par Alfons Calders

Reconnaissance et synthèse vocales : une science complexe

Dans notre édition de décembre dernier, nous vous annoncions le projet de robot Anty pour les enfants hospitalisés. Avec l'aide de la région de Bruxelles-Capitale, le groupe de recherche Robotics and Multibody Mechanics (R&MM) de la VUB développera un premier prototype d'ici 2007. Mais ce projet présente d'autres facettes que la seule application « robot ». Cet ami à câliner doit être doué de « sens » jusqu'à un certain niveau. Anty a aussi besoin d'une voix pour communiquer avec les petits patients.

Pour ce volet « vocal », le groupe de recherche Anty s'est vu adjoindre les services du groupe de recherche DSSP (laboratoire de parole numérique et de traitement audio), un service du département ETRO (Electronique et Informatique). Le moment venu, Anty fera « appel » aux résultats des travaux de recherches fondamentales et appliquées de ce groupe. Des travaux de fin d'études sont par ailleurs en cours dans le but de conférer à Anty une voix personnelle et un (balbutiement de) langage (en synthèse vocale). Ces travaux de fin d'études ont aussi pour but de conférer à cette voix une intonation spécifique ciblée sur le moral de l'enfant hospitalisé. *Industrie Technique et Management* s'est entretenu au sujet des recherches vocales avec le **professeur Werner Verhelst**, chef du laboratoire de parole numérique et de traitement audio de la **VUB**.

PAROLE NUMÉRIQUE ET TRAITEMENT AUDIO

La parole numérique présente deux aspects : la compréhension (reconnaissance vocale) et la parole (synthèse vocale). En Belgique, on aurait plutôt tendance, dès qu'on aborde le sujet, à évoquer la fameuse débâcle d'Ypres, mais toutes les universités mènent des recherches fondamentales sur les deux aspects. Et il existe déjà des applications où cette technologie représente socialement bien plus qu'un simple gadget.

Songeons par exemple à un projet de recherche stratégique de base SBO, une coopération financée par l'IWT entre la VUB, la RUG, la KULeuven et l'UA, dont les travaux portent sur une combinaison de la reconnaissance vocale et de la synthèse vocale destinée à aider des personnes souffrant de dysphasie (difficultés d'élocution) et de dyslexie (avec les problèmes de lecture afférents) dans le cadre de leurs exercices. On constate en effet d'une part qu'il existe trop peu de logopèdes pour accueillir les enfants à problèmes et, d'autre part, que les parents n'ont souvent pas assez de temps (et de patience) de faire répéter les exercices par leurs enfants. Il convient aussi de mentionner « l'accoutumance » des thérapeutes qui compliquent les exercices aussitôt qu'ils voient une amélioration, avec comme corollaire que l'enfant finit par se décourager. Un ordinateur peut être utilisé par un enfant ou un adulte présentant un handicap de lecture (et « neutre », « sans contrainte ») ce qui offre l'avantage supplémentaire d'être un observateur objectif. Cela nécessite un ordinateur qui peut reconnaître ce que la personne en question lit et qui peut également la corriger. Et pas un seul exercice, mais toute une panoplie. Pour les exercices d'amélioration de la lecture, on travaille à l'heure actuelle avec des enregistrements existants. L'idéal est de travailler au départ de textes écrits qui sont reconnus et débités par l'ordinateur.

L'ami Anty (ici au Salon de l'Auto) doit pouvoir parler pour communiquer avec ses jeunes patients.

Dans le volet « reconnaissance », on en est à l'heure actuelle à la « reconnaissance de mots ». Cela signifie que l'ordinateur peut reconnaître des mots prononcés par une personne ou un nombre limité de personnes après un « processus d'apprentissage » (où ces personnes répètent un grand nombre de fois les mots en question). On travaille à la « reconnaissance de phrases » (par ex., la transcription de lettres dictées). Comme ces systèmes sont basés sur une grammaire préprogrammée, l'ordinateur éprouve de grosses difficultés si la personne bafouille ou se reprend au milieu d'une phrase. Mais à qui cela n'arrive-t-il pas ? Dans ce projet, cela va encore plus loin : l'ordinateur doit disposer d'une « intelligence artificielle » pour reconnaître des mots bafouillés ou mal lus pour ensuite pouvoir les corriger et les prononcer correctement.

Dans le volet de la VUB, la synthèse vocale, on cherche à créer une langue « naturelle » au départ de « textes écrits ». Cela peut sembler évident, mais on arrive aujourd'hui à générer des textes courts avec une voix (de synthèse) assez égale. Il s'agit, pour l'essentiel, de phrases courtes dans un domaine d'application restreint. Exemples : « le numéro que vous avez demandé... », « tournez à droite », etc. Même dans ces applications simples, le système n'est pas infaillible. Cette étude vise à franchir un pas de plus avec, d'une part, le volet « instructions parlées/systèmes d'aide » (comme il s'agit de personnes avec des problèmes de lecture, un manuel imprimé ne sera pas d'une grande utilité) et, d'autre part, de « donner la lecture du texte de

l'exercice » et « la correction des erreurs de lecture du texte ». Cela requiert un très haut niveau de qualité de lecture, car elle remplit une fonction d'exemple. La seule « simplification » est que l'on peut travailler avec un vocabulaire simple (simplifié) dans une langue sans phrases à la syntaxe complexe. Le projet Anty nécessitera lui aussi « une synthèse vocale de qualité supérieure » si l'on veut que le robot à câliner joue un « rôle d'informateur » ou pour lui « faire répondre spontanément des phrases sensées ».

EVOLUTION DANS LA RECHERCHE VOCALE

Il va encore falloir de nombreuses recherches fondamentales et stratégiques avant d'en arriver là. Il n'en reste pas moins vrai que les recherches sur la synthèse vocale ont déjà bien avancé. Et même si on pensait le contraire, les problèmes sous-jacents de la synthèse vocale sont similaires et tout aussi complexes que ceux liés à la reconnaissance vocale.

La reconnaissance vocale implique un problème qui n'apparaît pas dans la synthèse vocale, à savoir les « bruits de fond » qui rendent plus difficile à distinguer ce qui doit être reconnu. Les problèmes concernant l'intonation, les accents toniques, les liaisons dans une phrase parlée, représentent quant à eux des difficultés tant pour la reconnaissance vocale que pour la synthèse vocale. Pour la reconnaissance vocale, on travaille avec des phonèmes (des sons) qui sont associés de manière à former des mots intelligibles. Pour la synthèse vocale, il faut choisir les bons phonèmes pour prononcer correctement le mot en question dans une phrase spécifique. Les deux approches reposent sur un « modèle linguistique » permettant de déterminer les mots qui peuvent être placés en séquence, car l'ordre des mots dans un « dialogue » diffère de l'ordre des mots dans un « exposé » et est même encore différent que lorsque l'on « achète un billet de train ». Le plus difficile, dans la reconnaissance vocale, semble être la transcription d'une « interview » car la reconnaissance de la langue parlée comporte le problème que l'on se répète parfois, que l'on insère des pauses, que l'on place l'accent sur le mot principal, etc. Dans la synthèse vocale informatique, on essaie de reproduire de telles attitudes de comportement à la parole, et cela n'a rien de facile.

Selma Yilmazyildiz avec une maquette d'Anty dans ses bras, en compagnie des collaborateurs du projet Anty (de g. à dr. : Werner Verhelst, Ivan Hermans, Kristof Goris, Jelle Saldien et Lukas Latacz).

LA PAROLE EN DIT PLUS LONG QUE LE TEXTE

Au fil des études sur la parole numérique, on a renoncé pour la pratique à l'option très fondamentale de la synthèse articulatoire. Une « simulation physique » de la bouche, des cordes vocales et du nez pour parvenir à une « machine parlante » n'est vraiment pas simple, et aujourd'hui même irréaliste : comment mesurer le fonctionnement de tous les muscles de la tête pour ensuite simuler la « parole » ? Il y a les lenteurs de réaction des muscles qui influencent la cavité résonante permettant de produire des phrases complètes, où des lettres sont prononcées très différemment dans une combinaison et dans une autre. Prenons par exemple le « a » dans les mots « balle » et « pâtes », qui peut être ouvert ou fermé. Et décomposer la « séquence des sons » en composantes vocales ou en lettres n'est pas réaliste.

C'est pourquoi les applications actuelles reposent sur une approche plus pragmatique, fondée sur des bases de données de phonèmes, de mots et même de séquences de phrases connus. Pour la reconnaissance vocale, on essaie d'identifier ces séquences. Pour la synthèse vocale, elles sont liées entre elles par un « filtrage » afin que la voix de synthèse les prononce de la façon « la plus naturelle possible ». Comme la base de données des possibilités s'étoffe continuellement et que les ordinateurs sont de plus en plus puissants, on arrive peu à peu à une reconnaissance « acceptable » et à des voix de synthèse qui n'ont plus rien de dérangeant, du moins pour les « applications limitées ».

Outre la composition (ou la reconnaissance) du texte, les chercheurs, dans le domaine de la « parole numérique » et donc tant au niveau de la reconnaissance que de la synthèse, sont également confrontés au problème que le « texte » ne représente qu'une partie de la « langue parlée ». Un texte, c'est du vocabulaire et du contenu. Ce qui est déjà assez difficile dans la synthèse vocale, car comment l'ordinateur peut-il prononcer la phrase écrite : « Il est mon hôte » ? Vient-il chez moi ou est-ce moi qui vais chez lui ? Ou encore le nombre 1018 dans « la collection 1018 » (la collection dix dix-huit) ? Outre le vocabulaire, il y a « l'intonation » et « l'émotion ». On peut dire quelque chose textuellement, mais l'intonation de la voix et le contexte de la conversation font que l'interlocuteur comprend le contraire : « Et tu crois cela ? »

Dans la synthèse vocale, on était déjà satisfait que l'ordinateur puisse prononcer son message « sur un ton égal ». Dans la reconnaissance vocale, ce volet ne fait plus depuis longtemps l'objet de recherches aussi approfondies que la « reconnaissance du contenu » proprement dite. Il existe cependant des applications économiquement importantes pour lesquelles les recherches sur la « reconnaissance des émotions » sont cruciales. Un premier exemple en est le contrôle sur la base des intonations vocales lors d'une conversation par GSM pour déterminer, par exemple, si un chauffeur de poids lourd est fatigué. Un autre exemple, et une demande de l'industrie, est le contrôle des conversations téléphoniques basées sur ordinateur. Ce dernier ne comprend pas toujours le client dès la première fois, ce qui peut l'énerver et la conversation devient alors totalement incompréhensible pour l'ordinateur. Mais l'ordinateur ne fait que « demander confirmation ». Par exemple, vous voulez commander par téléphone un billet d'avion à destination de « Amsterdam » et l'ordinateur comprend « Atlanta ». Vous répétez alors votre demande une deuxième fois et vous lâchez ensuite un « merde ! », sur ce il vous propose « Berne ? » Et ainsi de suite. Personne ne veut perdre de clients à cause d'un ordinateur. Donc, « l'intonation vocale émotionnelle » doit pouvoir - indépendamment du message proprement dit - être identifiée, car l'appel doit alors être repris par un opérateur humain qui recollera les morceaux. Il en va de même dans les centres d'appels. A la VUB, on mène également des recherches sur la « reconnaissance émotionnelle », ce dont pourra bénéficier le projet Anty.

SYNCHRONISATION ET DIALOGUE

Une application actuellement réalisable - du moins pour le groupe de recherche de la VUB - est la synchronisation de deux voix qui disent la même chose, la suppression éventuelle d'une voix avec le maintien de plusieurs aspects de l'autre voix, etc.

Le karaoké en est un exemple pratique. Il est en effet possible de remplacer la fréquence vocale, le timbre, le rythme et les pauses respiratoires d'un amateur qui chante faux par le timbre du chanteur original. On a alors l'impression que l'amateur chante juste alors que ce n'est pas le cas. De la même manière, on peut postsynchroniser des films (substitution de la voix lors de prises de vue en extérieur, par le texte enregistré en studio sans le problème actuel de décalage entre le « mouvement des lèvres » et les « paroles »). Cela permet de mieux synchroniser le texte doublé avec le mouvement des lèvres. Ou rendre la voix de la doublure parfaitement similaire à celle de l'acteur original (ce que la plupart des acteurs locaux ne souhaitent pas vraiment). Au départ d'un enregistrement, il est possible de réaliser un chœur de plusieurs voix. Et, ici également, bien que les recherches ne soient pas menées en fonction du projet Anty, on peut très bien se figurer que cette technique y sera malgré tout intégrée : peut-être qu'Anty pourra chanter avec les enfants hospitalisés ?

TRAVAIL DE FIN D'ETUDES SUR ANTY

Parallèlement à ces « recherches dont les résultats pourront être appliqués ultérieurement », le DSSP a lancé des recherches spécifiquement axées sur Anty. Contrairement au projet de robot Anty de *Robotics and Multibody Mechanics* (R&MM), le DSSP a entamé les travaux sans disposer de fonds spécifiquement destinés à cet effet. Les recherches se font par l'intermédiaire de travaux de fin d'études.

VUB ETRO

Le service Electronique et Informatique constitue (de nouveau) un département unique à la VUB afin de pouvoir exploiter la synergie entre les deux : l'informatique repose sur l'électronique et les développements électroniques font largement appel à la programmation. Mais c'est surtout parce que la VUB se spécialise dans la recherche fondamentale du traitement des signaux. La conjonction des deux spécialités est encore plus fondamentale parce que tant l'informatique que l'électronique (même le développement matériel) reposent sur les « mathématiques », le développement d'algorithmes appropriés qui, une fois implémentés dans les logiciels ou le matériel, font dans l'application ce que l'on veut qu'ils fassent.
Le département Electronique et Informatique (ETRO) est subdivisé en plusieurs groupes de recherche. L'étude de « matériel électronique » est davantage assurée par le LAMI (puces et capteurs). Le plus grand est le groupe IRIS (traitement de l'image). Vient ensuite TELE (télécoms, et donc de l'informatique pure). Le dernier-né est Image & Son, un groupe pluridisciplinaire constitué de membres des groupes de recherche IRIS et DSSP. Bien entendu, il y a également le DSSP, parole numérique et traitement audio, une discipline en fait assez récente. L'audio numérique a commencé (en Belgique) au début des années '70 (dans les universités de Gand et de Mons). A la VUB, les recherches ont commencé avec le doctorat du professeur Verhelst qui a constitué un groupe d'une dizaine de chercheurs et d'étudiants. Il s'agit plutôt de « recherche stratégique » : partir d'interrogations pratiques sur les modèles de données de « parole » pour déboucher sur des applications utilisables et des perspecitves fondamentales.

Lukas Latacz du projet Anty et l'étudiante Selma Yilmazyildiz (master of applied computer science, ir dept. VUB), lors de l'enregistrement.

Un premier travail de fin d'études porte sur « une voix pour Anty ». Etant donné que les recherches portant sur des « réponses sensées, sans intervention humaine » ne sont encore qu'une perspective lointaine, et qu'un robot muet n'est pas très communicatif, on a songé à adopter une autre approche. Anty fonctionnera pour certaines opérations comme une « marionnette télécommandée », mais on souhaite également que le robot à câliner ait des « réactions spontanées » vis-à-vis de l'enfant. Anty a reçu pour cela un « charabia » avec lequel il peut réagir « spontanément » aux sentiments exprimés par l'enfant. Et il s'agit là d'un projet réalisable. Au départ d'une base de données d'enregistrements vocaux d'Ivan Hermans, le père spirituel d'Anty (non pas que cela doive être sa voix, mais cela serait assurément amusant), des sons et des parties de mots sont générés par synthèse vocale. Ils sont combinés de manière à former un langage enfantin dont certaines intonations qui suscitent une réaction « émotionnelle » sont réalisées via des algorithmes.

L'être humain reconnaît les émotions, même s'il ne comprend pas la langue. Il s'agit d'intonations, de rapidité d'élocution, de qualité de voix (par exemple un chat dans la gorge)... qui forment ensemble un certain nombre d'intonations de base. On a ainsi des intonations qui signifient « l'interdiction », la « réprimande » ou « attirent l'attention » (une voix informatique qui annonce, sur un ton neutre, « un incendie s'est déclaré et vous êtes invité à quitter le bâtiment » aura vraisemblablement moins d'impact que quelqu'un qui crie « au feu, au feu ! »). Il y a également l'intonation « informative » (qui sera utile à un stade ultérieur en combinaison avec l'image, lorsqu'Anty sera utilisé pour expliquer des choses telles que « tu vas être opéré »). Des voix émotionnelles directement utiles reflètent des attitudes comme « consoler » et « encourager », mais aussi des intonations « d'approbation » ou de « surprise » (par ex., en combinaison avec « vision » : si une personne « connue » entre dans la pièce).

L'identification, l'extraction et l'intégration dans le « langage Anty » de caractéristiques « pertinentes » devraient pouvoir être réalisées dans le cadre d'un travail de fin d'études, et si possible de travaux complémentaires. Anty pourrait ainsi recevoir une voix propre. La puissance informatique ne doit pas être intégrée dans Anty (il fonctionne sur batterie et doit être le plus léger possible). Le robot doit cependant pouvoir compter sur une communication fiable avec le serveur. La transmission du résultat de la synthèse ne requiert aucune puissance informatique. Ce qui n'est pas le cas de la génération en temps réel d'un langage émotionnel, mais c'est là un problème de puissance informatique sur le serveur.