La synthèse de la parole et le traitement automatique des langues.

Christophe d'Alessandro, Evelyne Tzoukermann

Editorial
La synthèse de la parole à partir du texte et le traitement automatique des langues

Christophe d'Alessandro (1),
Evelyne Tzoukermann (2)

(1) LIMSI-CNRS,
Bâtiment 508 - Université Paris XI, BP 133 - F91403 Orsay
cda@limsi.fr

(2) Bell Laboratories - Lucent Technologies
700 Mountain Avenue, Murray Hill, NJ 07974-0636
evelyne@research.bell-labs.com

RÉSUMÉ. *Dans cet éditorial, nous présentons le numéro thématique "Synthèse de la parole à partir du texte".*

ABSTRACT. *In this editorial, we introduce the "Text-To-Speech Synthesis" special issue.*

MOTS-CLÉS : *synthèse de la parole à partir du texte*

KEYWORDS: *text-to-speech synthesis*

1. Introduction

Le but de la synthèse de la parole à partir du texte est de calculer automatiquement un signal de parole correspondant à un énoncé écrit. Les sources du texte prononcé peuvent être diverses: lecture de journaux, système de réponse vocale, systèmes d'information, voire saisie au clavier de l'ordinateur. Ainsi les deux pôles de la synthèse de la parole sont d'un côté l'analyse et l'interprétation du texte, et de l'autre côté la prédiction des paramètres acoustico-phonétiques du son et la synthèse du signal. Le premier de ces pôles entre pleinement dans le champ du traitement automatique des langues (TAL) [COL97], et sera plus particulièrement abordé dans ce numéro thématique.

La transformation d'un texte en parole dépend de la capacité d'interpréter le texte écrit, tout comme le fait un orateur. Cela implique de comprendre le texte, ses nuances et ses connotations, la situation du discours et l'acte de parole à effectuer. Il serait parfaitement utopique d'imaginer qu'un tel programme est actuellement à la portée d'une machine . Néanmoins la synthèse de la parole à partir du texte nous a poussés à mettre en oeuvre des procédures automatiques pour plusieurs aspects de l'analyse linguistique. Ce besoin d'automatisation a par contrecoup mis en avant des problématiques jusqu'alors plus ou moins ignorées, et a fonctionné souvent comme une mise à l'épreuve par le biais de l'écoute des travaux théoriques. Le lecteur peu familier avec le traitement automatique de la parole pourra s'informer dans [LIE77, CAL89, MEL96, BOI00].

Un système de synthèse enchaîne d'assez nombreuses procédures de calcul [**?**]. Un premier ensemble de procédures, pour l'analyse et l'interprétation du texte, relève du traitement automatique de l'écrit. Tout d'abord, il s'agit de calculer la prononciation du texte. Lorsque les systèmes se sont trouvés confrontés à des sources de textes réelles et variées, il a fallu s'occuper des nombreuses "anomalies" graphémiques des textes, et de leur "normalisation": les sigles, les abréviations, les acronymes, les diverses sortes de mots composés, les symboles, les signes et les styles de ponctuation, les divers types de nombres et de chiffres. Tout cela nécessite des pré-traitements pour aboutir à une prononciation réaliste. De plus l'apparition de textes électroniques peu structurés, voire entachés de "fautes" plus ou moins systématiques (courrier électronique, textes sans accent etc.) pose des problèmes nouveaux que l'on se doit de traiter automatiquement.

Après cette normalisation, étape qui transforme un texte réel en suite de mots sous la forme orthographique, il s'agit de calculer la prononciation du texte. Dès la fin des années 60, les informaticiens se sont préoccupés de cette "orthographe inversée" qui permet de passer des lettres aux sons, la transcription automatique graphème-phonème ou "phonétisation": le premier programme de transription phonétique automatique pour la synthèse en français date de 1968. La phonétisation a été abordée par le biais de systèmes de règles, de lexiques spécialisés, de techniques d'apprentissage automatique etc. Mais cette question, assez simple en apparence, se révèle en fait très difficile si l'on veut traiter de problèmes comme les noms propres, les mots nou-

veaux et inconnus, les variantes de prononciation, les différents dialectes, idiolectes ou sociolectes. Cela pose d'importantes questions de phonologie, comme celles du "e" muet, de la coupe syllabique, des liaisons, de l'harmonie vocalique, de l'emprunt de phonèmes d'autres langues etc. De ce fait, la phonétisation est toujours un domaine de recherche actif, comme on le verra plus bas.

Les premiers systèmes de synthèse en français ne comprenaient que 2 procédures: une procédure de phonétisation (du texte aux phonèmes), et une procédure de synthèse acoustique de la suite de phonèmes. Ces systèmes étaient d'une façon générale intelligibles. Cependant, ils ont aussi montré qu'une suite de phonèmes, même correctement produite, ne suffisait pas pour former de la parole acceptable. Sans prosodie, la parole n'est pas vraiment humaine, même si dans certaines langues elle peut rester intelligible. Il fallait donc calculer également la prosodie: intonation et accentuation, c'est-à-dire le rythme de prononciation, les pauses dans l'énoncé, la mélodie, les variations de forces des syllabes. Ainsi la prosodie, domaine jusqu'alors relativement peu étudié a, si l'on peut dire, sauté aux oreilles des acousticiens, phonéticiens, linguistes et informaticiens s'occupant de synthèse, à partir des années 70-80. Mais comment calculer la prosodie, et quel est son statut linguistique ? Peut on définir une phonologie de la prosodie ? Quels sont les rapports entre syntaxe et prosodie ? Entre sémantique et prosodie ? C'est sans doute la synthèse de la parole qui a motivé et nourri tout un pan de recherche en linguistique, en phonétique, en TAL autour de ces questions.

La prosodie est par excellence le domaine de la nuance dans le discours et le royaume de l'orateur. Pour prédire la prosodie il faudrait comprendre parfaitement la situation de communication et d'énonciation, c'est-à-dire la structure et le sens de l'énoncé, les positions respectives du locuteur et de l'auditeur , le rapport du locuteur à son discours, avec ce que cela implique d'attidudes ou d'émotions. Cette analyse ne peut rester qu'assez incomplète aujourd'hui. Dans des situations d'applications bien spécifiées (par exemple la réponse vocale dans un système d'information, ou la lecture d'une notice technique) ces aspects sémantiques, pragmatiques, énonciatifs et discursifs seront bien sûr plus faciles à appréhender.

Lorsque les phonèmes à synthétiser sont connus, ainsi que les variations prosodiques, intervient l'étape de synthèse sonore à proprement parler. Les sons étant définis de manière abstraite, il reste à construire l'instrument, le synthétiseur vocal, et à le jouer. Au début de la synthèse, jusqu'aux années 70 [FLA73], ce sont les questions autour des traits et des règles phonologiques qui ont constitué une partie importante de l'enjeu en synthèse. Les acousticiens, phonéticiens et linguistes ont cherché où se trouvait (ou se cachait) l'invariance phonémique. Les systèmes de synthèse par règles sont apparus, avec une ambition d'universalité [ALL87] (systèmes multilingues, multilocuteurs, multistyles). Mais lorsqu'intervient le "locuteur" artificiel, ses caractéristiques et sa personalité vocale si l'on peut dire, on quitte le domaine du traitement automatique des langues pour entrer dans celui du traitement du signal acoustique, de la synthèse sonore. C'est ainsi qu'une part importante de la recherche en synthèse de la parole n'est pas le fait de laboratoires de linguistique ni d'informatique linguistique,

mais de laboratoires d'acoustique, de phonétique ou de génie électrique (traitement du signal, télécommunication, informatique).

Enfin intervient le domaine de l'évaluation: la compréhension et la perception de la parole synthétique par les auditeurs humains. Ces deux derniers domaines, celui de la synthèse du signal et celui de la perception, ne seront guère abordés directement dans ce volume, mais évoqués au fil des articles, par le biais de résultats d'évaluation ou de descriptions de systèmes.

2. Les articles de ce volume

Les auteurs discutent de travaux allant de la conversion orthographique à des environnements de développement en passant par des travaux sur la prosodie et des travaux présentant des systèmes généraux. Il a donc semblé logique d'organiser les ouvrages du particulier au général. Ainsi, une première partie comprend des travaux discutant des différents aspects de l'analyse des textes pour la conversion en parole de synthèse.

Nicolas Torzec, Thierry Moudenc et Françoise Émerard présentent leur système de prétraitement et d'analyse linguistique dans le système CVOX de France Telecom. Les auteurs font état des différents modules du système comprenant typiquement les étapes de normalisation, analyses morphologique, syntaxique et prosodique. Le système est adapté à la vocalisation automatique de messages électroniques, application qui pose de nombreuses difficultés dues à l'aspect hétérogène et mal formé de la langue écrite dans les messages de courrier électronique. De la sorte, le texte nécessite une étape de normalisation particulièrement adaptée à la forme d'entrée.

Frédéric Béchet présente aussi un système d'analyse de textes, en vue d'utilisations non seulement en synthèse de la parole mais aussi en reconnaissance automatique. Dans cette dernière application, la conversion phonétique de textes orthographiques nécessite souvent plus d'une réécriture pour refléter les différentes prononciations d'un mot par des locuteurs différents. Pour entraîner les systèmes de reconnaissance automatique, il est important de pouvoir produire un grand nombre de textes transcrits afin de construire les modèles acoustiques. À ces fins, l'auteur décrit un système de phonétisation à base de règles, produisant en sortie une transcription adaptée à l'application visée.

La deuxième partie comprend des articles traitant de la prosodie du français. Albert et Philippe Di Cristo présentent une approche métrique-autosegmentale de la prosodie, approche utilisée dans le système de synthèse SYNTAIX. La prosodie utilise l'information produite par une analyse syntaxique superficielle de la phrase et génère une représentation métrique de l'énoncé. La structure intonative, i.e. les séquences de tons et les frontières des groupes prosodiques, en découle, de même que l'organisation temporelle de l'énoncé. Le système est construit en vue de modéliser la prosodie de la synthèse multilingue; en effet, il devrait s'adapter par exemple à l'intonation de l'anglais avec le même formalisme métrique auto-segmental.

Si le système précédent prend en compte une analyse syntaxique superficielle, l'approche de Piet Mertens, Jean-Philippe Goldman, Eric Wehrli, et Arnaud Gaudinat génère les structures prosodiques à partir de structures syntaxiques riches. Deux sous-ensembles sont présentés, l'analyseur syntaxique Fips basé sur les modèles génératifs de Chomsky, et le système Mingus gérant la prosodie du français. A partir de l'arborescence syntaxique, la séquence de tons est associée à des groupes de mots, transformés ensuite en cibles acoustiques. Les cibles correspondent à un niveau de hauteur reflétant la position de l'accentuation dans le groupe syllabique.

Philippe Boula de Mareüil, Christophe d'Alessandro, Frédéric Beaugendre et Anne Lacheret-Dujour décrivent le produit de leur recherche consacrée à l'application d'une grammaire en tronçons à la génération de la prosodie. Le papier présente un étiquetage morpho-syntaxique, un analyseur "superficiel" et un parenthésage prosodique. Une interface syntaxe-prosodie est également décrite avec un nombre restreint de règles d'accentuation qui vont permettre de générer la prosodie. Des expériences quantitatives sont ensuite menées pour mesurer les tronçons définis par les frontières prosodiques et pour comparer l'approche à une autre basée essentiellement sur les mots outils et la ponctuation.

La troisième partie est constituée d'articles présentant de façon globale des systèmes de synthèse. Michel Morel et Anne Lacheret-Dujour décrivent leur système de synthèse vocale à partir du texte, nommé "Kali". L'approche, qui, contrairement à la plupart des systèmes existants, est développée pour les déficients visuels, rend compte de cinq modules couvrant prétraitement, syntaxe, prosodie, phonétisation, et traitement acoustico-phonétique. L'accent a été mis sur l'intelligibilité du système, même à un débit de parole élevé.

Philippe Boula de Mareüil, Philippe Célérier, Thierry Cesse, Serge Fabre, Carine Jobin, Pierre-Yves Le Meur, David Obadia, Benoît Soulage et Jacques Toen présentent le système de synthèse multilingue d'ELAN, qui comprend huit langues. Ce système de synthèse concaténative fonctionne suivant les principes de modification temporelle (algorithme TD-PSOLA). L'article soulève les questions intéressantes d'enregistrement d'une nouvelle voix et de variabilité dans la taille de l'inventaire acoustique, et discute de l'utilisation d'unités plus longues où les segments sont enregistrés avec leur information prosodique. De plus, il décrit un outil de recopie prosodique permettant de transférer l'information prosodique du signal sonore à la voix synthétisée.

Le troisième article de cette section présente un autre système complet, celui des Laboratoires Bell par Evelyne Tzoukermann. Le système est développé sur la plateforme multilingue, plateforme qui supporte actuellement plus d'une douzaine de langues [SPR98]. De plus, la partie analyse de texte est gérée à l'aide de transducteurs à états finis, tranducteurs fonctionnant avec des poids représentant la préférence d'une analyse sur une autre. Le système subit de constantes évaluations et traite avec aisance la plupart des problèmes de normalisations, y compris les adresses et sites internet. Le système peut aussi être consulté sur l'internet de manière interactive.

Finalement, la quatrième partie ne comprend qu'un seul article traitant d'environnement de développement, dans une perspective différente des autres articles. Les auteurs Nawfal Tounsi, Thierry Dutoit, Michel Bagein, Fabrice Malfrère, Alain Ruelle et Dominique Wynsberghe présentent un logiciel de synthèse vocale gratuitement mis à la disposition des chercheurs. Construit de manière modulaire, le logiciel permet aux chercheurs de remplacer différents modules du système et de bénéficier du travail d'autres chercheurs. La plateforme est également multilingue et a été testée dans plusieurs langues, comme le français et l'arabe.

3. Conclusion et perspectives

La possibilité d'une machine parlante n'est pas une question nouvelle: les premieres réalisations techniques datent du XVIIème siècle en Europe. Ces première machines, en fait des instruments imitant l'appareil vocal humain à l'aide de technologies dérivées de la facture d'orgue, ont été accompagnées d'une réflexion philosophique sur la parole, le language, et les automates [SER95]. Les moyens technologiques et les connaissances scientifiques contemporains. nous placent ils dans une situation épistémologique radicalement différente ? Ce n'est pas certain: ainsi Descartes envisageait tout-à-fait qu'une machine puisse imiter de la parole, certains oiseaux parleurs le font bien eux aussi. Cependant, il posait 3 conditions pour un langage (et donc une parole, forme sonore du langage) véritablement humain:

1. La première condition est que le langage doit être indépendant de son support: pour la parole, cela implique qu'elle puisse être multimodale. De fait, bien que cela n'apparaisse pas dans ce volume, de nombreuses recherches s'appliquent à la synthèse multimodale: visages parlants, langue des signes, génération de textes.

2. La seconde condition est celle de la pertinence de la parole en situation. Rapportée à la synthèse, il s'agit de la synthèse "située": la ligne de démarcation entre véritable parole et reproduction mécanique, plus ou moins sophistiquée, de la parole passe par cette possibilité de réponse intelligente, c'est-à-dire adaptée, expressive, voire émotive, modulée en fonction de la situation de communication. Ici encore, de nombreuses recherches, peu ou pas représentées dans ce volume, se développent autour des aspects d'adaptation et d'expression. Ces recherches sont à la fois dans le domaine de la phonétique au sens large (psychophonétique, traitement du signal vocal, perception ...) et dans celui du TAL, particulièrement de la compréhension des textes.

3. La troisième condition est la libre production des énoncés, en dehors de toute réponse automatique à un stimuli ou à une "passion". Des éléments dans ce sens se rencontreront peut-être dans les futurs "avatars" que nous promet la réalité virtuelle et autres "animats" de la "vie artificielle". Pour ce qui est du TAL, il s'agit bien sur de la génération automatique des textes.

Cependant, même si la parole de synthèse ne peut guère franchir pour le moment les limites que les philosophes lui ont assignées il y a plus de 3 siècles et demi, elle

reste un champ de recherche particulièrement fécond pour le traitement automatique des langues, la linguistique, la phonétique, l'acoustique, et elle n'a pas fini de nous poser problème et de nous aider à comprendre le langage.

En plus de ce rôle fondamental, les lecteurs/auditeurs constateront que les systèmes actuels peuvent déjà se préter à de nombreuses applications techniques. On peut aujourd'hui faire prononcer automatiquement n'importe quel texte, dans moins d'une vingtaine de langues, avec une qualité qui commence à s'approcher, sans l'atteindre encore, de la qualité d'une parole enregistrée et codée, dite par un locuteur virtuel comprenant peu le texte, et ayant peu de moyens expressifs. En ce sens la synthèse à partir du texte occupe une position médiane entre l'enregistrement/restitution de message, très peu adaptable au contexte, et la synthèse "située", qui serait capable non seulement de prononcer n'importe quel texte, mais aussi de le faire avec les nuances réellement appropriées, voire de le générer à propos et spontanément.

Ce numéro thématique ne peut prétendre couvrir toutes les recherches sur la synthèse de la parole, et nous devons en conclusion signaler les manques les plus importants. Tout d'abord, on ne trouvera pas ici d'articles sur les aspects d'acoustique ou de traitement du signal. Pourtant, cette partie des recherches sur la synthèse est très active: modèles articulatoires, modèles de glotte, méthodes d'analyse-synthèse et modification du signal etc. Pas d'articles non plus sur les aspects visuels de la synthèse, bien que la synthèse de visages parlants ou de langues des signes représente également un domaine important de recherche. Certaines recherches plus récentes, sur la synthèse des attitudes ou des émotions, sur l'utilisation de grandes bases de données de parole, sur la synthèse de textes enrichis de marques sémantiques ou pragmatiques, sur la synthèse de voix chantée, ne sont pas non plus représentées ici. Le domaine de l'évaluation n'est pas traité directement, bien qu'il soit abordé au cours de certains articles. Le lecteur trouvera donc un utile complément sur l'actualité de la recherche dans le domaine dans la série de conférences organisées périodiquement sur la synthèse de la parole [BAI92, VAN96]. Enfin, un disque compact contenant des exemples de 25 systèmes différents de synthèse en français accompagne ce volume. Ces exemples, de 1968 à nos jours, ont été réunis par Christophe d'Alessandro, et sont commentés dans une article final, qui invite le lecteur/auditeur à un promenade sonore dans le monde de la synthèse de la parole en français.

4. Bibliographie

[ALL87] Allen J., Hunnicutt M.S., Klatt D., "From text to speech: the MITalk system" Cambridge Univ. Press, 1987.

[BAI92] Bailly G., Benoît C. (Eds.) *TALKING MACHINES: Theories, Models, and Designs*. North Holland, 1992.

[BOI00] Boite R., Bourlard H., Dutoit T., Hancq J., Leich H., Traitement de la parole. Presse polytechniques et universitaires romandes, 2000.

[CAL89] Calliope. La parole et son traitement automatique. Masson, Paris, 1989.

[COL97] Cole R.A., Mariani J., Uskoreit H., Zaenen A., and Zue V., (Eds), Survey of the state of the art in human language technology, Cambridge University Press, and Giardini Editori e stampatori, 1997.

[DUT97] Dutoit T. An Introduction to Text-To-Speech Synthesis. Kluwer Academic Publishers, 1997.

[FLA73] Flanagan J.L. , Rabiner L.R., (Eds.) *Speech Syntheses*, Pennsylvania: Dowden, Hutchinson & Ross, 1973.

[MEL96] Méloni H., (Ed) Fondements et perspectives en traitement automatique de la parole, AUPELF-UREF, 1996.

[LIE77] Liénard J.S. Ls processus de la communication parlée: introduction à l'analyse et à la synthèse de la parole. Masson, 1977.

[SER95] Séris, J.P., Langages et machines à l'âge classique. Hachette, 1995.

[SPR98] Sproat R. (Ed), "Multilingual Text-To-Speech Synthesis: the Bell Labs Approach", Kluwer Academic Publishers, 1998.

[VAN96] Van Santen R., Sproat R., Hirschberg J., and Olive J. (Eds), Progress in Speech Synthesis. Springer Verlag, 1996.

Remerciements

Nous tenons à remercier les collègues dont la liste suit pour la qualité de leurs rapports, contributions précieuses à la sélection et à l'amélioration des articles publiés ici. Merci également aux collègues qui ont bien voulu nous procurer des exemples sonores de leurs systèmes de synthèse à partir du texte. En plus des auteurs des articles de ce volume, il s'agit de Gérard Bailly, Filip Deprez, Martine Garnier-Rizet, Eric Keller, Jean-Sylvain Liénard, Douglas O'Shaughnessy, Romain Prudon, Xavier Rodet, Nicklas Sajanti, Daniel Teil.

Rapporteurs externes

Véronique Aubergé	Institut de la Communication Parlée, Grenoble
Gérard Bailly	Institut de la Communication Parlée, Grenoble
Anelies Brafford	LIMSI-CNRS, Orsay
Francoise Emerard	France Télécom R&D , Lannion
Isabelle Guaitella	Laboratoire Parole et Langage Univ. de Provence
Jean-Sylvain Liénard	LIMSI-CNRS, Orsay
Thierry Moudenc	France Télécom R&D, Lannion
Patrick Paroubek	LIMSI-CNRS, Orsay
Janet B. Pierrehumbert	Dpt of Linguistics, Northwestern Univ., Evanston, USA
Gordon Ramsay	Institut de Phonétique, Univ. Libre de Bruxelles, Belgique
Jacqueline Vaissière	ILPGA, Université de Paris III

Rapporteurs du comité de rédaction

Christophe d'Alessandro	LIMSI-CNRS, Orsay
Philippe Blache	Laboratoire Parole et Langage Univ. de Provence
Danièle Clément	Bergische Univ. Gesamthochshule, Wuppertal, Allemagne
Anne Condamines	ERSS-CNRS, Université Toulouse Le Mirail
Marc El-Bèze	LIA, Université d'Avignon
Piet Mertens	Katholieke Universiteit, Leuven, Belgique
Evelyne Tzoukermann	Bell Labs., Lucent Technologies, Murray Hill, USA
Bernard Victorri	LTM CNRS Paris
Pierre Zweigenbaum	AP-HP, Université Paris VI