Les assistants personnels vocaux

Veille technologique

> L'évolution des technologies vocales

Le principe de reconnaissance vocale trouve ses origines au milieu du XXème siècle :

À l'époque, 3 ingénieurs de chez ***Bell Labs*** achève à partir de différent relais un dispositif permettant de traduire des nombres saisi par la parole sur une interface machine. 10 ans plus tard, les recherches ont pris suffisamment d'ampleur pour apporter une pleine maîtrise de la reconnaissance vocale numérique, et on commence tout doucement à implémenter les premiers phonèmes.

1971 voit la naissance du projet ***ARPA***, ce projet réunira pendant plus de 5 ans, ingénieurs, chercheurs et investisseurs dans le but de faire évoluer la technologie lié à la reconnaissance vocale. Nombre de professionnels ayant contribué à l'évolution du traitement de la parole par informatique n'imagine pas à l'époque que cette technologie pourrait un jour permettre le contrôle d'une machine. Pourtant, les premières machines de traitement vocale autonome ne tardent pas à être commercialisé dans cette décennie.

S'en suivra la production de systèmes de reconnaissance vocal contenant alors quelques milliers de mots, ces systèmes bien que indépendant à l'époque évolueront vite, parallèlement aux avancées de l'informatique, en module complétant d'autres outils informatique.

Pour exemple la société Texas Instrument qui produit en 1978 un jouet nommé "Speak & Spell" ou "Dictée Magique" en français; cette machine, d'apparence anodine était capable de combiner à la fois reconnaissance et synthèse vocale afin de familiariser les enfants à l'apprentissage de l'anglais.

Dans le début des années 80, et pour la première fois, le procédé de reconnaissance vocale est utilisé dans l'aviation militaire, cette innovation sera marquante de par son utilisation concrète dans un domaine d'importance majeure. /

Le procédé de reconnaissance vocale tel que nous le connaissons aujourd'hui est conceptualisé à partir de 1985 avec la naissance des premières sociétés dédiés à la conception de solution de reconnaissance vocale, des géants du secteur tel que Dragon System ou encore Lernout & Hauspie.

Petit historique :

1995 - ViaVoice (IBM)

1997- Dragon (Nuance Communications) actuellement Dragon NaturallySpeaking / Dragon Dictate (Anciennement MacSpeech)

1999 - Shazam (Reconnaissance vocale) Entertainment Limited a été créé en 1999 par Chris Barton, Philip Inghelbrecht, Avery Wang et Dhiraj Mukherjee.

2000 - Audacity

2003 - Acapela (Synthèse Vocale)

2008 - Silvia de Cognitive Code inc., sortie le 1 Janvier 2008

2011 - SIRI par SRI International et Apple Inc., sortie le 4 octobre 2011

2012 - S Voice de Samsung, sortie le 30 Mai 2012
- Voice Mate de LG, sortie le 20 juin 2012
- Google Now de Google, sortie le 9 Juillet 2012
- Maluuba de Maluuba inc (Android et Windows phone), sortie le 11 Septembre 2012
- iTranslate

2013 - Voxygen (Synthèse Vocale)

2014 - Cortana de Microsoft, sortie le 2 avril 2014
- Echo de Amazon, sortie le 6 Novembre 2014 (commande vocale / Wi-fi, Bluetooth)

De nos jours - Hound (Bêta anglaise)

> La reconnaissance vocale

La reconnaissance vocal est un processus permettent de convertir des phrases dictées sous forme écrite afin de pouvoir les interprétés. Pour ce faire l'information passe par plusieurs étapes et sous plusieurs états différents.

Initialement, l'information est émise dans notre cerveau qui l'extériorise par l'intermédiaire de la parole, cette dernière passe alors sous forme acoustique pour pouvoir être interprété par une autre personne. Les machines ont elles beaucoup de mal à reconnaître les mots et les phrases dans leurs ensembles ainsi que les pauses dans la parole et les silences. Dans le cas de la reconnaissance vocale, l'information acoustique est directement enregistrée par un microphone qui transforme alors cette information en signal électrique qui est directement retraduit en signal binaire par le module avec lequel on s'enregistre. La machine à donc en sa possession des donnés qui retranscrivent des donnés sonore, encore impossible à traduire. Pour trouver la signification de ces données un logiciel de reconnaissance vocal vas les découpés en plusieurs parties identifiable appelés phonèmes. Ce processus peut se faire de deux manières différentes, soit il est établis localement à partir d'une base de donnée personnelle établis au paravent par l'utilisateur. Dans ce cas précis seul les personnes ayant déjà enregistré leurs phonèmes peuvent être reconnus. Le deuxième processus ce fait extérieurement, après que les données sonores soit envoyer sur un serveur ayant pour but de les interprétés. A partir d'une très grande base de donné et d'un logiciel très complexe les

différentes phonèmes sont identifiées et assemblés pour déterminer les mots qu'elles composent permettant la conversion des mots sous forme écrite.

Pour interpréter les signaux, il existe plusieurs techniques d'extractions des phonèmes :

- La ZCR (Zero Crossing Rate) sert à informer où et quand un signal dépasse zéro, c'est une mesure de fréquence qui varie plus ou moins rapidement si la fréquence est forte ou faible. Les signaux vocaux étant compliqués à analyser de par le fait qu'ils varient en quelques millisecondes, cette analyse est basée sur des fragments de signal de 10 à 30 millisecondes. Chaque fragment démarrant à la moitié du temps de celui qui le précède (et dure lui aussi 10 à 30 millisecondes). Ce qui entraîne une superposition des fragments (la 2ème moitié d'un signale et la 1ère moitié du signal suivant sont donc identique), cette étape est appelé le framing et est utilisé dans cette situation pour améliorer la précision de l'analyse du signale. Après que le signal ait été "framé" et la ZCR modélisé, il faut différentier quel signal est considérer par de la parole et quel signal est considérer comme du bruit. On utilisera donc pour cela une limite considérant toutes les valeurs inférieures comme étant du bruit et les valeurs supérieurs comme étant de la parole.

- L'étude de l'énergie à court termes : Il est connus que chaque transfert d'information implique un transfert d'énergie, l'information vocale se transmet par énergie mécanique mais une fois interpréter par un micro on peut observer les variations d'énergie sous forme de signal électrique facile à interpréter. Les intérêts de cette analyse sont premièrement, comme pour la ZCR la détection de la parole et du bruit car la parole dégage plus d'énergie que le bruit dans la mesure où l'utilisateur est proche de son micro. Le deuxième intérêt réside dans la détection des voyelles car elles dégagent plus d'énergie.

- Il existe également d'autres techniques qui peuvent permettre l'isolation de phonèmes et la dissociation du bruit. Ces techniques étant particulièrement complexes nous allons donc les citer en passant les détails techniques. Il existe donc les techniques de l'encadrement, du pitch extractor, du Group delay ainsi que de l'analyse spectrographique.

> Les moteurs de recherche

- Le tampon acoustique, l'exemple parfait avec Shazam -

Pour reconnaître les différents sons, des Algorithmes de recherche et de détection ont été développé. L'un d'eux a été développé par l'agence à la célèbre application Shazam. Son principe est simple, détecter une musique à partir d'un court passage pour aider l'utilisateur à trouver le nom de sa musique, l'artiste ainsi que l'album. Pour se faire, Shazam a eu l'idée de créer un algorithme basé sur le tampon acoustique.

- Quel est son principe ? -

Il est tout simplement révolutionnaire ! En quelques secondes Shazam détecte une musique, avec un pourcentage d'erreur maximal de 0.0025%. Pour commencer, il faut savoir que Shazam code les données relatives à une empreinte en 32 bits. Cela signifie que l'application numérise la valeur du maximum parmi 2^32 valeurs possibles, soit environ 4 milliards de valeurs.
Ensuite, Shazam fait 10 empreintes acoustiques par seconde, 10 empreintes par secondes pour chacun de ses morceaux référencés.

La base de données de Shazam contient environ de 18 milliards d'empreintes. En moyenne une valeur d'empreinte est donc contenue dans 5 morceaux (18 milliards sur 4). Ainsi en recherchant dans la base de données un morceau qui a deux empreintes identiques avec le morceau inconnu on peut affirmer que les deux musiques correspondent avec une chance de 1 sur 400 000 de se tromper.
De plus, pour assurer sa fiabilité, Shazam fait une vérification sur 100 fichiers qui peuvent paraître similaires, en fonction aussi du bruit ambiant.

Car oui, Shazam peut détecter votre musique n'importe où, même si il y a une autre musique à coté de vous, ou un bruit ambiant assez fort (comme une conversation au ton normal, si le ton est élevé il y aura trop de perturbations sonores). Enfin, le dernier point fort de Shazam : s'il trouve 2 empreintes correspondantes à un même morceau, il considère que c'est celui-là !

En effet, il n'y a qu'environ un risque sur 25 000 que l'un des 300 morceaux suspecté à cause d'une empreinte bruitée corresponde à l'un des 3 morceaux identifié par une empreinte correcte.

Un autre exemple de moteur de recherche dans les bases de données est le célèbre robot Siri, créé et utilisé par Apple pour simplifier la vie des utilisateurs. Voici son fonctionnement :

- Les données binaires (autres qu'audio) sont compressées avec Zlib et sont soit des ping, soit des pong, soit des fichiers plist.
- Les données audio sont compressées avec Speex (un codec conçu pour la VoIP)
- Chaque échange est identifié grâce à un marqueur unique contenu dans chaque iPhone 4S qu'Apple vérifie constamment lors des échanges entre votre téléphone et Siri. On pourrait toutefois copier ce code sur un autre appareil afin de le faire passer pour un iPhone4S.
- Que Siri et le serveur d'Apple échangent beaucoup de données, y compris des scores de confiance pour chaque mot.

Siri traite et analyse l'ensemble de ces informations en utilisant un amalgame d'applications, comme la reconnaissance vocale, une restitution de sortie en voix naturelles, un algorithme de recherche de la langue couplé à un moteur de réponse sophistiqué. Fort de tout cela, il construit une compréhension contextuelle de vous et stocke l'information pour plus tard…

Il est indiqué que Siri recueille deux types de données: les données utilisateur et les données d'entrée vocale. En effet, il stocke les données utilisateur, les questions les plus demandées, les informations que les utilisateurs veulent le plus afin d'améliorer les recherches, les stockages, les informations envoyées.

De plus, à chaque demande, la base de données enregistre toutes les informations fournies par l'utilisateur et le téléphone : la voix, les mots compris, l'indice de confiance, de bonne compréhension. Mais également des informations statistiques, de lieu, de nombre de demandes, de la longueur de la conversation et de la demande.

De plus, Siri est de plus en plus performant car la compréhension des demandes est de plus en plus efficace, car il y a de plus en plus d'informations stockées, des recherches plus précises, des réponses plus simples et plus rapides.

> La recherche de l'information

La base de données est une enveloppe technico-fonctionnelle qui va héberger des tables. C'est en général l'objet le plus haut dans la hiérarchie des "choses" gérées par le SGBD, bien que certains SGBD proposent aussi le tablespace. La base est destinée à accueillir les tables.

Pour les SGBD par contre, il en existe plusieurs sortes:

- fichiers
- hiérarchiques (IMS/DB d'IBM par exemple)
- relationnels (Oracle, DB2, Interbase, MySQL, etc...)
- orientés Objets.

Une base de données est un « conteneur » stockant des données telles que des chiffres, des dates ou des mots, pouvant être retraités par des moyens informatiques pour produire une information ; par exemple, des chiffres et des noms assemblés et triés pour former un annuaire. Les retraitements sont typiquement une combinaison d'opérations de recherches, de choix, de tri, de regroupement, et de concaténation.

C'est la pièce centrale d'un système d'information ou d'un système de base de données (ou base de données tout court), qui régit la collecte, le stockage, le retraitement et l'utilisation de données. Ce dispositif comporte souvent un logiciel moteur le SGBD (Système de Gestion de Bases de Données), des logiciels applicatifs, et un ensemble de règles relatives à l'accès et l'utilisation des informations.

Le SGBD est une suite de programmes qui manipule la structure de la base de données et dirige l'accès aux données qui y sont stockées. Une base de données est composée d'une collection de fichiers ; on y accède par le SGBD qui reçoit des demandes de manipulation du contenu et effectue les opérations nécessaires sur les fichiers. Il cache la complexité des opérations et offre une vue synthétique sur le contenu. Le SGBD permet à plusieurs usagers de manipuler simultanément le contenu, et peut offrir différentes vues sur un même ensemble de données.

Les disques durs, mémoire de masse de grande capacité, ont été inventés en 1956. L'invention du disque dur a permis d'utiliser les ordinateurs pour collecter, classer et stocker de grandes quantités d'informations.

Les premières bases de données hiérarchiques sont apparues au début des années 1960. Les informations étaient découpées en deux niveaux de hiérarchie : un niveau contenait les informations qui sont identiques sur plusieurs enregistrements de la base de données. Le découpage a ensuite été étendu pour prendre la forme d'un diagramme en arbre.

En 1965, Charles Bachman conçoit l'architecture Ansi/Sparc encore utilisée de nos jours. En 1969, il créa le modèle de données réseau au sein du consortium CODASYL pour des applications informatiques pour lesquelles le modèle hiérarchique ne convient pas. Charles Bachman a reçu le prix Turing en 1973 pour ses « contributions exceptionnelles à la technologie des bases de données. ».

En 1968 Dick Pick crée Pick, un système d'exploitation contenant un système de gestion de base de données « multivaluée » (SGBDR MV).

En 1970, Edgar F. Codd note dans sa thèse mathématique sur l'algèbre relationnelle qu'un ensemble d'entités est comparable à une famille définissant une relation en mathématique et que les jointures sont des produits cartésiens. Cette thèse est à l'origine des bases de données relationnelles. Edgar F. Codd a reçu le prix Turing en 1981.

Le modèle entité-association a été inventé par Peter Chen en 1975 ; il est destiné à clarifier l'organisation des données dans les bases de données relationnelles.

Dans le modèle relationnel, la relation désigne l'ensemble des informations d'une table, tandis que l'association, du modèle entité-association, désigne le lien logique qui existe entre deux tables contenant des informations connexes.

Les premières bases de données étaient calquées sur la présentation des cartes perforées : répartis en lignes et colonnes de largeur fixe. Une telle répartition permet difficilement de stocker des objets de programmation ; en particulier, elles ne permettent pas l'héritage entre les entités, caractéristique de la programmation orientée objet.

Apparues dans les années 1990, les bases de données objet-relationnel utilisent un modèle de données relationnel tout en permettant le stockage des objets. Dans ces bases de données les associations d'héritage des objets s'ajoutent aux associations entre les entités du modèle relationnel.

Les étapes clés du cycle de vie d'une base de données sont la conception et la mise en service.

Avant la conception, les utilisateurs et les producteurs des informations sont interviewés en vue de prendre connaissance des caractéristiques des informations, des relations entre les informations, ainsi que les caractéristiques du système informatique qui accueillera la base de données. L'objectif de cette étape est de recueillir les caractéristiques des informations dans la pratique, et les besoins des usagers, et de les formuler d'une manière simple, compréhensible autant par les usagers que les administrateurs de base de données.

Puis sera créé un schéma d'ensemble du réseau d'informations et de relations, sous forme de diagramme comportant des entités, des attributs et des relations. Ce plan est ensuite transformé en instructions formulées dans le langage de commande du SGBD et les instructions sont exécutées en vue de créer la structure de la base de données et la rendre opérationnelle.

La définition de l'organisation interne d'une base de données - son modèle de données physique - est l'étape finale de sa construction. Cette opération consiste tout d'abord à définir des enregistrements correspondants au modèle de données logique. Les enregistrements sont stockés dans des fichiers, et chaque fichier contient typiquement un lot d'enregistrements similaires. Lors de cette étape diverses techniques sont utilisées en vue d'obtenir un modèle qui aboutit à une vitesse adéquate de manipulation de données, tout en garantissant l'intégrité des données.

La qualité du modèle de données physique a un impact majeur sur la vitesse des opérations sur la base de données. Une simple amélioration peut rendre les opérations sur les données 50 fois plus rapides, différence d'autant plus sensible qu'il y a une grande quantité de

données. Au début des années 2000 il existe des bases de données contenant plusieurs téraoctets de données et des ingénieurs indépendants dont l'activité consiste uniquement à aider des clients à accélérer leurs bases de données.

Une fois opérationnelle, des opérations de surveillance permettent de déceler des problèmes susceptibles de nécessiter des modifications du schéma. Des modifications peuvent également être apportées en cas de changement des besoins des utilisateurs.

> La synthèse vocale

La synthèse vocale est l'opposé de la reconnaissance vocale. Elle fonctionne de façon similaire mais dans un but différent. La synthèse mettra du son sur une réponse de l'unité pour rendre une réponse parlée proche de la voix humaine.

- L'analyse textuelle approfondie -

Pour commencer, nous disposons d'une base de données phonétique. C'est effectivement grâce à la découpe phonétique du langage que notre analyse sera possible.

En effet, un texte normalisé, c'est à dire l'entrée dans la machine pour une traduction en langage machine, sera découpé phonétiquement. Suite à la phonétisation, une analyse syntaxique permettra une correspondance aux différents phonèmes. Enfin, le texte passera par la prosodie, pour déterminer l'inflexion, le ton, la tonalité, l'intonation, l'accent et la modulation d'un langage oral incluant aussi le rythme d'élocution. Cela peut dépendre des émotions et de l'impact souhaité.

- L'utilisation de la base de données acoustique -

La base de données acoustique est nécessaire pour reproduire une voix. C'est cette base de données qui contient donc les voix qui seront reproduites lors de la synthèse vocale.

Pour commencer, afin de pouvoir reproduire au mieux chaque langue mais surtout chaque sonorité de cette langue, un narrateur enregistre une série de texte comportant tous les sons possibles de ladite langue. Les enregistrements subissent ensuite une segmentation, ils sont étudié en fonction de la prosodie et découpé en segment de sonorité afin d'y correspondre au mieux. Ces données, appelées tampons acoustiques, sont triés et constituent la base de données vocale.

- La synthèse à partir des bases de données -

La synthèse est la mise en relation des bases de données, en effet, lors de la réception des données à synthétiser, les phonèmes récupérés à partir de l'analyse textuelle sont associés aux tampons acoustiques correspondant.

Enfin, le son est généré à partir des tampons acoustiques, du timbre et des durées de prononciation qui auront été calculé au préalable.

Il existe deux formes de génération de voix :

- La première détermine un modèle numérique de l'ensemble des productions vocales d'un locuteur qui sont alors utilisés pour générer le signal de parole. Ceci concerne les méthodes de synthèse articulatoire, de synthèse par formants ou encore les récentes techniques de synthèse HMM (Hidden Markov Model). Mais ces associations par formants n'ont jamais pour s'approcher de l'illusion d'une voix naturelle

- La seconde approche, apparue à la fin des années 90 et la plus utilisée actuellement, procède par juxtaposition de segments de parole préalablement enregistrés. C'est de la synthèse par concaténation et c'est la seule qui permette de produire une parole véritablement naturelle.

Voxygen nous apprend par exemple :

" Depuis le début des années 2000, la technologie « Voxygen Expressive Speech » est basée sur le principe de la Synthèse par Corpus (SPC). La SPC s'appuie sur un dictionnaire acoustique créé de telle façon qu'il porte l'univers de production d'un locuteur. Cette base de données est constituée de l'enregistrement d'un ensemble de phrases bien choisies, i.e. offrant de bonnes propriétés de couverture des phénomènes linguistiques voire prosodiques pertinents. Du fait de la richesse du dictionnaire acoustique, on arrive ainsi à sélectionner des séquences acoustiques adéquates, qui ne nécessitent potentiellement aucune déformation, offrant ainsi une qualité de timbre parfaitement naturelle."

> Pour finir

Finalement comment ça marche ?

Les assistants personnels vocaux sont finalement l'assemblage de toutes ces technologies. En effet, il faut connaître les principes de la reconnaissance vocale pour pouvoir comprendre ce que souhaite l'utilisateur. Il faut des bases de données dans lesquels faire les recherches appropriées et même une connexion à internet pour répondre à certaines questions et forcément le logiciel qui s'occupera de gérer tout cela. Ensuite il faut répondre à l'utilisateur et donc savoir synthétiser une voix après avoir modélisé le texte.

Comment on en est arrivé là ?

Pour en arriver là, ce genre d'assistant a évolué de programme ne requérant aucune partie vocale pour ensuite écouter et analyser des questions simple ou des commandes simples. Ensuite seulement les réponses vocales ont été incluses.

Pour conclure, les assistants personnels vocaux sont en pleine expansion et évolution quant à leurs compétences. Ils tendent à devenir des IA et commencent à apprendre de leurs expériences. Ils sont surtout là pour nous faciliter la vie, nous servir de pense-bête et nous renseigner sur diverses choses. Qui peut savoir jusqu'où l'on pourra aller avec ces outils devenus presque indispensables pour certaines personnes ? Pourrions-nous un jour espérer d'avoir des assistants personnels, qui arriveraient à anticiper nos réactions et à tenir une

conversation logique, censée et intéressante ? Ce scénario est déjà pensé par George Lucas, par exemple, en créant C3PO.