L'AVENTURE DE L'IA AU SERVICE DE LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE : *« LA PLACE DES CONNAISSANCES DANS L'ÉLABORATION DES SYSTÈMES »*

Par Régine ANDRÉ-OBRECHT[1]

Si de nos jours il est évident que l'élaboration de systèmes de reconnaissance automatique de la parole relève du domaine de recherches en Intelligence Artificielle (IA) en relation directe avec les notions d'apprentissage-machine (*Machine Learning*) et apprentissage profond (*Deep Learning*), ce n'était pas le cas auparavant : en dehors de la période des systèmes experts développés dans les années 80, jusqu'au début des années 2000, les deux communautés scientifiques que sont celles du traitement automatique de la parole et celles de l'IA s'ignoraient ! Il faut attendre ces dernières années pour que les communautés « reconnaissance de la parole » et « traitement du langage naturel » travaillent ensemble.

Comme nous allons le voir au travers d'un survol historique, la reconnaissance automatique de la parole doit affronter les défis de l'apprentissage automatique de connaissances en parole. Selon les périodes, cet apprentissage prend plusieurs formes, mais il a très longtemps été fortement supervisé.

La « ReConnaissance » de la parole

La parole est le moyen de communication entre êtres humains le plus efficace parmi tous les possibles (écriture, dessin, geste…), mais aussi le plus complexe. La production, la perception, la reconnaissance et la compréhension de la parole sont des étapes essentielles du processus de communication parlée et elles sont étroitement liées à une phase d'apprentissage entreprise dès les premières heures de la vie. Il en est de même pour les systèmes automatiques qui tendent à créer une communication homme-machine ; il n'est pas absurde de penser que leur développement implique des modèles nécessitant d'importantes phases d'apprentissage. En fonction du développement technologique, cet apprentissage va plus ou moins s'appuyer sur les connaissances acquises *a priori* en parole.

Quelques connaissances en parole

La parole met en jeu un système de production et un système de perception, tous deux très complexes et très dépendants l'un de l'autre, impliquant de nombreux acteurs et procédures :

---

1 Communication présentée à l'Académie des Sciences, Inscriptions et Belles-Lettres de Toulouse à la séance du 10 novembre 2022.

- les organes de production sont les poumons, les cordes vocales, le conduit buccal (mâchoire et langue), épisodiquement le conduit nasal et les lèvres,
- l'onde produite est une onde acoustique, résultat de la mise en mouvement et de l'articulation des organes de production ; elle se propage de manière turbulente dans ces conduits, puis dans le milieu environnant, de manière dynamique et non reproductible à l'identique,
- la phrase prononcée est une suite de sons élémentaires ou phonèmes, nécessairement coarticulés du fait de leur production et enchaînés pour former une suite de mots obéissant à la syntaxe de la langue employée,
- l'oreille transforme le signal acoustique en un signal mécanique au niveau de l'oreille moyenne, puis en un signal électrique au niveau de l'oreille interne et du nerf cochléaire.

Il en résulte que le signal de parole est une combinaison « unique » d'informations de nature cognitive, neuromusculaire et physiologique, extrêmement variable. La physiologie du locuteur (dont le sexe et l'âge), son origine géographique (accent) comme sociale contribuent à la variabilité interlocuteur. La non reproductibilité des commandes neuronales, le contexte articulatoire dynamique, la prosodie empruntée (intonation, vitesse d'élocution, intensité sonore) et l'environnement extérieur (ambiance sonore, état émotionnel) sont les principaux facteurs de variabilité intra locuteur. Ajoutons que l'objectif essentiel de l'acte de parole étant la communication, les effets de la « loi du moindre effort » accentue cette variabilité : la production s'adapte de manière dynamique à la perception !

La préhistoire de la reconnaissance de la parole

L'apparition de la phonétique expérimentale ou instrumentale, à la fin XIX[e] siècle (abbé Jean-Pierre Rousselot[2], 1846-1924) a établi un lien très fort entre les connaissances *a priori* et la reconnaissance de la parole. Accompagnées de l'analyse acoustique et des avancées en électroacoustique avec les bancs de filtres analogiques, ces connaissances permettent la conception des premières machines dites parlantes[3] dans les années 50 et la création du phonétographe ou sténo-sonographe de Jean Dreyfus-Graf pour reconnaître des phonèmes (Dreyfus-Graf, 1953).

La reconnaissance automatique de la parole

À compter des années 60 jusqu'aux années 2010, l'architecture d'un système de Reconnaissance Automatique de la Parole (RAP), à l'image des systèmes de reconnaissance des formes, s'impose comme composée de deux modules : un module de paramétrage qui vise à extraire l'information pertinente des *stimuli* du monde réel et un module de décision qui la transforme en une suite de catégories définies *a priori*. Les deux modules sont élaborés à partir des connaissances du domaine. En parole, l'entrée est le signal de parole et les catégories sont, selon les applications, des phonèmes, des mots, des suites de mots ou des phrases.

2 *Principes de Phonétique expérimentale* 1897 et 1901

3 1952 : machine câblée pour reconnaitre 10 chiffres utilisable par un seul locuteur, réalisée par K.H. Davis, R. Biddulph et S. Baleshek.

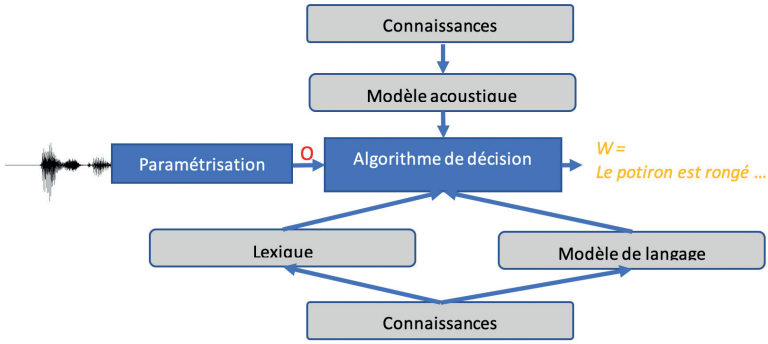

Fig. 1 RAP : architecture d'un système de reconnaissance automatique de la parole

La compréhension du système auditif humain et la phonétique expérimentale conduisent à privilégier, comme paramétrisation, les informations issues d'analyse fréquentielle à court terme, réalisée sur des fenêtres temporelles d'environ 20ms. La paramétrisation s'enrichit avec les connaissances issues de la psychoacoustique expérimentale : la prise en compte d'une échelle fréquentielle perceptive comme l'échelle Mel est essentielle.

Le module de décision exploite des connaissances acoustico-articulatoires et des connaissances linguistiques. Les premières sont issues de la phonétique expérimentale qui met en relation réalisation acoustique et caractérisation articulatoire : la mise en correspondance entre le triangle acoustique (caractérisation des voyelles par les deux harmoniques les plus significatives du spectre, appelées formants) et le triangle articulatoire (caractérisation des voyelles par le positionnement de la langue et l'ouverture de la mâchoire) en est un exemple. Les connaissances linguistiques sont d'ordre phonotactique, lexical et syntaxique ; elles précisent les enchaînements possibles de phonèmes et de mots en fonction de la langue utilisée.

De fait, dans tout système de RAP, il est nécessaire de savoir :

- quels mots peuvent être prononcés. Ils composent le lexique de l'application. Sa taille sera croissante pour passer 1000 dans les années 70 à plus de 250 000 en 2010, 600 000 en 2022,
- quels enchaînements de mots sont acceptables selon l'application. Trois situations sont répertoriées : tout mot peut suivre tout autre mot ; des règles donnent explicitement les phrases acceptables ; l'enchaînement des mots est évalué statistiquement.

La prise en compte exclusive de ces connaissances implique qu'aucun mot absent du lexique de l'application ne peut être reconnu. Les systèmes de reconnaissance automatique de suite de mots sont évalués au travers du nombre de mots non reconnus (*Word Error Rate* - WER) prenant en compte le nombre de mots omis, insérés ou substitués.

L'évolution des systèmes de RAP est liée à la modélisation utilisée au sein du modèle de décision, elle-même liée à la manière de prendre en compte des connaissances. Il est admis de distinguer trois grandes périodes. Au début, la décision est prise par comparaison à des références ou exploitation de connaissances *a priori*, puis elle devient probabiliste après estimation statistique des paramètres de modèles markoviens. Ces

dernières années, la prise en compte des connaissances dans la modélisation est remise en cause pour laisser place à un total « apprentissage machine » avec l'arrivée des réseaux de neurones.

Les années 60-80 : deux approches opposées - analytique *vs* globale

La recherche de connaissances sur la parole continue conduit à construire des systèmes analytiques basés sur leur exploitation explicite des connaissances, tandis que la reconnaissance d'un faible nombre de mots isolés ou connectés repose sur des méthodes de comparaison de type *« pattern matching »*.

Les systèmes analytiques type expert/blackboard

Un système analytique repose sur l'intégration explicite des connaissances acoustiques, phonétiques, phonologiques, lexicales et syntaxiques. Pour ce faire, des expert-phonéticiens sont mis à contribution. L'un des plus célèbres est Victor Zue[4] : il transcrit le signal de parole en une suite de phonèmes et de mots, à la seule lecture[5] des spectrogrammes[6] de parole continue.

L'architecture des systèmes est basée sur une approche de type « système-expert ». Les connaissances sont décrites sous forme de règle de décision pour chaque niveau d'abstraction (acoustique, phonétique, lexical, syntaxique). La stratégie de décodage est de type hiérarchique comme dans le système français KEAL développé par le CNET[7] (De Mori, 1984) ou elle emprunte des méthodes de l'IA de type multi-agent ou « tableau noir » comme dans un des systèmes américains le plus performant de l'époque, Hearsay II (Erman, 1980), développé dans le cadre du programme DARPA[8].

La reconnaissance globale

Dans le cadre d'interfaces de type de « commande par mots clés » où le nombre de mots est restreint (moins d'une centaine) et les phrases sont de courts enchaînements, la méthode de reconnaissance est globale : une phase d'apprentissage, très simplifiée, consiste à enregistrer au moins une prononciation de chaque mot à reconnaître et à transformer chacune en une suite de vecteurs ou codes spectraux, considérée comme une référence. En phase de reconnaissance, la prononciation inconnue est transformée en une suite de vecteurs de même nature pour être comparée par distance à chaque référence. Le plus proche est le mot reconnu. La vitesse d'élocution n'étant ni identique d'une prononciation à l'autre, ni homogène entre son début et sa fin, une distance « élastique » de type alignement temporel est utilisée et l'algorithme de programmation dynamique associé (Sakoe, 1978) donne naissance à des systèmes très performants.

---

4 Victor Zue (né en 1944) est directeur du MIT Computer Science and Artificial Intelligence Laboratory.

5 1979 : https://www.youtube.com/watch?v=cgUuUoqwGmA : lecture de spectrogrammes « *Speech as eyes see it* » par Victor Zue, au CMU.

6 Spectrogramme : représentation 2D d'une suite d'analyses spectrales réalisées toutes les 5ms sur des fenêtres de 10ms.

7 Centre National d'Études en Télécommunications – Lannion.

8 Programme d'une durée de cinq ans pour le développement des systèmes de compréhension de la parole, sponsorisé par le DARPA (Defense Advanced Research Projects Agency).

En France, le système Séraphine du CNET reconnaît 12 mots en mode indépendant du locuteur avec un taux de reconnaissance de 96% en conditions laboratoire et de 84% en conditions téléphoniques (Gagnoulet, 1982) ; le système Mozart du LIMSI[9] reconnaît 10 chiffres avec 84% de taux de reconnaissance en mode multi locuteurs (Gauvain, 1982). La société *Threshold Technologies* fut la première à commercialiser un système de reconnaissance de 32 mots, le *VIP100,* en 1972.

Les systèmes performants du moment sont contraints à la reconnaissance de mots isolés ou de courtes suites de mots, en mode monolocuteur ou multilocuteurs : les utilisateurs du système sont connus du système en contribuant à son apprentissage. Les conditions d'enregistrement doivent être identiques en phase d'apprentissage comme en phase d'utilisation, le plus souvent dans des conditions dites de laboratoire.

Les années 1980-2010 : l'intégration de la variabilité par approche probabiliste

La reconnaissance de la parole continue se heurte au fait que le signal ne présente aucune frontière : du fait même de la production de la parole liée à des mouvements articulatoires (cordes vocales, mâchoire, langue…), le signal est un *continuum* sonore (à l'exception de la prononciation des occlusives) ; les frontières entre phonèmes sont indéfinissables et le signal est non reproductible, donc aléatoire. La prise en compte de cette propriété avec des modèles de décision « probabilistes » et plus précisément avec les modèles de Markov cachés (Jelinek, 1976) permet un essor considérable de la RAP, dès les années 1980. En reprenant l'architecture d'un système de RAP (voir figure RAP), le problème revient à trouver la phrase la plus probable parmi toutes les phrases possibles, sachant que les observations réelles se traduisent par une suite de vecteurs de nature spectrale 0 :

$$P(\widehat{W}/\,O)\; = max_i\, P(W^i/\,0)$$

avec :

$W^i = \left(w_1^i, w_2^i, \dots, w_{n_i}^i\right)$ une phrase composée de n_i mots

$O = (o_1, o_2, \dots, o_t, \dots, o_T)$ une suite de T vecteurs de nature spectrale, chaque vecteur est calculé sur une fenêtre d'environ 20ms et T varie d'une prononciation à l'autre.

Leur principe de résolution repose de manière très simpliste sur la règle de Bayes :

$$P(\widehat{W}/\,O)\; = max_i \frac{P(O/\,W^i)\,P(W^i)}{P(O)} = max_i\, P(O/\,W^i)\,P(W^i)$$

et la définition de deux modèles probabilistes : le modèle dit acoustique évaluant $P(O/W^i)$, la probabilité d'observer une suite de vecteurs *O* sous l'hypothèse qu'il s'agit de la prononciation de la phrase W^i, et le modèle de langage évaluant $P(W^i)$, la probabilité de vouloir prononcer la phrase W^i *a priori*.

9 Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur.

La modélisation acoustique : calcul de P(O/W^i)

Chaque mot du langage est décrit comme une suite d'unités phonétiques, en général, le phonème. Chaque unité est associée à un modèle de Markov caché élémentaire (*Hidden Markov Model* - HMM), à savoir un double processus stochastique : un processus interne markovien non observable $(X_t)_{t=1,\dots,T}$ qui prend ses valeurs dans l'ensemble des états cachés de l'unité phonétique $(e_j)_{j=1,\dots,J}$ et un processus externe observable $(Y_t)_{t=1,\dots,T}$ dont les réalisations sont des vecteurs de même nature que « o ». Chaque modèle élémentaire est caractérisé par les probabilités de transition entre les état $(P(e_i/e_j) = a_{j,i})$ et les lois dites « d'émission » associées à chaque état e_i pour évaluer $(P(o_t/e_i))$, pour tout (t,i)) (voir ci-dessous figure 2 HMM). Ces lois sont des lois paramétriques, de type gaussienne ou mélange de lois gaussiennes (*Gaussian Mixture Model* - GMM). Le modèle probabiliste d'un mot est également un HMM obtenu en concaténant les unités de base le constituant et en intégrant les variantes de prononciation.

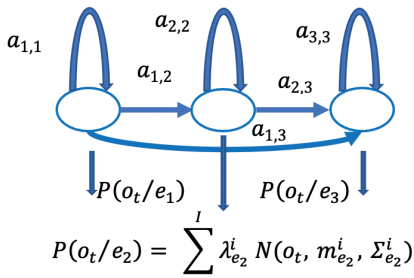

Fig. 2 HMM : modèle de Markov caché (trois états, un mélange de lois gaussiennes /état).

Le modèle de langage : calcul de P(W^i)

Le modèle de langage est défini à partir d'un lexique et d'une fonction de distribution probabiliste, qui sont appris à partir de grands corpus de textes écrits, que l'on notera *Text :*

- à l'issu d'un examen exhaustif de *Text*, le lexique rassemble les mots les plus fréquents. Seuls ces mots pourront être utilisés dans les phrases potentiellement acceptées par le système.
- le modèle probabiliste le plus utilisé jusqu'à nos jours est le modèle dit N-gram (N≥2) qui consiste à évaluer la probabilité d'apparition d'un mot à partir des « N-1 » mots précédents selon les équations suivantes :

$$P(W^i) = P\left(w_1^i, w_2^i, \dots, w_{n_i}^i\right)$$

$$= P(w_{n_i}^i \hat{0}\ 2), w_2^i, \dots, w_{n_i-1}^i) P(w_{n_i-1}^i / w_1^i, w_2^i, \dots, w_{n_i-2}^i) \dots P(w_2^i / w_1^i) P(w_1^i)$$

$$= P(w_{n_i}^i / w_{n_i-\mathrm{N}+1}^i, \dots, w_{n_i-1}^i) P(w_{n_i-1}^i / w_{n_i-\mathrm{N}}^i, \dots, w_{n_i-2}^i) \dots P(w_2^i / w_1^i) P(w_1^i)$$

Les fréquences d'occurrence des N-uplets de mots (w_1, w_2, ..., w_{N-1}, w) dans *Text* donnent une estimation des probabilités $P(w/w_1, w_2, ..., w_{N-1})$.

Le développement de l'approche probabiliste basée sur les HMM est lié au formalisme mathématique rigoureux associé à l'exploitation de deux algorithmes fondamentaux :

- l'algorithme de Baum-Welch (Baum, 1972) rend possible l'apprentissage automatique des paramètres du modèle,

- l'algorithme de Viterbi permet le décodage en temps réel (ou acceptable dans les années 90) de toute prononciation de phrases.

Tous les détails sont explicités dans un tutoriel très complet sur les HMM proposé par Lawrence Rabiner (Rabiner, 1989).

La reconnaissance de parole continue : de 1000 à plus 200 000 mots

Les performances de tels systèmes ne cessent d'augmenter depuis les années 80 jusqu'aux années 2010, conséquence de l'augmentation des corpus d'apprentissage, qu'il s'agisse du corpus « parole » recueilli pour apprendre les modèles acoustiques ou du corpus des données textuelles pour estimer les modèles de langage, et conséquence de la puissance de calcul des ordinateurs. Cette augmentation de puissance s'accompagne d'une complexification des HMM : le nombre de mélanges de lois gaussiennes (Juang, 1985) ne cesse de croître et de nombreuses techniques d'adaptation sont développées pour prendre en compte la diversité des locuteurs (Anastasakos, 1997) et/ou celle des environnements (Gales, 1998). Le modèle de langage 4-gram se substitue au modèle classique bi-gram. C'est l'âge d'or de l'approche dite GMM-HMM (*Gaussian Mixture model - Hidden Markov Model*).

Un des meilleurs systèmes de RAP en langue française, basé sur cette approche, est développé par le LIMSI (Gauvain, 2005). L'observation acoustique est une suite de vecteurs de nature spectrale de dimension 39, extraits sur une fenêtre d'analyse de 30 ms, toutes les 10ms. L'application est la reconnaissance automatique d'émissions de radio d'informations (France Inter, France Info, RFI…) :

- au niveau acoustique, 23 000 HMM composés de trois états chacun, modélisent les variantes de 35 phonèmes selon leur contexte. Les 12 000 mélanges de lois gaussiennes sont appris à partir d'un corpus audio de 190h, incluant 276 000 prononciations de mots,
- le modèle de langage est composé d'un lexique de 200 000 mots et d'un modèle 4-gram, appris à partir d'un corpus de 500 millions de mots.

En phase de reconnaissance, le taux d'erreurs mots WER est de 10,7% sur 10h d'émissions. Les résultats sont cependant très hétérogènes et varient en fonction du type d'enregistrement (plateau, terrain…). Un système comparable est développé pour la langue anglaise pour atteindre 10% de WER.

La place des connaissances

Les connaissances interviennent dans la construction des modèles probabilistes : elles sont souvent liées aux spécificités de la langue. Le nombre de modèles élémentaires phonétiques est dépendant de la langue ; les variantes de prononciation d'un même mot, des liaisons obligatoires ou facultatives dans l'élocution doivent être introduites au niveau de la modélisation acoustique (la langue française est une très bonne illustration des difficultés langagières engendrées). La définition du lexique est plus ou moins complexe et peut nécessiter des ressources conséquentes pour certaines langues (un lexique allemand demande plus de ressources qu'un lexique français compte tenu de l'existence des déclinaisons, chaque déclinaison engendrant un mot différent). Il faut aussi pallier le manque dû à l'extraction automatique à partir de textes, en particulier la sous-représentation des noms propres ; le nombre de mots dits « hors vocabulaire » est source d'erreurs. Néanmoins, ces contraintes diminuent avec l'augmentation des ressources mises à disposition en phase d'apprentissage, du fait de la grande automatisation.

À partir des années 2010 : de l'approche probabiliste à l'approche neuronale

L'utilisation du neurone formel (voir ci-dessous figure 3 MLP) en reconnaissance automatique des formes, date des années 50 et le perceptron multi couches (*Multi-Layer Perceptron* - MLP) se développe dès les années 80 avec l'apprentissage automatique de ses paramètres par l'algorithme de rétropropagation du gradient (Rumelhart, 1986).

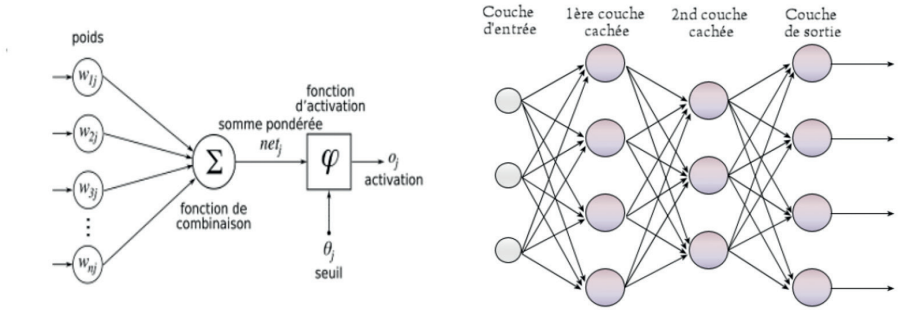

Fig. 3 MLP : neurone formel (vecteur d'entrée de dimension nj). Perceptron composé de deux couches cachées de neurones formels, d'un vecteur d'entrée de dimension 3 et d'une couche de sortie de dimension 4.

Le nombre de couches cachées du MLP, souvent très limité du fait du manque de puissance de calcul, rend la modélisation de représentations complexes impossible. De plus le MLP gère difficilement l'extrême variabilité de la dimension temporelle de la parole.

Du perceptron aux systèmes hybrides DNN-HMM et TDNN-DNN-HMM

Pour tenter de prendre en compte la dimension temporelle de la parole et les effets de coarticulation entre phonèmes, apparaissent les réseaux de neurones à délais temporels (*Time-Delay Neural Networks* - TDNN). La reconnaissance de phonèmes en mode dépendant du locuteur s'avère plus performante que l'approche HMM (Waibel, 1989). L'idée de remplacer les mélanges de lois gaussiennes par les sorties des réseaux de neurones artificiels apparait en 1990 (Morgan, 1990), néanmoins il faut attendre les années 2010 pour que les systèmes hybrides associant modèles de Markov cachés et réseaux de neurones profonds (*Deep Neural Network* - DNN) atteignent des performances comparables aux systèmes de type GMM-HMM. Le terme profond signifie que le réseau de neurones est constitué d'un nombre important de couches cachées, souvent de l'ordre de 6 à 10 ; l'apprentissage des milliers de paramètres (pondérations entre chaque couple de neurones) est rendu possible grâce à la montée en puissance de calcul et en stockage des ordinateurs.

La modélisation DNN permet des modélisations plus complexes et plus discriminantes que la modélisation GMM qui est basée sur des probabilités *a posteriori*, modélisations rendues possibles grâce à un apprentissage désormais acceptable, mais néanmoins toujours très complexe. Le cas particulier de l'approche TDNN-HMM est largement exploré et fournit de très bons résultats (voir ci-dessous figure 4 TDNN). Elle va également s'enrichir avec l'arrivée de nouvelles cellules issues des études dans le cadre de l'approche *End-to-End*.

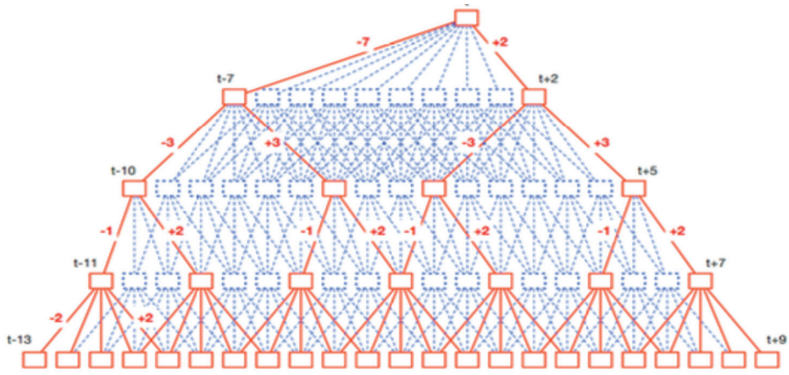

Fig. T4 DNN : architecture d'un TDNN composée de 3 couches cachées ; chaque entrée est un vecteur acoustique ; la sortie à l'instant t dépend des 23 observations acoustiques observées sur l'intervalle [t-13, t+9].

L'approche End-to-End (E2E)

L'approche E2E introduit un changement complet de paradigme en voulant relier directement l'entrée à la sortie au travers d'un seul modèle et en unifiant le processus d'apprentissage. Cette évolution se fait en trois étapes révélatrices d'architectures toujours de plus en plus sophistiquées, accompagnées du développement de différents types de Réseaux de Neurones Récurrents (*Recurrent Neural Networks* - RNN) comme les *Long Short* (LSTM) capables de traiter la variabilité temporelle de la parole (Graves, 2013) :

- l'approche CTC (*Connectionist Temporal Classification*) (2006),
- l'approche RNN-*Transducer* (*Recurrent Neural Network*) (2012),
- l'approche encodeur-décodeur avec attention (2015).

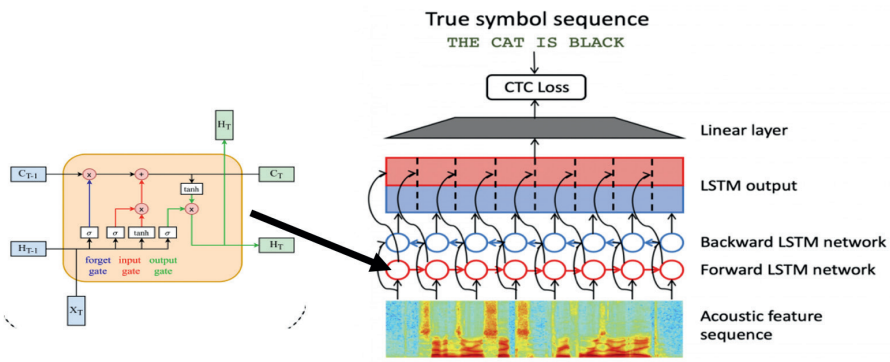

Fig. 5 RNN-LSTM : une cellule LSTM. Système de reconnaissance du signal de parole en une suite de caractères , incluant plusieurs niveaux de cellules LSTM.

Ces nouveaux réseaux de neurones profitent aussi à l'approche DNN-HMM. Citons à titre d'exemple le système DNN-HMM développé à l'université d'Aix-la-Chapelle[10]

10 RWTH Aachen University, Germany

(Luscher, 2019). Le HMM est construit à partir de modèles phonétiques de trois états chacun ; le DNN est composé de six couches bidirectionnelles, constituées de 1000 unités de type LSTM (voir ci-dessus figure 5 RNN-LSTM) dans chaque direction, et d'une couche de sortie de dimension 1000, pour caractériser les variantes acoustiques. Ce modèle est appris à partir de la base de données *LibriSpeech*[11], l'une des plus utilisées actuellement pour développer et évaluer les systèmes de RAP. Elle rassemble plus de 1000h d'enregistrements d'*audiobooks*, avec plus de 2000 locuteurs, et plus de 803 millions de phrases incluant 977 000 mots différents pour l'apprentissage des modèles de langage. Couplé à un modèle de langage 4-gram, le système atteint un WER de 2,6% sur les 5h de parole de *Librispeech* réservées à l'évaluation en condition propre.

Il est impossible de décrire en si peu de pages le foisonnement explosif des différents systèmes développés depuis 2016. Une course au meilleur taux de reconnaissance s'engage et les grands acteurs que sont Microsoft, Google et IBM développent des systèmes E2E de plus en plus complexes, demandant toujours plus de données. Il faut ajouter que ces dernières années, l'approche neuronale est également empruntée :

- pour trouver une meilleure paramétrisation du signal de parole et donc se substituer aux analyses spectrales classiques (Baevski, 2020),
- pour se substituer à l'approche probabiliste de type n-gram, dans le développement des modèles de langage, avec des performances stupéfiantes en génération de texte, en dialogue homme-machine et en traduction automatique. En témoignent les dernières réalisations comme ChatGPT (*Chat Generative Pre-trained Transformer*, développé par la société OpenAI[12]).

Les WER obtenus avec la base de données *Librispeech* par les grands acteurs commerciaux sont inférieurs à 2% en condition laboratoire et à 3% toutes conditions confondues[13].

La place des connaissances dans l'approche Deep Learning.

La conception des systèmes de RAP basés sur les techniques de *Deep Learning* ne nécessite l'introduction d'aucune connaissance *a priori* ; certains ingénieurs spécialistes n'ont d'ailleurs aucune formation en parole. Cependant, comme dans toutes les applications empruntant actuellement ces approches, il est frustrant de ne pas pouvoir donner la moindre interprétabilité ou explicabilité des résultats obtenus, qu'ils soient bons ou mauvais. Une question se pose naturellement : qu'est-ce que le système apprend ? Plusieurs études tentent d'y répondre. Des réponses rassurantes sont proposées : elles montrent que des neurones se spécialisent pour réagir à certains types de sons. Dans le cadre d'une étude de cas (Nagamine, 2015), à savoir la reconnaissance automatique des 40 phonèmes anglais, un réseau de neurones profond avec six couches cachées de 256 neurones chacune est appris. En phase d'évaluation, la sélectivité de chaque neurone face à chaque phonème est mesurée et il est démontré une forte spécialisation de chacun des neurones en correspondance avec la catégorisation linguistique *a priori* des phonèmes. Les neurones semblent découvrir ces connaissances…

11 https://paperswithcode.com/sota/speech-recognition-on-librispeech-test-clean, https://paperswithcode.com/sota/speech-recognition-on-librispeech-test-other.

12 https://openai.com/blog/chatgpt.

13 https://paperswithcode.com/task/speech-recognition.

« WER we are and WER we think we are » emprunté à Szymanski 2020

Néanmoins, les systèmes commerciaux montrent des limites et les performances des meilleurs systèmes peuvent chuter à 20% (Szymanski 2020). Parmi les multiples raisons, il est clair que la production de parole est différente dès lors que l'auditeur est une machine ou un homme et que la spécificité du domaine (banque, assurance…) et celle du milieu environnant ne sont pas correctement prises en compte.

La reconnaissance robuste de conversations en parole spontanée, la reconnaissance de parole d'enfants et de non natifs, et le réel dialogue restent des défis tout comme la compréhension automatique. Sous cet angle, l'apprentissage de toutes les connaissances nécessaires pour atteindre ces objectifs paraît encore utopique et la parole a quelques chances de rester un sujet de recherche multidisciplinaire quelques années encore.

Remerciements

Un grand merci à Abdelwahab Heba (Heba, 2021), chercheur chez Microsoft Speech Research, et à Lucile Gelin (Gelin, 2022), chercheuse associée IRIT-Lalilo, qui m'ont permis au travers des discussions autour de leur manuscrit de thèse de mettre mes connaissances en *Deep Learning* à niveau !

Bibliographie

Anastasakos T., Donough J.Mc, Schwartz R., Makhoul J., « A compact model for speaker-adaptive training », *In Proceeding of Fourth International Conference on Spoken Language Processing, ICSLP'96, volume 2,* IEEE, 1996, p. 1137-1140.

Baevski A., Zhou Y., Mohamed A., Auli M., « wav2vec 2.0 : A framework for self-supervised learning of speech representations », *Advances in Neural Information Processing Systems*, 33, 2020.

Baum L.E., « An inequality and associated maximization technique in statistical estimation for probabilistic functions of a Markov process», *Inequalities III*, 1972, p. 1-8.

Dreyfus-Graf J., « Phonétographe et Phonétique », *Folia Phoniatr Logop volume 5, Issue 4*, 1953, p. 223-232.

Erman Lee D., Hayes-Roth Frederik, Lesser Victor R., Reddy RRaj D., « The Hearsay-II Speech-Understanding System: Integrating Knowledge to Resolve Uncertainty », *ACM Computing Surveys, volume 12, n° 2*, ACM, June 1980, p. 213-253.

De Mori R., Gilloux M., Mercier G., Simon M.A., Tarridec C., Vaissière J., Gillet D., Gérard M., « Integration of acoustic, phonetic, prosodic and lexical knowledge in an expert system for speech understanding », *Proc of ICASSP*'84, San Diego, Californie, IEEE, March 1984.

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

Gauvain J.L., Mariani J., « Mozart : un système de reconnaissance globale de parole continue », *Proc of the 11e ICA*. Paris, 1982.

Gauvain J.L., Adda G., Adda-Dekker M., Allauzen A., Gendner V., Lamel L., Schwenk H., «Where Are We in Transcribing French Broadcast News? », *Proc. of INTERSPEECH 2005*, Lisbon (Portugal), ISCA, September 2005, p. 1665-1668.

Gelin L., « Reconnaissance automatique de la parole d'enfants apprenant.e.s lecteur. ice.s en salle de classe : modélisation acoustique de phonèmes », thèse de doctorat de l'Université de Toulouse, février 2022.

Heba A., « Reconnaissance automatique de la parole à large vocabulaire : des approches hybrides aux approches End-to-End » , thèse de doctorat de l'Université de Toulouse, mars 2021.

Jelinek F., « Continuous speech recognition by statistical models », *Proc of IEEE, vol. 64, Issue: 4,* IEEE, April 1976, p.532-566.

Luscher C., Beck E., Irie K., Kitza M., Michel W., Zeyer A., Scluter R., Ney H., « RWTH ASR Systems for LibriSpeech : Hybrid vs Attention », *Proc. of INTERSPEECH 2019*, Graz (Austria), ISCA, September 2019, p. 231-234.

Morgan N., Bourlard H., « Continuous speech recognition using multilayer perceptrons with hidden Markov models », *Proc. IEEE Intl. Conf. on Acoustics Speech & Signal Processing, vol.1, 1990*, Albuquerque (USA), IEEE, April 1990, p. 413-416.

Nagamine T., Seltzer M.L., Mesgarani N., « Exploring How Deep Neural Networks Form Phonemic Categories », *Proc. of INTERSPEECH 2015*, Dresden (Germany), ISCA, September 2015, p. 1912-1915.

Rabiner L.R., « A tutorial on hidden Markov models and selected applications in speech recognition », *Proc of IEEE, vol. 77, Issue : 2,* IEEE, February 1989, p. 286-257.

Rumelhart D. E., Hinton G. E., Williams R. J., « Learning representations by back-propagating errors », *Nature 323 (6088)*, October 1986, p. 533–536.

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

Szymanski P., Zelasko P., Morzy M., Szymczak A., Zyła-Hoppe M., Banaszczak J., Augustyniak L., Mizgajski J., Carmiel Y., « WER we are and WER we think we are », in *Finding of the Association for Computational Linguistics, EMNLP 2020*, https://aclanthology.org/2020.findings-emnlp.295, doi : 10.18653/v1/2020.findings-emnlp.295.

Waibel A., Hanazawa T., Hinton G., Shikano K., Lang K., « Phoneme Recognition: Neural Networks vs Hidden Markov Models », *Proc. IEEE Intl. Conf. on Acoustics Speech & Signal Processing*, vol. 1, pp. 107-110, 1988.