Synthèse Vocale et Reconnaisance de la Parole : Droites Gauches et Mondes Parallèles

Thierry Dutoit (*), Laurent Couvreur(**), Fabrice Malfrère(***),
Vincent Pagel(***), Christophe Ris(*)

(*) Faculté Polytechnique de Mons, Service TCTS, 9 rue de Houdain, B-7000 Mons, Belgique
(**) MULTITEL asbl, 1 av. Copernic, B-7000 Mons, Belgique
(***) Babel Technologies SA, 33, Boulevard Dolez, B-7000 Mons, Belgique

RESUME

Parler et reconnaître ce qui est dit sont indiscutablement deux problèmes liés.
De fait, les réponses apportées depuis plus de 40 ans par la synthèse vocale et la reconnaissance de la parole puisent leurs ressources dans un substrat mathématique et linguistique commun (le traitement du signal et la linguistique informatique), motivé par un objectif identique (le dialogue homme-machine) et un objet d'étude unique (la parole).
Aujourd'hui encore, cependant, ces deux domaines évoluent comme deux mondes parallèles. Les personnes impliquées (chercheurs et ingénieurs) appartiennent à l'un ou (xor) l'autre de ces mondes, très rarement aux deux (la séparation étant souvent, et de façon assez paradoxale, matérialisée par un changement d'étage dans un bâtiment commun). Les transfuges ne sont pas légion. Les outils eux-mêmes diffèrent autant qu'ils se ressemblent.
Partant d'une analyse des convergences et des divergences entre les deux problèmes étudiés, cet article dresse une perspective historique des idées et des outils mis en oeuvre dans chaque domaine, et tente de faire le point sur les raisons et les conséquences de leurs évolutions parallèles.
Un reconnaisseur de parole ne devrait-il pas ressembler à un synthétiseur vocal fonctionnant à l'envers (et réciproquement)? Un synthétiseur ne devrait-il pas inclure automatiquement des fonctions de reconnaissance? Ni plus ni moins, bien au contraire!

I. INTRODUCTION

Parler et reconnaître ce qui est dit sont indiscutablement deux problèmes liés (section II). Il serait cependant faux de croire qu'il s'agit de problèmes inverses l'un de l'autre. Partant d'une analyse des convergences et des divergences entre les deux problèmes étudiés, cet article dresse une perspective historique des idées et des outils mis en oeuvre dans chaque domaine (section III), et tente de faire le point sur les raisons et les conséquences de leurs évolutions parallèles (section IV).

II. DES PROBLEMES LIES

Sur un plan purement physiologique, et indépendamment des techniques « modernes » utilisées en traitement automatique de la parole, la production et la reconnaissance de parole sont intimement liées.

Le modèle de Wernicke-Geschwind (Fig. 1) donne un aperçu fonctionnel de l'organisation des traitements réalisés par le cerveau pour ces deux tâches. Ce modèle a pour seul intérêt de mettre en évidence l'existence de tâches intermédiaires, d'ailleurs associées à des zones particulières du cortex. Ainsi, les stimuli visuels (lecture) ou auditifs (écoute) sont intégrés avant d'être transmis à l'aire de Wernicke, qui se charge de la compréhension, multimodale, distribuée, et associative (par opposition à analytique) du sens des mots. Cette tâche consiste à associer aux mots, naturellement dépendants de la langue, des « images » abstraites, qui peuvent être au contraire assez peu liées à la langue utilisée comme support. Ces images, produites dans l'aire de Wernicke à partir des stimuli visuels ou auditifs ou par la pensée abstraite, sont alors transmises grâce à un faisceau de neurones à l'aire de Broca (située, comme l'aire de Wernicke, dans l'hémisphère gauche). Elle y sont soumises *a posteriori* à un traitement linguistique complexe (cette fois plus analytique : sémantique, syntaxique, morphologique, phonétique), dépendant de la langue, et visant à énoncer une phrase.

Un tel modèle trouve sa justification dans son pouvoir de prédire les disfonctionnements du langage. Ainsi, il prédit l'*aphasie de Wernicke*, due à une lésion dans l'aire de Wernicke provoquant tout naturellement un rupture de la compréhension de la parole et du texte, en maintenant intacte la capacité de parler (mais la parole est alors vide de sens). Il prédit également l'*aphasie de Broca*, qui maintient les facultés de compréhension orale ou visuelle, mais handicape gravement la capacité de construire des phrases complexes (les patients utilisent des mots clés, non fléchis). Le modèle explique l'*aphasie de conduction*, liée à une lésion du faisceau de neurones reliant les aires de Wernicke et de Broca, et handicapant gravement la lecture orale, mais pas la compréhension ni la parole spontanée. Enfin, une lésion dans les régions du cortex moteur empêchent la production de toute parole articulée (lue ou spontanée), par impossibilité de contrôle des nombreux muscles intervenant dans ce processus.

Le séquencement du modèle de Wernicke-Geschwind peut ne pas être respecté dans certains cas. En particulier, on peut très bien lire un texte ou répéter une phrase sans rien en comprendre. C'est d'ailleurs ce que fait un jeune enfant, ou un adulte lisant un texte comme *Jabberwocky*, poème de Lewis Caroll (1871), traduit par H. Parisot :

Il était grilheure; les slictueux toves
Gyraient sur l'alloinde et vriblaient:
Tout flivoreux allaient les borogoves;
Les verchons fourgus bourniflaient.

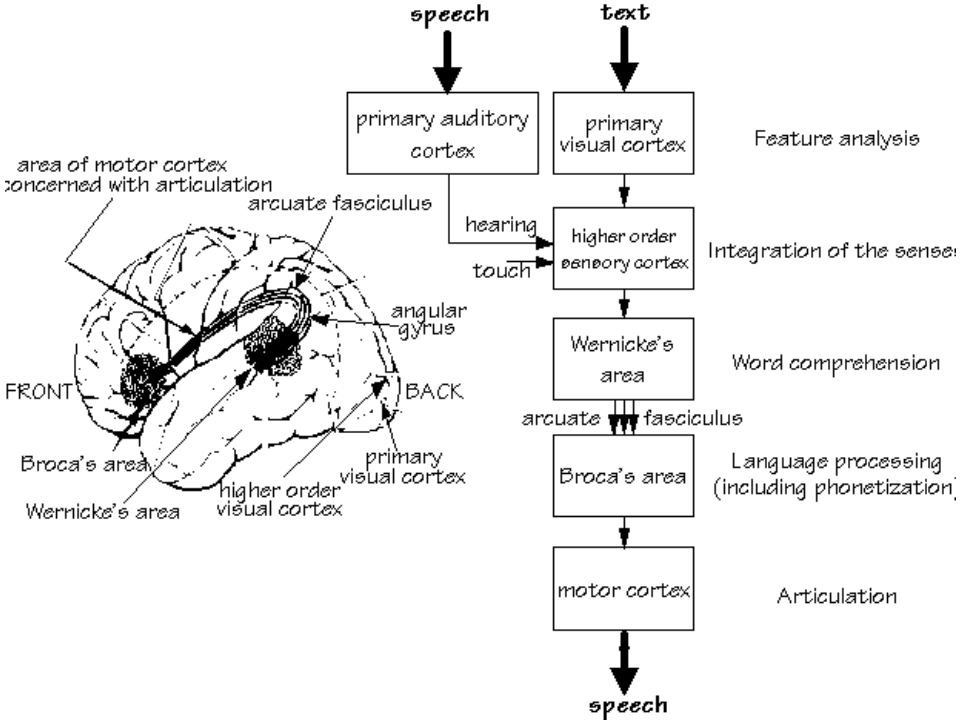

Figure 1. Le Modèle de Wernicke-Geschwind (d'après [1]).

On obtient un effet similaire en remplaçant les mots d'une phrase par des mots génériques portant la même nature syntaxique mais dépourvus de sens : « on verbe un nom adjectif en verbant les noms d'un nom par des noms adjectifs verbant l'adjectif nom adjectif mais verbés de nom ». C'est également la situation dans laquelle se trouve lorsqu'on cherche à répéter une phrase dans une langue que l'on ne comprend pas. La lecture ou la répétition est cependant nettement plus difficile dans ces conditions, ce qui suggère que la compréhension (aire de Wernicke) joue un rôle majeur dans les mécanismes de reconnaissance (visuelle ou auditive) du langage humain.

La compréhension du message en permet en effet la prédiction (partielle), qui est mise à profit pour :

- Résoudre les *ambiguités* du langage : les phrases que nous lisons ou entendons possèdent un niveau d'ambiguïté (phonétique, morphologique, syntaxique ou sémantique) énorme en regard de ce que nous en percevons. La plupart du temps, nous n'en avons même pas conscience : nous prédisons plus vite que nous n'analysons.
- Accroître la *robustesse* du processus de reconnaissance. Un auditeur ou un lecteur humain est en effet capable de reconnaître des mots mal produits (prononcés ou écrits) ou masqués (par du bruit auditif ou raturés). Mieux encore, il est capable de suivre une conversation lorsque plusieurs personnes parlent, ou de suivre une écriture parmi plusieurs superposées.

On peut aller plus loin : bien au delà de cette simple mise en commun de certaines fonctions cérébrales en lecture et en écoute, on peut présumer de l'existence d'une boucle de rétroaction entre la production et perception (Fig 2). Ainsi, lorsque l'enfant, après avoir appris à reconnaître les stimuli auditifs et visuels qu'on lui soumet en vue de communiquer, cherche à les produire par lui-même, il ajuste son geste (articulatoire) à l'écoute qu'il en a. On a d'ailleurs mis en évidence l'importance de cette boucle en traitement automatique de la parole, sous le nom d'*effet Lombard* : notre parole s'ajuste automatiquement à l'environnement dans lequel elle est produite (caractérisé par un niveau de bruit, convolutif et additif) grâce au retour auditif que nous en avons. Les personnes atteintes de surdité ont ainsi (et paradoxalement) plus de mal à se faire comprendre : du fait de la dégradation de cette boucle de rétroaction, leur parole ne s'adapte plus à l'environnement dans lequel elle est produite. Un effet similaire existe dans le processus d'écriture, qui requiert une analyse en temps réel des résultats du geste scriptural. On écrira différemment sur une page blanche et sur une page à l'arrière plan très texturé.

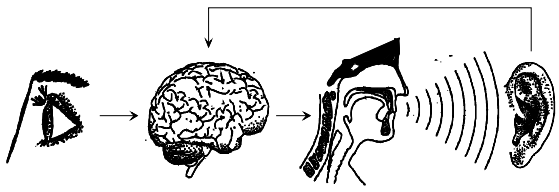

Figure 2. Un diagramme schématique du processus de lecture (d'après [2]).

Par contre, on comprend qu'il est fallacieux de croire que production et reconnaissance (de parole ou d'écriture) sont deux processus inverses l'un de l'autre : s'ils le sont bien fonctionnellement (c.-à-d. en termes d'entrées-sorties), il n'en va pas de même structurellement (en termes de mis en œuvre). Production et reconnaissance humaines sont fonctionnellement distinctes, avec l'aire de Wernicke pour point commun, et l'utilisation des fonctions de

reconnaissance pour guider les fonctions de production (boucle de réaction).

III. DES HISTOIRES PARALLELES

Si reconnaissance et production sont ainsi liées chez nous, il n'en va pas de même en ingénierie de la parole. Le cerveau humain intègre en effet très fortement la compréhension (dans l'aire de Wernicke) dans les processus de lecture et de reconnaissance, alors que l'ordinateur n'est encore que très peu apte à comprendre le texte qu'il est sensé lire ou reconnaître ; il peut tout au plus en saisir l'organisation syntaxique, et le plus souvent de façon peu hiérarchisée (le texte peut être segmenté en groupes, mais il reste difficile de retrouver les relations de dépendance entre ces groupes, et pour cause : la mise en évidence de ces relations requiert le plus souvent l'accès aux sens des mots). Le lieu de croisement naturel des activités cérébrales de reconnaissance et de production se trouve donc absent dans l'organisation des machines appelées à simuler ces activités : les synthétiseurs et les reconnaisseurs.

Il s'ensuit une évolution parallèle des recherches en reconnaissance vocale et en synthèse de parole, depuis près de 40 ans. Les personnes impliquées (chercheurs et ingénieurs) appartiennent à l'un ou (xor) l'autre de ces mondes, très rarement aux deux (la séparation étant souvent, et de façon assez paradoxale, matérialisée par un changement d'étage dans un bâtiment commun). Les transfuges ne sont pas légion. Les outils eux-mêmes diffèrent autant qu'ils se ressemblent.

En particulier, on est bien loin de la machine qui parlerait en fonction du retour « auditif » qu'elle aurait du son qu'elle produit.

III.1. Une brève histoire de la synthèse de la parole

L'objectif de la synthèse de la parole est de produire un signal intelligible et naturel. Le principal problème pour y parvenir est de simuler correctement la *coarticulation* entre les sons, et de gérer naturellement la *prosodie* (intonation et durée) qu'on leur applique

La synthèse vocale a connu trois grandes étapes technologiques, qui coexistent aujourd'hui commercialement : la synthèse *par règles*, la synthèse par *concaténation de diphones*, et le synthèse *par sélection d'unités dans une grande bases de données.*

Les synthétiseurs par règles sont basés sur l'idée que, si un phonéticien expérimenté est capable de «lire» un spectrogramme, il doit lui être possible de produire des règles permettant de créer un spectrogramme artificiel (Fig. 3) pour une suite de phonèmes donnée. Une fois le spectrogramme « dessiné », il ne reste plus alors qu'a générer le signal correspondant (à l'aide de générateurs et de résonateurs électriques) . Cette technique a été en vogue entre 1965 et 1985, surtout sous l'impulsion du MIT [3]. Elle est fort peu gourmande en mémoire (à peine 10 koctets pour les règles décrivant la coarticulation d'une voix). Basée sur la seule expertise humaine, elle fournit très difficilement un signal naturel.

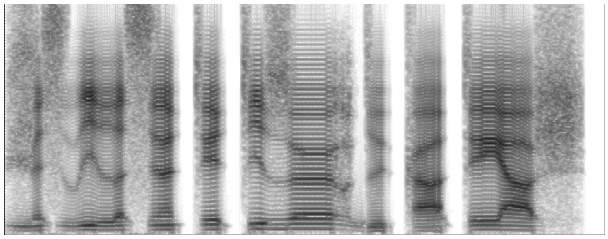

Figure 3. Spectrogramme d'une phrase synthétisée par règles.

Les synthétiseurs par concaténation de diphones (un diphone est une unité acoustique qui commence au milieu de la zone stable d'un phonème et se termine au milieu de la zone stable du phonème suivant) procèdent au contraire par mise bout à bout de segments acoustiques *déjà* coarticulés, extraits d'une base de données de signaux de parole (et non pas modélisé par un expert). Il s'ensuit que, contrairement aux cibles phonétiques de l'approche précédente, la production de parole fluide en synthèse par concaténation ne requiert qu'une étape de *concaténation* qui s'accompagne d'un *lissage* purement *acoustique* des discontinuités au droit des points de concaténation. Un problème supplémentaire apparaît cependant, du fait que les diphones utilisés ne respectent pas en général la prosodie que l'on cherche à produire. Il faut donc en modifier la durée et l'intonation avant de procéder à la concaténation, sans que ces opération ne dégradent la qualité des unités. Diverses techniques se sont succédées depuis 25 ans pour permettre ce type d'ajustement : synthèse basées sur le modèle LPC [4], synthèse PSOLA dans le domaine temporel [5], synthèse mixte MBROLA [6]. Ce type de technique fournit une très bonne intelligibilité et un naturel acceptable pour de nombreuses applications, mais la parole est souvent hyper-articulée, et l'intonation (produite par règles) reste peu naturelle. Un système de synthèse par diphones de bonne qualité nécessite entre 1 et 5 Moctets par voix (pour stocker les quelques 1500 diphones correspondants, soit environ 3 minutes de parole).

Enfin, on assiste depuis quelques années à un important bouleversement, avec l'arrivée de techniques de sélection d'unités dans une grande base de données [7, 8]. Plutôt que de garder qu'un exemplaire de chaque diphone de la langue, on puise ici dans plusieurs heures de parole, préalablement segmentée phonétiquement. Au moment de choisir les segments à mettre en œuvre (souvent des diphones), plusieurs instances d'une même unité phonétique sont alors disponibles, avec des prosodies différentes et positionnées (dans le corpus) dans des contextes phonétiques différents. Il faut donc, pour réaliser au mieux la synthèse, choisir les segments dont le contexte est le plus proche de la chaîne phonétique à synthétiser, dont la prosodie se rapproche également le plus de la prosodie à produire, et dont les extrémités ne présentent pas trop de discontinuité spectrale l'une par rapport à l'autre. On procède donc en général par programmation dynamique (algorithme de

Viterbi) dans le treillis des segments utilisables, de façon à minimiser un coût de synthèse global, qui tient compte (Fig. 4): du *coût de représentation* (dans quelle mesure les segments choisi correspondent-ils au contexte phonétique et prosodique dans lequel on les insère?) et d'un *coût de concaténation* (dans quelle mesure la juxtaposition des segments choisis amène-t-elle des discontinuités ?). Ces techniques ont permis récemment de produire de la parole dont l'intelligibilité et le naturel rendent possible la confusion avec une prononciation humaine. Elle impliquent cependant un accès très rapide à plusieurs Goctets de données.

On constate donc qu'en synthèse de parole, la technologie a évolué d'une approche basée sur des modèles (règles) vers une approche basée sur des exemples (diphones). Le paradigme gagnant semble être celui qui laisse le dernier mot aux données.

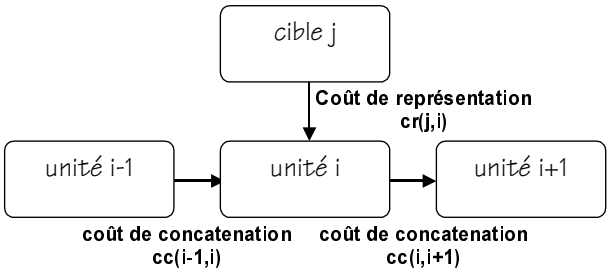

Figure 4. Sélection d'un exemplaire de diphone (unité *i*) pour représenter le diphone cible (*j*) et coûts afférents.

III.1. Une brève histoire de la reconnaissance vocale

Les premiers succès en reconnaissance vocale ont été obtenus dans les années 70 à l'aide d'un paradigme de reconnaissance de mots « par l'exemple ». On fait prononcer un ou plusieurs exemples de chacun des mots susceptibles d'être reconnus, et on les analyse sous forme de *vecteurs acoustiques* (typiquement : un vecteur de coefficients LPC ou assimilés toutes les 10 ms). L'étape de reconnaissance proprement dite consiste alors à analyser le signal inconnu sous la forme d'une suite de vecteurs acoustiques similaires, et à comparer la suite inconnue à chacune des suites des exemples préalablement enregistrés (Fig. 5). Ce principe de base n'est cependant pas implémentable directement : un même mot peut en effet être prononcé d'une infinité de façons différentes, en changeant le rythme de l'élocution. La superposition du signal inconnu aux signaux de base doit dès lors se faire en acceptant une certaine « élasticité » temporelle, formalisée mathématiquement par l'algorithme DTW (*Dynamic Time Warping*) [9]. C'est le principe de la reconnaissance implémentée dans nos GSM. On obtient de faibles taux d'erreurs (dasn un environnement calme), mais la reconnaissance est intrinsèquement monolocuteur, le vocabulaire est petit (< 100 mots), et les mots doivent être prononcés isolément.

Dès que l'on cherche à concevoir un système réellement multilocuteurs, à plus grand vocabulaire, et s'adaptant facilement à une application, il devient nécessaire de mener la reconnaissance sur base d'*unités de parole* de plus petite taille, typiquement des phonèmes ; on parle alors de *reconnaisseurs phonétiques*. On ne se contente plus alors d'exemples de ces unités, mais on cherche plutôt à en déduire un *modèle* (un modèle par unité), qui sera applicable pour n'importe quelle voix.

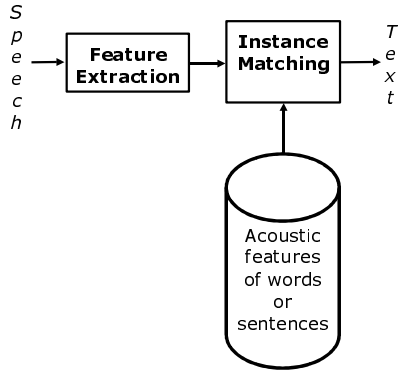

Figure 5. Reconnaissance "par l'exemple" (DTW).

Le formalisme de reconnaissance de la parole est souvent décomposé en plusieurs modules, généralement au nombre de quatre (Fig. 6):

- Un *module de traitement du signal* (qui produit typiquement un vecteur de coefficients LPC ou assimilés toutes les 10 ms) ;
- Un *module acoustique* qui peut produire une ou plusieurs hypothèses phonétiques pour chaque segment de parole de 10 ms, associées en général à une probabilité. Ce générateur d'hypothèses locales est généralement basé sur des *modèles statistiques* d'unités élémentaires de parole (typiquement des phonèmes), le plus souvent constitués de lois statistiques paramétriques dont on ajuste les paramètres pour « coller » au mieux aux données, ou de réseaux de neurones artificiels [10]. Un tel générateur d'étiquettes phonétiques intègre toujours un *module d'alignement temporel* qui transforme les hypothèses locales (prises sur chaque vecteur acoustique indépendamment) en une décision plus globale (prise en considérant les vecteurs environnants). Ceci se fait le plus souvent via des modèles de Markov cachés [11] ;
- Un *module lexical* qui interagit avec le module d'alignement temporel pour forcer le reconnaisseur à ne reconnaître que des mots existants effectivement dans la langue considérée. Un tel module lexical embarque en général des *modèles des mots* de la langue (les modèles de base étant de simples dictionnaires phonétiques ; les plus complexes sont de véritables *automates probabilistes*, capables d'associer une probabilité à chaque prononciation possible d'un mot).
- Un *module syntaxique* qui interagit avec le module d'alignement temporel pour forcer le reconnaisseur à intégrer des contraintes syntaxiques, voire sémantiques. Les connaissances syntaxiques sont généralement formalisés dans un *modèle de la langue*,

qui associe une probabilité à toute suite de mots présents dans le lexique.

Cette approche donne de bons résultats pour de la parole continue, à moyen ou grand vocabulaire, et indépendamment du locuteur, à condition que le signal analysé soit exempt de bruit. Les problèmes qui restent ouverts pour le moment sont essentiellement :

- Ceux liés à la robustesse de tels systèmes [21] : aux bruits d'environnement (bruits, musique, autres locuteurs), à la réverbération, aux changements de microphone, aux variations de conditions de transmission téléphonique, ainsi qu'aux conditions d'élocution difficiles (stress, effet Lombard, bruit de respiration, vitesse d'élocution inhabituelle).
- Ceux liés à la modélisation des contraintes syntaxiques et sémantiques de la langue, indispensable à la mise au point de systèmes de dictée vocale réellement efficaces.

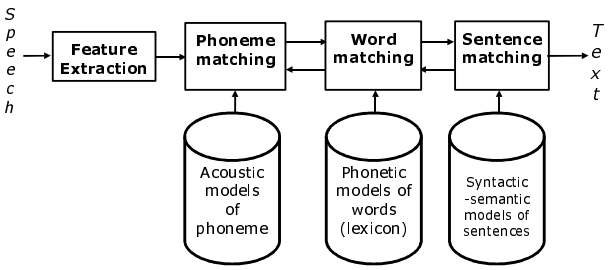

Figure 6. Reconnaissance par modélisation statistique de séquences d'unités acoustiques.

Ainsi, en reconnaissance vocale, l'évolution technologique s'est faite tout naturellement d'une reconnaissance basées sur des exemples à une reconnaissance basée sur des modèles (statistiques, et entraînés automatiquement).

IV. DES POINTS DE CONVERGENCE

S'il est clair que synthèse et reconnaissance ne sont en principe pas deux opérations inverses, il n'en reste pas moins que ces deux problèmes embrassent un ensemble de champs d'études commun, et que leur résolution met parfois en œuvre des techniques similaires.

Ainsi, les reconnaisseurs vocaux actuellement commercialisés embarquent de solides connaissances acoustiques, phonétiques, phonologiques, et lexicales. Ils ont également accès à un embryon de syntaxe et de sémantique, grâce au calcul de la probabilité d'une suite de *n* mots par les *n-grams* [23]. Les synthétiseurs actuels possèdent évidemment de solides connaissances acoustiques, phonétiques, phonologiques, lexicales, et syntaxiques (à travers les *n-grams* de natures grammaticales, et la structuration en groupes prosodiques), mais n'utilisent que très peu de données sémantiques.

L'objectif de la formalisation de ces connaissances est cependant assez différent dans les deux cas. La reconnaissance vise le général (toutes les variations possibles de prononciations d'une même phrase, laquelle sera d'ailleurs souvent mal formée sur le plan linguistique), tandis que la synthèse s'intéresse au particulier (une seule prononciation, par une seule voix, avec le plus d'intelligibilité et de naturel possible, ce qui implique de modéliser finement les caractéristiques vocales humaines). Il s'ensuit que les bases de données nécessaires en reconnaissance sont nettement plus importantes qu'en synthèse, et qu'elles seront le plus souvent fort différentes. C'est également la raison pour laquelle, jusque récemment, la synthèse faisaient peu appel à la modélisation statistique.

Parmi les points de rencontre actuels entre ces deux mondes, on trouve notamment :

- Les outils d'analyse (acoustico-phonétique) du signal de parole. Contrairement à ce que l'on pourrait attendre, les modèles du signal de parole ne sont pas identiques : ils sont plus analytiques en synthèse, alors qu'on cherche à les rendre indépendants du locuteur en reconnaissance. Des points de rencontre existent cependant. Ainsi par exemple, la segmentation phonétique d'un corpus peut être menée par comparaison entre le signal à segmenter et la sortie d'un synthétiseur prononçant ces phonèmes à un rythme imposé [12]. La modélisation multibande, initialement développée en codage et utilisée par la suite en synthèse, est maintenant utilisée pour accroître la robustesse des reconnaisseurs [13].
- Les outils d'analyse de texte, pour la préparation de bases de données, indispensables à l'entraînement des modèles de mots et de langue des reconnaisseurs, et à la mise au point de phonétiseurs et de générateurs de prosodie en synthèse. Les *n-grams* sont utilisés de part et d'autres, encore que pas pour les mêmes raisons. Même chose pour les arbres de décision (*CART : classification and regression tree*).
- En particulier, la formalisation de contraintes séquentielles sous la forme de transducteurs d'états finis pondérés (WFST : weighted finite state grammars) [14], qui permettent de transcrire une séquence de symboles en une autre. Ce formalisme simple est équivalent aux modèles de Markov cachés discrets. Les WFST ont l'avantage d'être très bien cernés sur le plan mathématique, et on dispose d'algorithmes pour les composer (entre eux), les déterminiser (les reformuler sous une forme ne présentant pas d'ambiguïté locale, et les minimiser (en taille). Ils peuvent être *réversibles*. Ils sont utilisés, par exemple, pour effectuer la phonétisation, la lemmatisation (décomposition d'un mots en racines et affixes), l'étiquetage syntaxique, ou leurs inverses.
- La réversibilité des algorithmes d'analyse de textes, qui reste un sujet porteur (autrefois le plus souvent associé au moteur d'inférence de PROLOG, aujourd'hui aux transducteurs probabilistes). On en trouve des exemples en phonétisation [15, 16] ou en analyse de l'intonation [17].
- Enfin, les recherches en dialogue homme-machine [22], ou en traduction orale. Pour de telles

applications, il est pratique que le formalisme de description de l'objet 'parole' soit identique pour les étapes de reconnaissance et de synthèse. Il est même indispensable que ce formalisme implique une certaine compréhension des messages (on en revient au modèle de Wernicke !). On trouve, parmi les groupes s'intéressant à ces sujets, des tentatives d'unification de représentation des connaissances [18, 19]. Il est probable que ces recherches constituent une voie d'avenir. Après tout, synthèse et reconnaissance se réconcilient naturellement à travers le dialogue.

CONCLUSION

Il y a 10 ans, Frank Fallside, professeur à l'université de Cambridge, fondateur de la revue internationale Computer Speech and Language (1983), et pionnier dans l'utilisation des réseaux de neurones pour la reconnaissance de la parole, écrivait déjà [20] :

"Human beings evidently learn to produce and perceive speech ``simultaneously" (…) By contrast in speech communication by machine, speech synthesis systems are designed quite separately from speech recognition systems, employing at present quite different techniques and importantly using labeled data. (…) A decomposition technique has been established that allows, in principle, the acquisition of speech to be built up for successively higher levels. The method offers one way of bringing together the specialist techniques of synthesis and recognition and in particular the use of prosody in each. Recent results will be given in the paper."

Disparu prématurément en 1993 (d'une crise cardiaque alors qu'il était en poste), ses derniers papiers sont tombés dans l'oubli, et ses idées n'ont jamais vraiment été appliquées. Jamais, ou … pas encore ?

RÉFÉRENCES

1. Kandel, E.K., Schwartz, J.H. (1985). *Principles of Neural Science*, Appleton & Lange.
2. Holmes, J.N., (1988), *Speech Synthesis and Recognition*, Van Nostrand Reinhold, London.
3. Allen, J., Hunnicut, S., Klatt D. (1987). *From Text to Speech: The MITTALK System*, Cambridge University Press.
4. Olive, Joseph. P., Liberman, Mark Y. (1985) Text to Speech--An overview, *J. Acoust. Soc. Am*. Suppl. 1 78, S6.
5. Moulines, E., & Charpentier, F., (1990). "Pitch Synchronous waveform processing techniques for Text-To-Speech synthesis using diphones", *Speech Communication*, vol.9, 5-6.
6. Dutoit, T., et Leich, H., (1993). "MBR-PSOLA : Text-To-Speech Synthesis based on an MBE Re-Synthesis of the Segments Database", *Speech Communication*, v. 13, n° 3-4, pp.435-440.
7. Hunt, A. and Black, A., (1996). "Unit selection in a concatenative speech synthesis system using a large speech database", *Proc. of ICASSP*, Atlanta, Georgia, pp 373-376.
8. Möbius, B. (2000). "Corpus-based speech synthesis: methods and challenges" *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart)*, 6 (4), 87-116.
9. Silverman, H.F., Morgan, D.P. (1990). "The application of dynamic programming to connected speech recognition", *IEEE ASSP magazine*, vol.7, pp.6-25.
10. Bourlard, H. and Morgan, N., (1994). *Connectionist Speech Recognition*, Kluwer Academic Publishers, Boston.
11. Rabiner, L. R., "A Tutorial on Hidden Markov Models and Select Application in Speech Recognition", *Proceedings of the IEEE*, 1989, vol. 77, n°2, pp. 257-286.
12. Malfrère, F. , Deroo, O., Dutoit, T., (1998). "Phonetic Alignement : Speech Synthesis Based Vs. Hybrid HMM/ANN", *Proc. International Conference on Speech and Language Processing*, Sidney, Australia, pp. 1571-1574.
13. Dupont, S., (2000). "Etude et développement d'architectures multi-bandes et multi-modales pour la reconnaissance robuste de la parole", *Thèse de doctorat*, Faculté Polytechnique de Mons.
14. Mohri, M., (1997). Finite State Transducers in Language and Speech Processing, *Computational Linguistics*, 23:2.
15. Meng, H., Hunnicutt, S., Seneff, S., and Zue, V. (1996). "Reversible Letter-to-sound / Sound-to-letter Generation Based on Parsing Word Morpology". *Speech Communication*, 18, pp. 47-63.
16. Galescu, L., and Allen, J., (2001). "Bi-directional Conversion Between Graphemes and Phonemes Using a Joint N-gram Model", in *Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Blair Atholl.
17. Véronis, J., & Campione, E. (1998). Towards a reversible symbolic coding of intonation, 5th International Conference on Spoken Language *Processing ICSLP'98*, pp. 2899-2902, Sidney.
18. Penn, G., Carpenter, B. (1999). "ALE for Speech: A Machine Translation Prototype". *Proc. Eurospeech'99*, Budapest.
19. Seneff, S. (1998). "The Use of Linguistic Hierarchies in Speech Understanding, " *Proc. ICSLP 98*, Sydney, Australia.
20. Fallside, F, (1992), "Spoken language acquisition by computer", *Proc. ASA 124th Meeting* New Orleans.
21. Couvreur, L., Ris, C. (2002). "Model-based independent componenet analysis for robust multi-microphone automatic speech recognition", soumis à *ICSLP'02*, Septembre 2002, Denver.
22. Pietquin, O. Dutoit, T., (2002). "Modélisation d'un Système de Reconnaissance dans le cadre de l'Evaluation et l'Optimisation Automatique des Systèmes de Dialogue", accepté à *JEP 2002*, Nancy.
23. Rosenfeld R. (2000). "Two decades of statistical language modeling : Where do we go from here" *Proceedings of the IEEE*, 88(8), 2000.