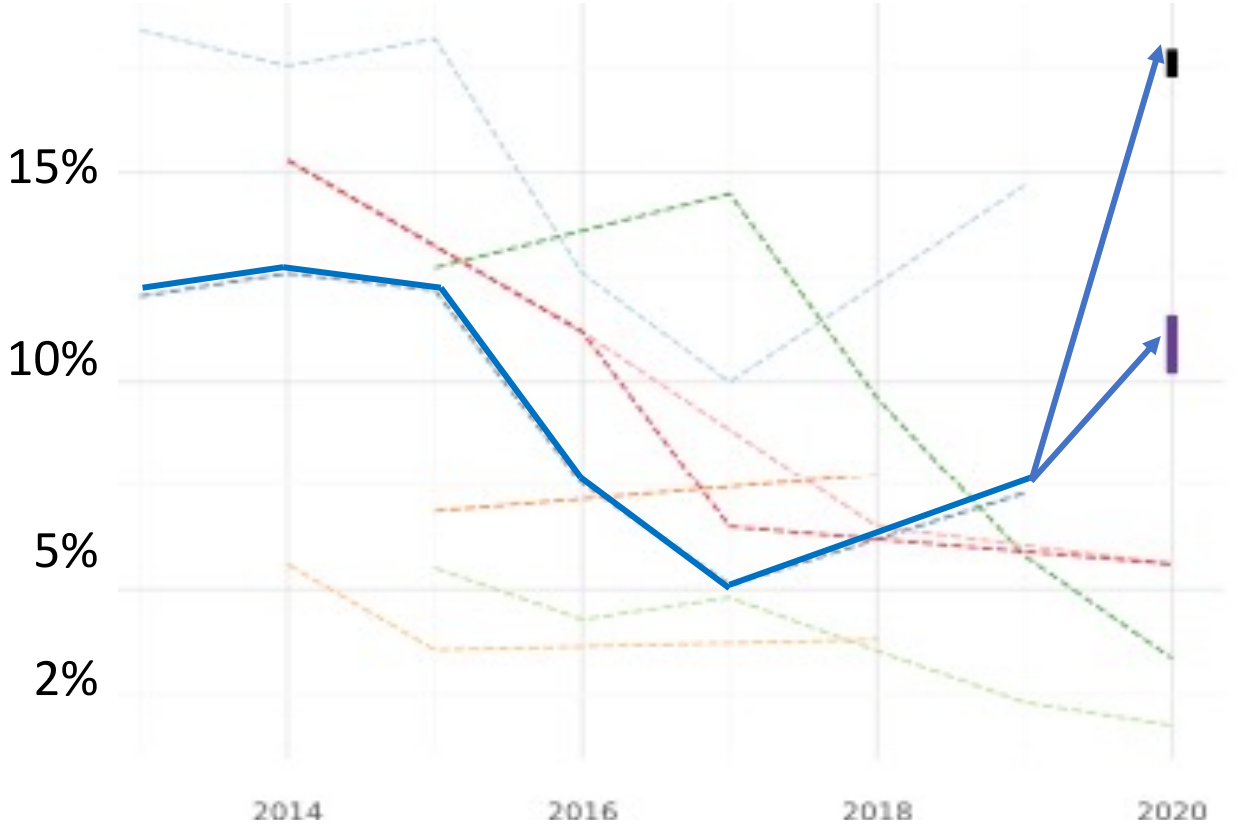
1024 unités/couche,

2 pour le décodeur

Mais il reste un long chemin à parcourir pour atteindre une reconnaissance solide de conversations humaines spontanées

… pour quelle réelle performance ?

« taux-erreur-mot » en fonction de la base de données



Expérimentation avec des systèmes commerciaux

Switchboard 10% - 12%

Call Center 17% - 19%
Assurance. 18% - 20%

Extrait de « WER we are and WER we think we are » 2020

Piotr Szymanski, Piotr Zelasko, Mikołaj Morzy, Adrian Szymczak, Marzena Zyla-Hoppe, Joanna Banaszczak, Łukasz Augustyniak, Jan Mizgajski, Yishay Carmiel, « WER we are and WER we think we are » October 2020
Conference: Findings of the Association for Computational Linguistics: EMNLP 2020

Quelle conclusion ?

Des performances remarquables grâce au Deep Learning, égales à un humain dans des contextes spécifiques (assistant vocal).

Limites actuelles :

- La parole atypique (enfants, non natifs …)
- La compréhension dans n'importe quel contexte
- La communication et le dialogue, la **vraie parole spontanée**
- La composante visuelle

→ Toujours plus de données – des données enrichies et libres d'accès ?
→ De nouveaux algorithmes en Deep Learning?

- Une augmentation artificielle des données (**Data Augmentation**)
- Une meilleure adéquation au contexte d'application par des apprentissages successifs (« **Fine Tuning** »)
- …

reConnaissance → Analyse Globale → Approche Probabiliste → Approche Connexionniste

Quelle conclusion ?

?

Existe-t-il une limite dans l'apprentissage machine des connaissances ?
Quelles connaissances sont apprises ?

Aide précieuse pour acquérir d'autres connaissances :

- Apprentissage d'une langue (CP ou L2)
- Voix pathologique
- …

Un grand merci à Abdelwahab Heba, mon dernier doctorant, chercheur chez Microsoft Speech Research