33 ans de synthèse de la parole à partir du texte: une promenade sonore (1968-2001)

Christophe d'Alessandro

33 ans de synthèse de la parole à partir du texte: une promenade sonore (1968-2001)

Christophe d'Alessandro

LIMSI-CNRS
Bâtiment 508 - Université Paris XI - BP 133 - F91403 Orsay Cédex
cda@limsi.fr

RÉSUMÉ. *Cet article présente un disque compact de 69 exemples sonores de synthèse de parole. Des exemples de 25 systèmes de synthèse automatique à partir du texte, principalement en français, sont décrits, avec 54 voix différentes. Dans une première partie, l'article présente des systèmes anciens (de 1968 à 1992). La seconde partie contient des exemples liés aux articles de ce numéro. La troisième partie contient des exemples d'autres systèmes contemporains. Ensuite, des parcours d'écoute sont proposés au lecteur/auditeur, afin d'apprécier divers aspects de la synthèse de parole: par types de synthétiseurs, sur le calcul de la prosodie, par type d'unités de synthèse. Enfin, un texte commun a été prononcé par 20 voix de synthèse différentes.*

ABSTRACT. *This article presents a compact disk containing 69 text-to-speech (TTS) synthesis sound examples. Examples from 25 automatic TTS systems, mainly in French are described, featuring 54 different voices. In a first part, ancient systems are presented (from 1968 to 1992). The second part describes sound examples linked to the papers of the present volume. The third part describes sound examples produced by other contemporary systems. Then, possible ways for listening to the examples are proposed to the reader/listener. This may be helpful for paying attention to specific aspects of TTS: e.g. synthesizer types, synthesis units, prosodic synthesis, regional accents. Finally, a common paragraph has been synthesized by 20 different voices.*

MOTS-CLÉS : *synthèse de la parole à partir du texte, exemples sonores, histoire de la synthèse.*

KEYWORDS: *text-to-speech synthesis, sound examples, history of synthesis*

1. Introduction

Le but de la synthèse de la parole à partir du texte est de produire automatiquement de la parole à partir d'un texte quelconque, c'est-à-dire de produire du son. Il nous a donc semblé indispensable de faire entendre le résultat sonore de ces recherches en synthèse, et ce depuis les premiers travaux jusqu'aux systèmes les plus récents, pour la langue française. Un disque compact d'exemples sonores est donc joint au volume. Le propos de cet article est de présenter et de commenter ces exemples, tout en indiquant les références bibliographiques correspondantes.

C'est à notre connaissance la première fois que des exemples sonores sont ainsi rassemblés de façon systématique et publiés pour le français. Une revue des systèmes de synthèse à partir du texte en anglais a été publiée, avec un disque souple 33 tours d'exemples sonores dans [KLA87]. On trouvera bien-sûr facilement de nombreux exemples d'autres langues en recherchant sur la Toile, et en particulier à l'adresse [MOL0101], probablement la plus complète à ce jour, qui présente des exemples sonores pour 25 langues.

Les exemples de ce disque sont tous strictement des exemples de synthèse à partir du texte, c'est-à-dire de synthèse automatique et à vocabulaire illimité. Cela exclut donc à la fois les exemples de synthèse non-automatique, comme la copie de parole par un synthétiseur, et les exemples de synthèse automatique avec un vocabulaire limité, comme par exemple les systèmes d'annonce du type horloge parlante ou information ferroviaire.

Le disque est constitué de 3 parties. La première partie rassemble des exemples de synthèse produits par des systèmes anciens, pour la plupart désormais disparus. Presque tous les systèmes de synthèse automatiques anciens dont l'auteur a eu connaissance sont représentés. Les systèmes qui ont existé mais dont il n'a pas été possible de retrouver d'exemples sonores, sont les suivants: le système de synthèse par règles de l'ICP, décrit dans [BAI88] (le système actuel de ce laboratoire figure ici), les premiers systèmes par diphones de la société Electrel (dont figure ici le système KALI), le système Mockingboard commercialisé dans les années 80 par BIP [MAC85], qui était relié à un ordinateur Apple II, le prototype de synthèse par diphones développé par Telic-Alcatel dans les années 80 autour d'un microprocesseur Nec7720 [MAC85], le système par diphones ICO85 de Vecsys [MAC85]. La seconde partie est directement liée aux autres articles de ce volume. En effet, les auteurs ont été invités à accompagner leurs textes d'exemples sonores de synthèse, et ils ont largement répondu à cette invitation. La troisième partie du disque contient des exemples de systèmes de synthèse contemporains, qui ne correspondent pas à un article du volume. Ici encore, l'auteur a sollicité des exemples de tous les systèmes actuels parlant français dont il a pu avoir connaissance, en ce début du 21ème siècle (il manque néanmoins certainement des systèmes, comme celui de la récente société Loquendo [LOQ01] par exemple , apparu trop tard pour figurer dans le disque ...).

Le disque présente 69 exemples sonores, correspondant à plus de 54 voix différentes, produites par 25 systèmes différents, issus de 16 laboratoires (certains labora-

toires relativement anciens, ont développé plusieurs systèmes de synthèse au cours des années). Certains systèmes ont fait l'objet de variations ou d'amélioration successives. C'est pourquoi il y a plus de 54 "voix" différentes: une voix est différente lorsqu'un systèmes possède plusieurs jeux de règles prosodiques, plusieurs possibilités de synthèse acoustique, plusieurs bases de voix, voire plusieurs langues. Ces variations sont indiquées dans les tableaux 1, 2, 3. Nous espérons que ce travail de récolement sera utile au moins pour les points suivants:

Illustration sonore des articles: c'est le but premier de ces exemples sonores. Ainsi le lecteur pourra apprécier non seulement la subtilité théorique ou la qualité technique des recherches exposées ici, mais aussi en écouter le résultat.

Comparaison des approches: le matériel rassemblé ici permet de comparer différentes approches pour différentes composantes des systèmes de synthèse. On trouvera en effet à la fois plusieurs systèmes utilisant des techniques de synthèse différentes (synthèse à formant, synthèse par diphones, synthèse par sélection et concaténation), plusieurs systèmes utilisant la même technique de synthèse (par exemple synthèse par diphones, avec des techniques de traitement acoustique différentes), voire plusieurs systèmes utilisant la même technique de synthèse et la même voix, mais des approches différentes pour la prosodie et la phonétisation.

Histoire des sciences et techniques: il reste finalement relativement peu de traces sonores des assez nombreux travaux menés sur la synthèse. Que de temps passé, de moyens mis en oeuvre pour la recherche, de pages de texte produites (sous forme d'articles, communications, thèses etc.) pour ne conserver que quelques minutes, parfois quelques secondes de son ! Les papiers restent, mais que les systèmes informatiques disparaissent, et avec eux la possibilité de produire de la parole de synthèse. Ce qui n'a pas été enregistré est perdu. Et ce qui a été enregistré il y a déjà 20 ou 30 ans commence à se dégrader.

Afin de pouvoir comparer également les systèmes entre eux à l'aide d'un matériel commun, les auteurs ont bien voulu synthétiser le court texte suivant:

"Synthèse de la parole à partir du texte. Le nouveau millénaire commence pour la revue Traitement Automatique des Langues par un numéro thématique consacré à la synthèse de la parole. Vous allez ainsi pouvoir apprécier la voix de plusieurs systèmes de synthèse automatique en français. Une façon directe d'appréhender les progrès et les limites du traitement automatique des langues."

L'article présente d'abord les trois sections du disque (systèmes anciens, exemples associés aux articles de ce numéro, autres systèmes contemporains). La dernière partie propose des itinéraires d'écoute dans le disque, invitant à la promenade dans divers paysages sonores, comme ceux des méthodes de synthèse, de la synthèse de la prosodie, des accents régionaux, de la synthèse expressive. Une courte conclusion suit.

No	Système	origine	Date	Synthèse	H/F	kHz
4	syst. 1	LAM/LIMSI (F)	1968	dip./additive	H	4,4
6	syst. 2	LIMSI (F)	1974	dip./additive	H	4,4
8	syst. 3	CNET (F)	1976	dip./voc. canaux	F	4,2
10	syst. 4(1)	INRS (CA)	1977	règles/synth. Klatt	H	5
11	syst. 4(2)	INRS (CA)	1980	règles/synth. Klatt	H	5
13	syst. 5	CEA (F)	1977	règles/synth. FOF	H	5
15	syst. 6	INFOVOX	1982	règles/synth. KTH	H	5
17	syst. 4(3)	INRS (CA)	1984	règles/synth. Klatt	H	5
19	syst. 6(1)	CNET (F)	1985	dip./LPC	H	5
20	syst. 6(2)	CNET (F)	1989	dip./LPC	H	5
21	syst. 6(3)	CNET (F)	1989	dip./LPC	H	5
23	syst. 7(1)	CNET (F)	1989	dip./PSOLA	H	8
24	syst. 7(2)	CNET (F)	1989	dip./PSOLA	H	8
26	syst. 8(1)	INFOVOX (S)	1991	règles/synth. KTH	H	8
28	syst. 9	LIMSI (F)	1992	règles/synth. Polyglot	H	8

Tableau 1. *Première partie: exemples sonores de systèmes anciens*

2. Exemples sonores de systèmes anciens

2.1. *LAM/LIMSI- Icophone III 1968 (4)*

Le premier système de synthèse automatique à partir du texte en français a été conçu et réalisé au Laboratoire d'Acoustique Musicale (LAM) de Paris VI [LAM01] et au Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI-CNRS) [LIM01] (avant 1972 ce laboratoire portait le nom de Centre de Calcul Analogique du CNRS). Ce système a été opérationnel en décembre 1968 [LCL68], à la suite du dépot d'un brevet [LEI68]. Il s'agit d'un système hybride analogique et numérique, nommé Icophone III, utilisant un ordinateur IBM 1130. La partie électronique et les concepts phonético-acoustiques ont été développés au LAM, alors que les programmes de transcription phonétique et la partie informatique ont été développés au LIMSI. La technique de synthèse est la synthèse par diphonèmes (ou diphones), utilisant un banc d'oscillateurs fixes.

L'ordinateur effectue la transcription phonétique (le premier programme écrit pour le français [LIE70]), la recherche des diphonèmes en mémoire, le séquencement des diphonèmes, puis leur sortie sur une interface numérique. De l'autre côté de l'interface se trouve le synthétiseur acoustique : un banc de 44 oscillateurs (tout les 100 Hz de 100 à 4400 Hz) analogiques pilotés numériquement par l'ordinateur. Les diphonèmes sont stockés sous forme de spectrogrammes stylisés, qui pilotent en 0 ou 1 l'activation des oscillateurs. Il n'y a donc ni modulation d'intensité, ni variation de fréquence fondamentale. La **piste 4** donne un exemple de la parole ainsi produite. Il s'agit donc de la première voix de synthèse automatique en français. Au niveau international,

c'est également en 1968 que sont publiés les travaux sur les deux premiers systèmes automatiques de synthèse pour l'anglais: le système américain par diphones de Dixon et Maxey, démontré en 67 et publié en 1968, et le système par règles japonais de Umeda, Matsui, Teranishi et Suzuki, publié aussi en 1968 [KLA87].

L'Icophone III possède d'emblée tous les traits d'un système de synthèse automatique à partir du texte. Cependant, un défaut majeur de ce système est l'absence de prosodie: la fréquence fondamentale est fixée à 100hz.

2.2. *LIMSI- Icophone V, 1974 (6)*

L'Icophone IV de 1971 est un nouvel appareil, qui permet de modifier la fréquence fondamentale des oscillateurs par programme ou manuellement. 16 valeurs de fréquence autour de la fréquence nominale sont possibles. Il faut noter que tous les oscillateurs restent en relation harmonique (ils sont tous multiples d'une même fréquence fondamentale), donc modifier la fréquence fodamentale va affecter les formants. Les travaux sur l'intonation aboutissent avec l'icophone V [TEI74, TEI75], qui est la première unité de réponse vocale autonome. A partir de 1974-1975 [CHO75, LIE77] la synthèse à partir du texte de l'Icophone V est dotée d'un algorithme de prosodie automatique. La longueur des mots, et une courte liste de mots outils permettent de découper les textes en groupes intonatifs. Les paramètres prosodiques sont ensuite calculés par combinaison de schémas intonatifs élémentaires. Un exemple de la synthèse ainsi obtenue est donné sur la **piste 6**. C'est un enregistrement tardif (de 1979) du système. La série des Icophones s'achève en 1979 avec l'Icolog [AST79], ou icophone logiciel, tout numérique, qui sera commercialisé ultérieurement par la société VECSYS, avec une carte spécialisée à microprocesseur: la carte Ico85.

2.3. *CNET 1976 (8)*

Le Centre National d'Etude des Télecommunicationq (CNET) de Lannion a consacré très tôt des efforts importants à la synthèse de la parole, pour des applications téléphoniques. Ses activités continuent aujourd'hui au sein de la compagnie France Télécom Recherche & Développement (FTR&D).

L'approche choisie au début par le CNET est également une approche par diphones. Le problème de la synthèse de la prosodie est posé d'emblée, et la technique de synthèse choisie a été le vocodeur à canaux. Avec cette technique, les paramètres prosodiques sont facilement modifiés, avec une qualité supérieure à celle, par exemple, d'un banc d'oscillateurs. Le vocodeur du CNET comprend 14 canaux, couvrant une bande de 250 à 4200 Hz. La fréquence fondamentale est codée sur 256 valeurs (8bits). Le débit total est de 4800 bits/seconde.

Les travaux décrits dans [LAR76, EME77] s'attachent à l'élaboration de la base de diphones, et surtout au calcul de la prosodie. Dès 1976 sont utilisées des techniques d'analyse syntaxique superficielle et partielle pour le marquage prosodique,

telles qu'on les retrouve dans beaucoup de systèmes actuels. A la même époque, un programme de conversion graphème-phonème est développé [DIV77]. La **piste 8** est un exemple du résultat produit en 1976 par le système du CNET.

2.4. *CNET-LPC 1985 (19, 20, 21)*

Le premier système du CNET a été suivi de plusieurs autres [BAR87], avec des techniques de synthèse différentes. Le vocodeur à canaux a été remplacé par une technique de synthèse de meilleure qualité: la prédiction linéaire (LPC).

Le système de synthèse ainsi réalisé a donné lieu à des développements commerciaux importants, sous forme de logiciels et de cartes informatiques spécialisées. Plusieurs compagnies ont utilisé les recherches du CNET pour développer leurs produits de synthèse. La qualité sonore de ces systèmes peut être appréciée sur les **pistes 19, 20, 21**, la **piste 19** date de 1985, la **piste 20** est due au logiciel MULTIVOX, enregistré en 1989, et la **pistes 21** à la carte TELEVOX, enregistrée en 1989.

2.5. *CNET PSOLA 1989 (23, 24)*

Les années 1980 ont vu un développement considérable des moyens informatiques. On a donc envisagé de stocker directement des bases de diphones sur les ordinateurs, avec des techniques de codage de haute qualité. La qualité de parole s'en trouve nettement améliorée, et la bande passante des systèmes peut s'élargir, pour passer à 8 kHz [STE85].

Pour la synthèse du français, puis de nombreuses autres langues, l'année 1988 est une année importante, puisque une nouvelle technique, la synthèse PSOLA, est présentée [HAM88, HAM89]. Il s'agit de modifier directement dans le domaine temporel une base de diphones. La technique est à la fois économique en temps de calcul, et d'une qualité remarquable pour les modifications prosodiques. Les **pistes 23 et 24** sont les premiers exemples, de 1989, de synthèse PSOLA. Cette méthode a démontré que l'on pouvait modifier la prosodie avec une qualité quasi-transparente, sur de la parole numérisée, avec des techniques conceptuellement et algorithmiquement très simples. Avec la méthode PSOLA, il semble que la synthèse par diphones ait atteint, vers 1989-1990, son plein potentiel de qualité [MOU90].

2.6. *INRS 1977-1984 (10, 11)*

Le système développé à l'Institut National de la Recherche Scientifique - Télécommunication [INR01] , au Québec, est un système de synthèse par règles. Contrairement aux systèmes précédents, il utilise un synthétiseur à formant, c'est-à-dire un modèle de l'appareil vocal humain. La synthèse est réalisée par règles: chaque phonème et les transitions entre phonèmes sont modélisés par un ensemble de valeurs acoustiques, et

par des règles de coarticulation [OSH81, OSH82]. Ce synthétiseur est donc beaucoup plus riche du point de vue acoustico-phonétique que les synthétiseurs par diphones. De même, la prosodie est calculée par règles [OSH89].

Ce système issu d'un laboratoire Canadien synthétise du français Québécois. De fait, la prosodie est particulièrement soignée. Elle se base sur la ponctuation et les catégories syntaxiques des mots. Une autre particularité de ce système est d'être dans la lignée des recherches effectuées au M.I.T. de Boston sur la synthèse par règles de l'anglais, et d'utiliser le synthétiseur à formants de Klatt [KLA87]. Il s'agit sans doute du meilleur système de synthèse par règles pour le français au milieu des années 1980. Les **pistes 10 et 11** sont des exemples produits à cette époque.

Dans la **piste 17**, l'accent québécois du synthétiseur est particulièrement affirmé. C'est un des rares exemples de synthèse d'une variété de français qui s'éloigne sensiblement du français parlé en Ile-de-France. Cet exemple démontre d'une part la souplesse de la synthèse par règles, et d'autre part la qualité des règles prosodiques employées.

2.7. *CEA - Sara 1979 (13)*

Le système Sara a été développé au CEA dans les années 70. C'est un système complet de synthèse par règles. Le système fonctionne sur un mini-ordinateur Intertechnique Multi-20, sous forme de logiciel [ROD77].

Une originalité de ce système est le type de synthétiseur à formants employé. Il s'agit d'une méthode de synthèse à formant en parallèle dans le domaine temporel, qui a été développée à cette occasion [ROD80]. Le signal est calculé pour chaque période de voisement, par superposition de formes d'ondes formantiques, dans le domaine amplitude-temps. La **piste 17** est un exemple du système Sara de 1977.

2.8. *KTH - Infovox 1982-1991 (15, 26)*

Les systèmes commercialisés par Infovox [INF01] sont issus des recherches menées au KTH de Stockholm [CAR76]. Ce système s'appuie sur un langage de programmation spécifiquement conçu pour mettre en oeuvre le formalisme de la synthèse par règles phonologiques. Chaque phonème est représenté par un jeu de traits phonologiques, que le synthétiseur doit utiliser pour calculer les valeurs acoustiques. Le synthétiseur lui-même est un synthétiseur à formants, dérivant directement de la théorie acoustique de production de la parole.

Une des ambitions de ces recherches était de développer un système de synthèse universel, indépendant de la langue. De fait le système SA-101 a été le premier système multilingue utilisant le même formalisme pour toutes les langues, en adaptant seulement les règles spécifiques pour chaque langue. La **piste 15** est une exemple du système SA-101, de 1982, où on entendra du français, puis la même voix dans

plusieurs autres langues [CAR82]. Pour l'auditeur français, ce système possède un fort accent nordique ou germanique, qui n'était probablement pas recherché par ses auteurs, et qui trahit son origine suédoise. C'est la contrepartie d'un résultat remarquable: toutes les langues sont synthétisées avec le même système, la même voix, en changeant seulement les règles phonologiques et prosodiques.

La **piste 26** est une version plus récente du synthétiseur par règles d'Infovox [CAR90]. De nouvelles langues ont été ajoutées, ainsi que de nouvelles règles pour la source glottique et la prosodie. Cet exemple date de 1991.

2.9. *LIMSI - Polyglot 1992 (28)*

Le système de synthèse par règles du LIMSI [LIM01] a été développé dans le cadre du projet Européen Polyglot. Ce projet visait en particulier à réaliser un synthétiseur multilingue, sur une plate forme matérielle commune, et avec un environnement logiciel commun. Le synthétiseur à formants développé est très proche de celui de Klatt. Les traits phonétiques, valeurs cibles, règles de coarticulations et de nombreux exemples se trouvent dans [GAR91, GAR93]. La **piste 28** est un exemple de la synthèse obtenue par le synthétiseur Polyglot en français.

3. Exemples sonores d'articles de ce volume

3.1. *Nicolas Torzec, Thierry Moudenc, Françoise Emerard "Prétraitements et analyses linguistiques dans le système de synthèse de parole à partir du texte CVOX: application à la vocalisation automatique d'Emails" (31, 32, 33, 34, 35, 36, 37, 38, 39, 40)*

Les travaux rapportés dans cet article sont menés chez (FTR&D) [FTR01]. Ils se situent dans la continuité des travaux du CNET-Lannion, dont l'évolution est retracée plus haut. Ainsi, les **pistes 31, 32, 36, 37, 38 et 39** sont des exemples de la version actuelle du synthétiseur PSOLA, en français **31 et 32**, anglais **36**, espagnol **37**, allemand **38** et russe **39**. Contrairement aux systèmes multilingues de synthèse par règles, chaque langue différente possède une voix différente, puisque le signal est construit à l'aide d'échantillons de parole réelle, et non de règles abstraites. Les **pistes 31 et 32** synthétisent le texte de référerence, avec une voix de femme et d'homme respectivement.

Les **pistes 33 et 34** font entendre des exemples de services commerciaux: service vocal d'annuaire inverse **33** et vocalisation de méls **34**. Les difficiles problèmes de traitement linguistique posés par ces applications sont discutés dans l'article.

Les **pistes 35 et 40** concernent deux des directions de recherche poursuivies actuellement chez FTR&D. Pour la **piste 35**, il s'agit, en utilisant le synthétiseur PSOLA par diphones, de rechercher des règles prosodiques expressives. Pour la **piste 40**, la

No	Système	Origine	Synthé.	H/F	kHz
31	syst. 10(1)	FTR&D (F)	PSOLA	F	8
32	syst. 10(2)	FTR&D (F)	PSOLA	H	8
33	syst. 10(3)	FTR&D (F)	PSOLA	H	8
34	syst. 10(4)	FTR&D (F)	PSOLA	F	8
35	syst. 10(5)	FTR&D (F)	PSOLA	H	8
36	syst. 10(6)	FTR&D (F)	PSOLA	H	8
37	syst. 10(7)	FTR&D (F)	PSOLA	H	8
38	syst. 10(8)	FTR&D (F)	PSOLA	H	8
39	syst. 10(9)	FTR&D (F)	PSOLA	H	8
40	syst. 11	FTR&D (F)	PSOLA/UNU	F	8
42	syst. 12	LPL (F)	MBROLA	H	8
43	syst. 12	LPL (F)	MBROLA	H	8
44	syst. 12	LPL (F)	MBROLA	H	8
46	syst. 13(1)	LIMSI (F)	PSOLA	H	8
47	syst. 13(2)	LIMSI (F)	MBROLA	H	8
48	syst. 13(1)	LIMSI (F)	PSOLA	H	8
49	syst. 13(2)	LIMSI (F)	MBROLA	H	8
50	syst. 13(3)	LIMSI (F)	MBROLA	H	8
51	syst. 13(4)	LIMSI (F)	PSOLA	H	8
53	syst. 14(1)	KUL/LATL (B/CH)	MBROLA	H	8
54	syst. 14(2)	KUL/LATL (B/CH)	MBROLA	H	8
55	syst. 14(3)	KUL/LATL (B/CH)	MBROLA	H	8
56	syst. 14(3)	KUL/LATL (B/CH)	MBROLA	H	8
57	syst. 14(1)	KUL/LATL (B/CH)	MBROLA	H	8
59	syst. 15(1)	ELECTREL (F)	PSOLA	F	11
60	syst. 15(2)	ELECTREL (F)	PSOLA	H	11
61	syst. 15(1)	ELECTREL (F)	PSOLA	F	11
62	syst. 15(1/2)	ELECTREL (F)	PSOLA	H/F	11
64	syst. 16(1)	ELAN (F)	PSOLA	H	8
65	syst. 16(2)	ELAN (F)	PSOLA	H/F	8
66	syst. 16(3)	ELAN (F)	PSOLA	H/F	8
68	syst. 17	LUCENT (USA)	GELP	H	5,5
70	syst. 18(1)	TCTS (B)	MBROLA	H	8
71	syst. 18(2)	TCTS (B)	MBROLA	F	8
72	syst. 18(1)	TCTS (B)	MBROLA	H	8
73	syst. 18(1)	TCTS (B)	MBROLA	H	8
74	syst. 18(2)	TCTS (B)	MBROLA	F	8
75	syst. 18(2)	TCTS (B)	MBROLA	F	8

Tableau 2. *Deuxième partie: exemples sonores de systèmes décrits dans ce volume*

synthèse utilise non plus seulement des diphones, mais des unités de synthèse plus longues, par exemple ici pour les chiffres, lorsque c'est possible.

3.2. *Albert Di Cristo, Philippe Di Cristo "SYNTAIX: une approche métrique-autosegmentale de la prosodie" (42, 43, 44)*

Les travaux rapportés dans cet article ont été menés au Laboratoire Parole et Langage d'Aix-en-Provence [LPL01]. Les **pistes 42 et 43** contiennent des textes de présentation du système, et la **piste 44** synthétise le texte de référence. Les travaux des auteurs ont porté sur la prosodie et l'analyse linguistique du texte, et le synthétiseur utilisé est celui mis à disposition dans le projet MBROLA [MBR01].

3.3. *Philippe Boula de Mareüil, Christophe d'Alessandro,Frédéric Beaugendre, Anne Lacheret-Dujour "Une grammaire en tronçons appliquée à la génération de la prosodie" (46, 47, 48, 49, 50, 51)*

Un nouveau système de synthèse, basé sur un synthétiseur par diphones a été développé au LIMSI [LIM01] dans les années 1990. Cet article discute de l'analyse du texte et du traitement de la prosodie dans ce système. La partie acoustique utilise soit un synthétiseur de type PSOLA développée au LIMSI, soit le synthétiseur mis à disposition dans le projet MBROLA [MBR01].

Les **pistes 46 et 47** contiennent un texte de présentation, avec les mêmes traitements sauf la synthèse acoustique, qui est réalisée en PSOLA **46** ou MBROLA **47**. Les **pistes 48, 49 et 50** contiennent le texte de référence, synthétisé à travers trois synthétiseurs: en PSOLA avec le locuteur FB du LIMSI (**piste 48**), en MBROLA (**piste 49**) avec le locuteur FB du LIMSI, en MBROLA avec la voix française standard de MBROLA (**piste 50**).

Les idées développées dans cet article pour l'analyse et la prosodie peuvent s'appliquer à d'autres langues, ce que l'on peut entendre sur la **piste 51**, qui est le synthétiseur du LIMSI en espagnol.

3.4. *Piet Mertens, Jean-Philippe Goldman, Eric Wehrli, et Arnaud Gaudinat "La synthèse de l'intonation à partir de structures syntaxiques riches" (53, 54, 55, 56, 57)*

Cet article traite de la prosodie du français, et en particulier de l'utilisation poussée de la syntaxe pour la synthèse prosodique. L'aspect prosodique est dans la lignée des travaux développés au département de linguistique de l'université Catholique de Leuven (système Mingus)[MIN01], alors que la partie syntaxique provient du Laboratoire d'Analyse et de Technologie du Langage, de l'université de Genève (système FIPS) [LAT01].

La **piste 53** contient une série de phrases affirmatives isolées, avec des structures syntaxiques variées. . La synthèse est produite avec l'analyse syntaxique de FIPS et le système prosodique Mingus. La **piste 54** donne des exemples, pour la même phrase, de plusieurs réalisations prosodiques acceptables différentes, avec et sans l'aide d'annotations manuelles du texte. La **piste 55** contient une série de phrases isolées, affirmatives et interrogatives. Elles sont générées par le système Mingus-TTS, mais sans l'analyse de FIPS. La **piste 56** démontre l'effet de changements dans les paramètres acoustiques du modèle mélodique utilisé pour synthétiser l'intonation. Les hauteurs tonales et les valeurs extrêmes haute et basse de l'intonation sont variées pour une même phrase.

L'exemple de la **piste 57** synthétise le texte de référence. Le synthétiseur utilisé dans tous ces exemples est celui mis à disposition dans le projet MBROLA [MBR01].

3.5. *Michel Morel, Anne Lacheret-Dujour ""Kali", synthèse vocale à partir du texte" (59, 60, 61, 62)*

La société Electrel commercialise depuis plusieurs années des systèmes de synthèse à partir du texte [ELE01]. Ces systèmes visent en particulier des applications de lecture automatique pour l'aide aux déficients visuels. Le système KALI décrit ici est issu d'une collaboration entre Electrel, le laboratoire CRISCO de l'université de Caen [CRI01], le laboratoire GREYC de l'université de CAEN [GRE01] et le club MICROSON, qui regroupe des utilisateurs de systèmes de synthèse à partir du texte.

Le système utilise un synthétiseur de type PSOLA à base de diphones, et une analyse syntaxique partielle pour la prosodie. La **piste 59** contient le texte de présentation, avec une voix de femme. La **piste 60** est un texte de présentation, avec une voix d'homme. La **piste 61** est à nouveau la voix de femme, avec d'importantes différences de vitesse d'élocution, afin de démontrer les performances de KALI pour l'intelligibilité en lecture rapide. La **piste 61** est un autre texte de présentation en voix féminine qui s'achève par un duo des voix d'homme et de femme du système.

3.6. *Philippe Boula de Mareüil, Philippe Célérier, Thierry Cesse, Serge Fabre, Carine Jobin, Pierre-Yves Le Meur, David Obadia, Benoît soulage, Jacques Toen "ELAN Text-To-Speech: un système multilingue de synthèse de la parole à partir du texte" (64, 65, 66)*

La société ELAN consacre une partie de ses activités au développement et à la commercialisation de produits de synthèse de parole multilingues [ELA01]. Ces produits sont basés sur les algorithmes du CNET, et utilisent de la synthèse par diphones avec le synthétiseur PSOLA.

La **piste 64** contient le texte de référence avec une voix d'homme. La **piste 65** commence avec une phrase en anglais, puis donne des exemples en français contenant beaucoup de difficultés pour la normalisation du texte, puis continue sur de l'italien,

et s'achève avec du français prononcé avec un fort accent espagnol. La **piste 66** donne des exemples en espagnol.

3.7. *Evelyne Tzoukermann "La synthèse de la parole du Français: le système des Laboratoires Bell" (68)*

Les laboratoires Bell de Lucent Technologies ont développé un système de synthèse de la parole multilingue, sur une plate-forme logicielle commune [BEL01]. Parmi les langues traitées se trouve le français, et l'article détaille cette composante du système.

La technique de synthèse est basée sur des diphones, en utilisant la prédiction linéaire pour les traitements acoustiques. La prosodie est calculée par une analyse morpho-syntaxique partielle. La **piste 68** contient un texte de présentation de ce système.

3.8. *Nawfal Tounsi, Thierry Dutoit, Michel Bagein, Fabrice Malfrère, Alain Ruelle et Dominique Wynsberghe "Le projet EULER: vers une synthèse de parole générique et multilingue" (70, 71, 72, 73, 74, 75)*

Le projet MBROLA [MBR01] a eu un impact considérable sur les recherches en synthèse de la parole. Au départ MBROLA est un algorithme de modification des durées et de la fréquence fondamentale de la parole, qui permet la synthèse par diphones avec une qualité et un principe assez voisin de PSOLA. MBROLA est apparu quelques années après PSOLA, mais avec le projet de diffuser librement pour la recherche la partie acoustique d'un système de synthèse (à l'exclusion des applications commerciales et militaires), c'est-à-dire la base de diphones et le système de modification prosodique. Pour faire un système de synthèse à partir du texte il ne manque donc que la phonétisation automatique du texte et le calcul automatique de la prosodie.

Comme cette partie acoustique est longtemps restée une sorte de verrou pour les systèmes de synthèse, et comme la qualité de MBROLA est comparable à celle des meilleurs systèmes par diphones, la diffusion libre de ce système a été une ouverture véritable. Elle a permis à de nombreuses équipes, qui n'avaient pas de système complet, d'accéder assez rapidement à la synthèse à partir du texte, en ne se concentrant que sur des problèmes spécifiques, comme la prosodie. Ainsi, dans ce disque il n'y a pas moins de 7 laboratoires qui utilisent exclusivement ou partiellement le synthétiseur acoustique MBROLA, dont 5 qui n'avaient jamais développé de système de synthèse à partir du texte avant ce projet.

Dans cet article, les auteurs présentent la continuation du projet MBROLA, le projet EULER, qui vise maintenant à développer et à diffuser également des outils génériques et libres d'utilisation pour toutes les procédures de synthèse multilingue. Il faut noter que le projet MBROLA a également permis le développement de bases de

No	Système	Origine	Synthé.	H/F	BP
77	syst. 19(1)	ICP (F)	PSOLA	H	8 kHz
78	syst. 19(2)	ICP (F)	PSOLA	H	8 kHz
79	syst. 19(3)	ICP (F)	PSOLA	H	8 kHz
81	syst. 20	LATL (CH)	MBROLA	H	8 kHz
83	syst. 8(2)	INFOVOX (S)	FORMANT	H	8 kHz
84	syst. 21	INFOVOX (S)	MBROLA	H	8 kHz
86	syst. 22	LAIP (CH)	MBROLA	H	8 kHz
88	syst. 23	L&H (B)	GELP	F	5,5 kHz
89	syst. 24(1)	L&H (B)	UNU	F	5,5 kHz
90	syst. 24(2)	L&H (B)	UNU	F	11 kHz
92	syst. 25(1)	LIMSI (F)	UNU	H	8 kHz
93	syst. 25(2)	LIMSI (F)	UNU	F	8 kHz
94	syst. 25(3)	LIMSI (F)	UNU	H	8 kHz
95	syst. 25(1)	LIMSI (F)	UNU	H	8 kHz
96	syst. 25(2)	LIMSI (F)	UNU	F	8 kHz
97	syst. 25(3)	LIMSI (F)	UNU	H	8 kHz

Tableau 3. *Troisième partie: exemples sonores d'autres systèmes contemporains*

diphones et de systèmes de synthèse dans de nombreuses langues jusqu'alors dépourvues de synthétiseur.

Les exemples sonores présentent de la parole synthétisée par MBROLA, avec des procédures de traitement intégrées à EULER. Les **pistes 70 et 71** contiennent le texte de référence, avec une voix d'homme et de femme respectivement. Les **pistes 71 et 72** contiennent des textes de présentation, avec une voix d'homme, et les **pistes 74 et 75** contiennent les mêmes textes avec une voix de femme.

4. Autres systèmes contemporains

En dehors des articles publiés dans ce volume, il existe bien-sûr d'autres systèmes de synthèse à partir du texte en français. Nous avons donc proposé à tout ceux que nous connaissions de figurer également dans ce disque, et si possible de synthétiser le texte de référence. Tous les laboratoires contactés ont répondu positivement.

4.1. *ICP*

Le système actuel de l'Institut de la Communication Parlée de Grenoble (ICP) est basé sur la synthèse par diphones, avec la méthode PSOLA [ICP01, BAI92]. Une attention particulière est portée sur les problèmes de prosodie [BAI89], [BAR94]. En

effet la prosodie permet de varier les styles de parole en fonction du contexte ou du type d'énoncé à synthétiser.

La **piste 77** est un exemple de synthèse des attitudes. Il s'agit de rendre la synthèse plus expressive, en adaptant la prosodie aux aspects pragmatiques de la lecture. Ainsi, la prosodie d'une même phrase est variée automatiquement en fonction de marqueurs d'attitudes placés dans le texte [MOR01]. La **piste 78** est un autre exemple de synthèse de la prosodie dans un contexte pragmatique particulier. Il s'agit de synthèse automatique de prosodie pour des formules mathématiques [HOL00]. Dans ce cas particulier, on peut en effet déduire de l'énoncé une prosodie bien structurée. La **piste 79** contient le texte de référence.

4.2. *LATL-Fipsvox*

La **piste 81** est issue du système Fipsvox développé au Laboratoire d'Analyse et de Technologie du Langage, de l'université de Genève [GAU97, GAU98, LAT01]. Il contient le texte de référence. Le synthétiseur utilisé est celui mis à disposition dans le projet MBROLA [MBR01].

4.3. *TELIA-Infovox*

Après les exemples anciens du système multilingue Infovox [INF01], les versions plus récentes de ce système sont représentées par les **pistes 83 et 84**. Deux systèmes assez différents sont utilisés. La **piste 83** contient un texte de présentation, produit par un synthétiseur par règles à formant. Cet exemple est dans la continuité de l'exemple Infovox de 1991 dans la première partie du disque.

Au contraire, l'exemple de la **piste 84** utilise de la synthèse par diphones. Il s'agit du même texte de présentation, mais l'algorithme de synthèse est MBROLA [MBR01].

4.4. *LAIP*

Le Laboratoire d'Analyse Informatique de la Parole de l'université de Lausanne s'intéresse particulièrement à la synthèse des aspects rythmiques de la parole [LAI01] [KEL97] [ZEL98]. La **piste 86** contient le texte de référence, augmenté de quelques phrases de protestation. Le synthétiseur utilisé est celui mis à disposition dans le projet MBROLA [MBR01].

4.5. *L&H*

La compagnie Lernout & Hauspie développe depuis plusieurs années des systèmes de synthèse à partir du texte multilingues [LEH01]. Ici sont représentés des exemples pour le français.

La **piste 88** contient le texte de référence, synthétisé par diphones avec une technique de prédiction linéaire à excitation glottique. Cet exemple a été produit par le système TTS-3000 qui date de 1996 (logiciel de 1,5 Moctets).

Les **pistes 89 et 90** utilisent une technique de synthèse tout à différente. Il s'agit de synthèse à partir de gros corpus de parole enregistrée. L'unité de synthèse n'est plus strictement le diphone, mais un segment de longueur variable. La synthèse est obtenue par optimisation de la chaîne de segments sur toute la base de donnée [COO00, RUT00]. La **piste 89** contient le texte de référence, synthétisé par le système à base de corpus RealSpeak Compact, qui date de 2001 (logiciel de 16 Moctets). La **piste 90** contient le texte de référence, synthétisé par le système à base de corpus RealSpeak Classic qui date de 2000 (logiciel de 64 Moctets). La taille de ces systèmes est bien entendu plus importante que celle du système par diphones.

4.6. *LIMSI - SeLimsi*

Les derniers exemples de synthèse de ce disque correspondent au nouveau système de synthèse à partir de gros corpus développé au LIMSI [LIM01]. Ce système SeLimsi [PRU01] est basé sur la recherche d'une suite optimale de diphones dans une grosse base de donnée de parole. En particulier, on recherche des suites contigües de diphones, afin si possible de tirer le meilleur parti du contenu de la base. Ce système traite également la prosodie par sélection: il n'y a pas de traitement du signal pour la synthèse, si ce n'est un lissage des points de montage des différents segments. Les bases de données de parole sont étiquetées automatiquement par reconnaissance de la parole et analyse du signal. Ce type de système n'utilise donc plus du tout de règles déterministes de synthèse, même pas pour la prosodie. Tout le travail est dans l'analyse et l'étiquetage des bases de données, et dans l'optimisation de la sélection au moment de la synthèse.

Trois voix existent actuellement pour ce système. Les **pistes 92, 93 et 94** synthétisent un texte de présentation, avec deux voix d'homme et une voix de femme. Les **pistes 95, 96 et 97** synthétisent le texte de référence, avec les mêmes deux voix d'homme et la voix de femme.

5. Quelques itinéraires d'écoute

Le disque contient donc 69 exemples sonores. On peut l'écouter linéairement, du début à la fin, ou bien en recherchant un système particulier. Cependant, le matériel réuni ici est suffisament riche pour permettre des itinéraires d'écoute variés, instructifs

ou même distrayants. Quelques un de ces itinéraires invitant à la promenade sonore sont indiqués ici et seront facilement parcourus en programmant la séquence des pistes d'un lecteur de disques compacts.

5.1. *Par type de synthétiseurs et d'unités de synthèse*

Du point de vue acoustico-phonétique, il existe plusieurs familles de systèmes de synthèse. Une première façon de procéder est d'assembler des fragments de sons, pour le français généralement des diphones (c'est-à-dire un segment de son compris entre le milieu d'un premier phonème, et le milieu d'un second phonème). De nombreux systèmes par diphones ont été développés pour le français. Ce type de synthèse est relativement gourmand en mémoire (environ 1200 fragments de son à stocker, soit plusieurs minutes de parole), mais par contre il ne demande que des connaissances élémentaires en phonétique. La différence entre ces systèmes vient des deux aspects de la synthèse par diphone: la technique pour coder/modifier les diphones, et la base de diphone proprement dite. Sur ce disque sont représentés 5 types de techniques de synthèse par diphones (avec des variantes).

Diphones/Banc d'oscillateurs : sur la **piste 4**, les oscillateurs sont tous accordés, avec une fréquence fondamentale fixe de 100Hz. Sur la **piste 5**, il sont accordés sur une fréquence fondamentale variable.

Diphones/Vocodeur à canaux: sur la **piste 8**.

Diphones/Prédiction linéaire: il existe plusieurs variantes de cette méthode qui diffèrent surtout par la façon de coder la source d'excitation. On rencontre de la prédiction linéaire sur les **pistes 19, 20 21, 68 et 88**.

Diphones/PSOLA: technique brevetée par le CNET-Lannion, elle a été implémentée ensuite par d'autres laboratoires. Diverses versions de PSOLA se trouvent sur les **pistes 23, 24, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 48, 51, 59, 60, 61, 62, 64, 65, 66, , 63, 77, 78, 79**.

Diphones/MBROLA: technique dérivée de PSOLA, mais brevetée par le laboratoire TCTS de la FPMS [DUT97]. Au contraire de PSOLA, il n'y a qu'une seule implémentation de l'algorithme MBROLA, diffusée librement (voir plus haut dans la partie consacrée à EULER), mais beaucoup de bases de diphones différentes à travers le monde. Ecouter les **pistes 42, 43, 44, 47, 49, 50, 53, 54, 55, 56, 57, 70, 71, 72, 73, 74, 75, 81, 84, 86**. Les **pistes 46, 47, 48, 49** permettent d'entendre les mêmes textes, avec la même base de diphones en PSOLA et en MBROLA.

La synthèse par règles à formants est une technique tout à fait différente. Le synthétiseur à formants est un modèle numérique de l'appareil vocal, construit d'après la théorie acoustique de production de la parole. Ce modèle est piloté par un jeu de

règles acoustico-phonétiques qui décrivent explicitement le signal de parole. Ce type de synthèse est le plus économique, en terme de ressources informatiques, puisque le synthétiseur à formants contient quelques centaines de lignes de programme et puisque quelques centaines de règles de synthèse suffisent. Par contre c'est celui qui demande le plus de connaissances scientifiques, en acoustique et en phonétique. Sur ce disque sont représentés 4 systèmes de synthèse par règles à formants.

Règles/Synthétiseur de Klatt: le synthétiseur de Klatt est un synthétiseur à formants en série et en parallèle, diffusé librement, qui a donné naissance à des systèmes de qualité remarquable, et dont on trouve des exemples sur les **pistes 10, 11, 17**.

Règles/Synthétiseur de Polyglot: très proche et dérivé du synthétiseur de Klatt, sur la **piste 28**.

Règles/Synthétiseur du KTH: issu des travaux mené au KTH (Stockholm), synthétiseur à formants en série qui applique les résultats fondamentaux de ce laboratoire sur l'acoustique de la parole, un des premiers systèmes commerciaux multilingues, sur les **pistes 15, 26, 83**.

Règles/Synthétiseur FOF: une technique originale de synthèse à formants qui s'est développée par la suite pour la synthèse de voix chantée, sur la **piste 13**.

Avec le développement considérable de l'informatique ces dernières années une nouvelle technique de synthèse s'est largement développée: la synthèse par sélection et concaténation. Une grande base de donnée de parole est enregistrée, puis analysée et segmentée (automatiquement ou non). La synthèse consiste à sélectionner et à concaténer des fragments de parole de tailles variables (UNU, unités non-uniformes), qui représentent le mieux le texte à synthétiser étant donné la base de parole disponible [SAG88, TAL92, BLA95]. Ce type de synthèse ne demande quasiment pas de connaissance en phonétique, mais des outils puissants d'étiquetage, des algorithmes d'optimisation et de l'espace mémoire, toutes choses dont l'ingénieur contemporain est largement pourvu.

UNU/PSOLA: à la suite des travaux en diphones avec PSOLA, la **piste 40** montre le mélange de diphones et d'unités plus longues.

UNU: ces systèmes sont basés purement sur la sélection et la concaténation d'unités. La prosodie également est traitée par sélection, **pistes 88, 89, 90, 92, 93, 94, 95, 96, 97**.

5.2. *Iitinéraires prosodiques*

Le texte de référence: permet d'écouter 20 exemples différents avec exactement le même paragraphe de quelques phrases. Ce texte ne pose pas vraiment de problème de phonétisation, et permet donc d'apprécier à la fois la qualité de voix

globale des systèmes, et la qualité prosodique, sur les **pistes 2, 31, 32, 44, 48, 49, 50, 57, 59, 64, 70, 71, 79, 81, 88, 89, 90, 95, 96, 97**.

Variations prosodiques: les exemples représentent 3 types d'approches pour la prosodie. Sur la **piste 4** se trouve un système sans calcul d'intonation. Les durées phonémiques sont fixées par la longeur des diphonèmes. La plupart des systèmes utilisent une prosodie calculée par règles et réalisée par traitement du signal. Les règles se basent sur une analyse du texte par mots outils, ou sur une analyse syntaxique plus poussée. Les exemples de synthèse par règles de la prosodie sont sur les **pistes 6 à 88**. La troisième approche est de traiter la prosodie par sélection: on n'impose pas le détail de la prosodie, mais on cherche dans la base les éléments adéquats. Il n'y a pas de règles de calculs prosodiques, mais seulement un étiquetage prosodique de la base de donnée et une sélection des meilleurs segments: **pistes 89, 90, 92, 93, 94, 95, 96, 97**.

La prosodie selon MBROLA: pour une écoute encore plus concentrée de la prosodie, on trouvera exactement la même voix (mêmes diphones, même système de synthèse) grâce au projet MBROLA, pour le texte de référence, avec 5 réalisations prosodiques différentes sur les **pistes 44, 50, 57, 70, et 86** (ce dernier exemple est le texte de référence augmenté de quelques phrases).

Variations régionales: des systèmes de synthèse ont été développés pour le francais, dans plusieurs pays francophones ou non (Belgique, Canada, Etats-Unis, France, Suède, Suisse). Cela se traduit par des accents régionaux différents pour les synthétiseurs. Le système multilingue d'Infovox possède un accent germanique ou nordique prononcé, sur les **pistes 15, 26 et 83**. Pour une trace d'accent Wallon ou Lorrain, peu marqué toutefois, on écoutera les pistes **42, 43, 44, 50, 53, 54, 55, 56, 57, 70, 71, 72, 73, 74, 75, 81, 84, 86**. Les **pistes 10, 11, 17** possèdent un accent canadien marqué et la **piste 68** un accent canadien plus léger. La **piste 17** est très remarquable du point de vue de la prosodie du français canadien. La **piste 65** contient un exemple de français prononcé avec un fort accent ibérique. La **piste 81** provient de Suisse Romande et possède un accent relativement marqué. Pour les autres pistes, il s'agit d'un accent français peu marqué aux oreilles de l'auteur, tel qu'on le rencontre actuellement en Ile-de-France ou dans les grands médias nationaux.

Prosodie expressive: certains systèmes cherchent à rendre la synthèse plus expressive en étudiant les styles prosodiques. De par les limites de l'analyse du texte, il faut néanmoins introduire des annotations manuellement dans le texte pour spécifier l'expression voulue. On écoutera des exemples de prosodie expressive sur les **pistes 35, 54, et 77**.

Enfin, dans les tableaux 1, 2, 3, on trouvera mention du genre du synthétiseur (femme ou homme) et de la bande passante (qui joue un rôle important pour la qualité perçue).

5.3. *Conclusions*

Une écoute verticale du disque, entre disons les plus anciens et les plus récents exemples de synthèse, montre une évolution importante de la qualité. Cependant, la synthèse est clairement identifiable comme de la parole artificielle, au bout de quelques secondes. Dès les premiers systèmes, la parole était intelligible, bien qu'assez peu agréable. Elle pouvait donc aider dans les cas où elle était le plus nécessaire, par exemple comme système de lecture pour les déficients visuels, mais était difficilement acceptable sinon.

Depuis quelques années, la qualité de la parole de synthèse a été jugée acceptable pour les services téléphoniques, comme en témoignent plusieurs exemples du disque. Elle remplace alors le locuteur humain dans des tâches d'information téléphonique automatique: lire un texte, sans expression, contexte ni situation (mél, annuaire inverse, état d'un stock, par exemple).

En franchissant encore une étape, la synthèse pourrait être utilisée dans les interfaces intelligentes. Il faudrait alors améliorer les systèmes dans le sens d'une synthèse "située" qui soit capable de prendre mieux en compte la situation d'élocution et d'énonciation, l'ancrage pragmatique de l'acte de parole.

Le disque met aussi en évidence deux aspects plus pessimistes de la synthèse. En premier lieu, les progrès sont manifestement lents. Pendant les 33 ans de recherches représentés ici, on ne trouve finalement que peu de ruptures importantes dans les méthodes ou dans le résultat sonore. Je laisse à l'auditeur le soin de chercher pour lui-même où se situent ces ruptures. Si on compare ce résultat avec par exemple l'évolution de la technologie informatique, les progrès peuvent être qualifiés de lents. En second lieu, depuis 33 ans les progrès en synthèse ne proviennent-t-ils pas autant, sinon plus, de progrès technologiques en électronique et en informatique que de progrès scientifiques fondamentaux ? Il est manifeste en termes d'applications que les meilleurs systèmes actuels utilisent un bagage phonétique et linguistique assez mince et un bagage acoustique encore plus mince et il n'est pas certain que cette situation soit appelée à évoluer dans les années à venir.

Remerciements

Cet article n'aurait pas été possible sans l'aide des collègues qui ont bien voulu mettre à disposition leurs exemples de synthèse. Cela a parfois impliqué une recherche d'enregistrements déjà anciens. L'auteur tient donc à exprimer toute sa sincère reconnaissance aux auteurs des articles de ce volume, ainsi qu'à Gérard Bailly, Filip Deprez, Martine Garnier-Rizet, Eric Keller, Jean-Sylvain Liénard, Douglas O'Shaughnessy, Romain Prudon, Xavier Rodet, Nicklas Sajanti, Daniel Teil. Les dates pour les exemples anciens sont parfois incertaines. Lorsqu'une date n'était pas certaine sur l'enregistrement, j'ai choisi la date de la publication correspondant à l'exemple, date qui peut ne pas être la date véritable de l'enregistrement en question. Je tiens

également à remercier Daniel Teil et Jean-Sylvain liénard pour leurs remarques et corrections sur ce texte.

6. Bibliographie

[AST79] Asta, V., Liénard, J. S.: L'icophone logiciel - un synthétiseur par formes d'ondes", actes des 10e Journée d'Etude sur la Parole, Grenoble, mai 1979.

[BAI88] Bailly, G., Murillo, G., Dakkak, O.A., and Guérin, B. A text-to-speech synthesis system for French by formant synthesis. In 7th FASE Symposium, pages 225-260, 1988.

[BAI89] Bailly, G. Integration of rhythmic and syntactic constraints in a model of generation of French prosody. Speech Communication, 8:137-146, 1989.

[BAI92] Bailly, G. and Alissali, M. Compost: a server for multilingual text-to-speech system. Traitement du Signal, 9(4):359-366, 1992.

[BAR94] Barbosa, P. and Bailly, G. Characterisation of rhythmic patterns for text-to-speech synthesis. Speech Communication, 15:127-137, 1994.

[BAR87] Bartkova K., Sorin C. A model of segmental duration for speech synthesis in French. Speech communication, vol. **6**, pp. 245-260. 1987.

[BLA95] Black, A.W., Campbell, N. Optimizing selection of units from speech databases for concatenative synthesis. Proc. of Eurospeech 95, 581-584, 1995.

[CHO75] Choppy C., Liénard J. S., Teil D. Un algorithme de prosodie automatique sans analyse syntaxique, actes des 6e Journées d'Etude sur la Parole, Toulouse, mai 1975.

[CAR76] Carlson, R. Granström, B. A text-to-speech system based entirely on rules, Proc. IEEE-ICASSP'76, pp. 686-688, Paris, 1976.

[CAR82] Carlson, R. Granström B., & Hunnicutt, S, A multi-language text-to-speech module, Proc. IEEE-ICASSP'82, Vol. 3, pp. 1604-1607, Paris, 1982.

[CAR90] Carlson, R., Granström, B. & Nord, L. (1990). Segmental evaluation using the Esprit/SAM test procedures and mono-syllabic words. In *Talking Machines* (Bailly, G. & Benoit C. eds), pp 443-453. 1990.

[CHO77] Choppy, C. Introduction de la prosodie dans la synthèse vocale automatique Thèse de docteur-ingénieur, Université Paris VI, 1977.

[COO00] Coorman, G., Fackrell, J., Rutten, P., Van Coile, B. Segment Selection in the L&H RealSpeak Laboratory TTS System, 6th International Conference on Spoken Language Processing: ICSLP-2000, Beijing, China, Vol. 2, pp. 395-398, 2000.

[DIV77] Divay, M, Guyomard, M. Conception et réalisation sur ordinateur d'un programme de transcription graphémo-phonétique du français. Thèse de 3ème cycle (informatique), Université de Rennes, 1977.

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

[EME77] Emerard, F. Synthèse par diphones et traitement de la prosodie. Thèse de 3ème cycle (linguistique), Université Grenoble III, 1977.

[GAR91] Garnier-Rizet, M. A rule-based segmental synthesis module for French. Proc. of Eurospeech 91, Genova, Vol. 1 51-54, 1991.

[GAR93] Garnier-Rizet, M. Elaboration d'un module de règles phonético-acoustiques pour un système de synthèse à partir du texte pour le français. Thèse de doctorat, sciences du langage, Université Paris III, 1993.

[GAU97] Gaudinat, A. & E. Wehrli, 1997. 'Analyse syntaxique et synthèse de la parole: le projet FIPSVox',prosodie et syntaxe, t.a.l, 38:1, pp. 121-134.

[GAU98] Gaudinat,A. , Goldman J-P., & E. Wehrli, 1998. 'Le système de synthèse FIPS-Vox:syntaxe, phonétisation et prosodie', XXIIèmes JEP, juin 98, Martigny, Switzerland, pp.139-142.

[HAM88] Hamon, C., Synthèse de la parole par concaténation de formes d'ondes. 17ème Journées d'étude sur la parole de la Société Française d'Acoustique, Nancy, 1988.

[HAM89] Hamon Ch., Moulines E. & Charpentier F. A diphone synthesis system based on time-domain prosodic modifications of speech. Proc. IEEE-Int. conf. Acoust., Speech, Signal Process. ICASSP 89, pp. 238-241. 1989.

[HOL00] Holm, B. and Bailly, G. Generating prosody by superposing multi-parametric overlapping contours. In Proceedings of the International Conference on Speech and Language Processing, Beijing - China, 2000.

[KLA87] Klatt D., Review of text-to-speech conversion for English, (avec disque souple 33 tours) J. Acoust. Soc. Am., Vol. 82, 737-793. 1987.

[KEL97] Keller, E. & Zellner B. (éds.). Les défis actuels en synthèse de la parole. Études des Lettres, 3.. Université de Lausanne. 1997.

[LAR76] Lareur, D., Emerard, F. Speech synthesis by Dyads and automatic intonation processing. Proc. of IEEE-Int. Conf. Acoust. Speech, Signal Process., ICASSP'76, 694-697, 1976.

[LCL68] Leipp, E. Castellengo, M., Liénard, J.S. La synthèse de la parole à partir de digrammes phonétiques, 6e Int. Cong. on Acoust., Tokyo, août 1968.

[LEI68] Leipp, E, Castellengo, M., Liénard, J.S., Quinio, J., Sapaly, J., Teil D. Générateur synthétique de parole. Brevet ANVAR nř 182925, décembre 1968.

[LIE70] Liénard, J.S., Teil, D. Les éléments phonétiques et la traduction automatique du message écrit en message parlé, Automatisme, nř10, octobre 1970.

[LIE77] Liénard, J.S., Choppy, C., Teil, D., Renard, G., Sapaly, J. Diphone synthesis of French: vocal response unit and automatic prosody from the text. Proc. of IEEE-Int. conf. on Acoust., Speech, Signal Process. ICASSP'77 Hartford, May 1977.

[MAC85] Machines Parlantes 85. Catalogue de la 1ère exposition internationale Synthèse et reconnaissance de la parole. 6-12 mai 1985, le Carrefour interational de la communication, Paris, la Défense, 1985.

[MOR01] Morlec, Y., Bailly, G., and Auberg=E9, V. Generating prosodic attitudes i= n French: data, model and evaluation. Speech Communication, 357-371, 2001.

[MOU90] Moulines E. and Charpentier, F., Pitch-synchronous waveform processing techniques for text-to-speech synthesis using dyphones, Speech Communication, Vol.9,Nos.5/6, pp.453-467. 1990

[OSH81] O'Shaughnessy, D. A study of French vowel and consonant durations. J. of Phonetics, 9(4), 385-406 (1981).

[OSH82] O'Shaughnessy, D. A study of French spectral patterns for synthesis. J. of Phonetics, 10(4), 377-399 (1982).

[OSH89] O'Shaughnessy, D. Parsing for Text-to-Speech Synthesis using only a Small Dictionary, Computational Linguistics, 15, no. 2, 97–108, 1989.

[PRU01] Prudon R. & d'Alessandro, C. A selection/concatenation text-to-speech synthesis system: databases development, system design, comparative evaluation. 4th Int. Workshop on Speech Synthesis. Perthshire, Scotland, August 29th - September 1st, 2001.

[ROD77] Rodet X., 1980. Analyse du signal vocal dans sa dimension amplitude-temps. Synthèse de la parole par règles. Rapport CEA-R-4875, Thèse de doctorat d'état, Université Paris VI, 1977.

[ROD80] Rodet X. Time-domain formant-wave-function synthesis. In Spoken language generation and understanding, J. C. Simon éditeur, D. Reidel publishing company, Dordrecht, Pays-Bas. Republié dans Computer Music Journal,Vol. 8, No. 3, pp. 9-14. 1980.

[RUT00] Rutten, P., Coorman, G., Fackrell J.and Van Coile, B. Issues in Corpus based Speech Synthesis. In IEE Seminar on State of the Art in Speech Synthesis, 13 April 2000, Savoy Place, London, pp. 16/1-16/7

[SAG88] Sagisaka, Y., Speech synthesis by rule using an optimal selection of non-uniform units. Proc. IEEE-Int. conf. Acoust. Speech and Sig. CASSP'88, pp. 679-682, 1988.

[STE85] Stella, M., Charpentier, F. Synthèse par diphones utilisant le codage prédictif multi-impulsionnel et un vocodeur de phase. 14ème journées d'étude sur la parole, JEP'85, Paris, 1985.

[TAL92] Takeda, K., Abe, K., Sagisaka, Y., On unit selection algorithms and their evaluation in non-uniform unit speech synthesis. Proc. 1rst ESCA Workshop on Speech Synthesis, Autrans, 35-38, 1992.

[TEI74] Teil, D., Castellengo, M., Sapaly, J., L'unité à réponse vocale Icophone V. 5ème journées d'étude sur la parole, JEP'74, Orsay, 1974.

[TEI75] Teil, D. Conception et réalisation d'un terminal à réponse vocale. Thèse de docteur-ingénieur, Université Paris VI, 1975.

[ZEL98] Zellner, B. Caractérisation et prédiction du débit de parole en français. Une étude de cas. Thèse de Doctorat. Faculté des Lettres, Université de Lausanne. 1998.

[BEL01] http:www.bell-labs.com/project/tts/

[CRI01] http://www.elsap.unicaen.fr/

[ELA01] http://www.elan.fr/

[ELE01] http://www.electrel.fr/

[FTR01] http://www.rd.francetelecom.fr/

[GRE01] http://www.info.unicaen.fr/greyc/

[ICP01] http://www.icp.inpg.fr/

[INF01] http://www.infovox.se/

[INR01] http://www.inrs-telecom.uquebec.ca/

[LAI01] http://www.unil.ch/imm/docs/LAIP/LAIP.html

[LAT01] http://www.latl.unige.ch/

[LEH01] http://www.lhsl.com/default2.htm

[LIM01] http://www.limsi.fr

[LOQ01] http://www.loquendo.com

[MIN01] http://bach.arts.kuleuven.ac.be/pmertens/prosody/mingus.html
[MOL0101] http://www.ims.uni-stuttgart.de/ moehler/synthspeech/examples.html
[LPL01] http://www.lpl.univ-aix.fr/
[LAM01] http://www.lam.jussieu.fr
[MBR01] http://tcts.fpms.ac.be/synthesis/mbrola.html

Piste No	Contenu	Système	Durée
TOTAL			**59:01:52**
Piste 1	Introduction	(locuteur naturel)	00:20:00
Piste 2	**Texte référence 1**	(locuteur naturel)	00:23:69
Piste 3	*Présentation 1*	(locuteur naturel)	00:31:32
Piste 4	Texte libre 1	Icophone III - 1968	02:01:36
Piste 5	*Présentation 2*	(locuteur naturel)	00:15:16
Piste 6	Texte libre 2	Icophone V - 1974	01:37:59
Piste 7	*Présentation 3*	(locuteur naturel)	00:12:12
Piste 8	Texte libre 3	CNET-1976	00:50:67
Piste 9	*Présentation 4*	(locuteur naturel)	00:11:43
Piste 10	Texte libre 4	INRS-1977	02:28:22
Piste 11	Texte libre 5	INRS-1980	02:56:10
Piste 12	*Présentation 5*	(locuteur naturel)	00:12:71
Piste 13	Texte libre 6	SARA-1977	02:01:27
Piste 14	*Présentation 6*	(locuteur naturel)	00:10:30
Piste 15	Texte libre 7	Infovox-SA101-1982	06:12:38
Piste 16	*Présentation 7*	(locuteur naturel)	00:11:17
Piste 17	Texte libre 8	INRS-1984	01:23:68
Piste 18	*Présentation 8*	(locuteur naturel)	00:12:47
Piste 19	Texte libre 9	CNET-1985	00:33:48
Piste 20	Texte libre 10	CNET-Multivox-1989	00:46:61
Piste 21	Texte libre 11	CNET-Televox-1989	00:17:42
Piste 22	*Présentation 9*	(locuteur naturel)	00:07:49
Piste 23	Texte libre 12	CNET-PSOLAKDG-1989	01:44:22
Piste 24	Texte libre 13	CNET-PSOLAKDG-1989	01:27:57
Piste 25	*Présentation 10*	(locuteur naturel)	00:08:74
Piste 26	Texte libre 14	Infovox-1991	01:17:40
Piste 27	*Présentation 11*	(locuteur naturel)	00:13:53
Piste 28	Texte libre 15	LIMSI-Polyglot-1992	00:10:23

Tableau 4. *Première partie: exemples sonores de systèmes anciens*

A. Pistes du disque compact

Piste No	Contenu	Système	Durée
Piste 29	*Présentation 12*	(locuteur naturel)	00:10:04
Piste 30	*Présentation 13*	(locuteur naturel)	00:14:68
Piste 31	**Texte référence 2**	FTR&D - TAL2001	00:23:06
Piste 32	**Texte référence 3**	FTR&D - TAL2001	00:22:61
Piste 33	Texte libre 16	FTR&D - TAL2001 (QuiDonc)	01:00:14
Piste 34	Texte libre 17	FTR&D - TAL2001 (LeMel)	00:25:10
Piste 35	Texte libre 18	FTR&D - TAL2001	00:06:27
Piste 36	Texte libre 19	FTR&D - TAL2001 (Anglais)	00:16:53
Piste 37	Texte libre 20	FTR&D - TAL2001 (Espagnol)	00:05:36
Piste 38	Texte libre 21	FTR&D - TAL2001 (Allemand)	00:25:54
Piste 39	Texte libre 22	FTR&D - TAL2001 (Russe)	00:12:53
Piste 40	Texte libre 23	FTR&D - TAL2001	00:19:34
Piste 41	*Présentation 14*	(locuteur naturel)	00:07:68
Piste 42	Texte libre 24	SYNTAIX	00:32:00
Piste 43	Texte libre 25	SYNTAIX	00:21:54
Piste 44	**Texte référence 4**	SYNTAIX	00:20:46
Piste 45	*Présentation 15*	(locuteur naturel)	00:10:45
Piste 46	Texte libre 26	LIMSI - TAL2001	00:57:16
Piste 47	Texte libre 27	LIMSI - TAL2001	00:54:01
Piste 48	**Texte référence 5**	LIMSI - TAL2001	00:22:73
Piste 49	**Texte référence 6**	LIMSI - TAL2001	00:22:73
Piste 50	**Texte référence 7**	LIMSI - TAL2001	00:24:11
Piste 51	Texte libre 28	LIMSI - TAL2001 (Espagnol)	00:24:14
Piste 52	*Présentation 16*	(locuteur naturel)	00:10:73
Piste 53	texte libre 29	MINGUS/FIPS - TAL2001	01:28:68
Piste 54	Texte libre 30	MINGUS/FIPS - TAL2001	00:11:40
Piste 55	Texte libre 31	MINGUS/FIPS - TAL2001	00:41:69
Piste 56	Texte libre 32	MINGUS/FIPS - TAL2001	00:30:13
Piste 57	**Texte référence 8**	MINGUS/FIPS - TAL2001	00:20:06
Piste 58	*Présentation 17*	(locuteur naturel)	00:10:71
Piste 59	**Texte référence 9**	KALI - TAL2001	00:25:07
Piste 60	Texte libre 33	KALI - TAL2001	00:25:61
Piste 61	Texte libre 34	KALI - TAL2001	00:24:61
Piste 62	Texte libre 35	KALI - TAL2001	00:18:18
Piste 63	*Présentation 18*	(locuteur naturel)	00:20:04
Piste 64	**Texte référence 10**	ELAN - TAL2001	00:23:35
Piste 65	Texte libre 36	ELAN - TAL2001	00:51:70
Piste 66	Texte libre 37	ELAN - TAL2001	00:18:01
Piste 67	*Présentation 19*	(locuteur naturel)	00:06:26

Tableau 5. *Deuxième partie (1): exemples sonores de systèmes décrits dans ce volume*

Piste No	Contenu	Système	Durée
Piste 68	Texte libre 38	Bell Labs - TAL2001	00:39:09
Piste 69	*Présentation 20*	(locuteur naturel)	00:14:04
Piste 70	**Texte référence 11**	EULER - TAL2001	00:18:09
Piste 71	**Texte référence 12**	EULER - TAL2001	00:18:06
Piste 72	Texte libre 39	EULER - TAL2001	00:12:51
Piste 73	Texte libre 40	EULER - TAL2001	00:13:15
Piste 74	Texte libre 41	EULER - TAL2001	00:12:30
Piste 75	Texte libre 42	EULER - TAL2001	00:13:58

Tableau 6. *Deuxième partie (2): exemples sonores de systèmes décrits dans ce volume*

Piste No	Contenu	Système	Durée
Piste 76	*Présentation 21*	(locuteur naturel)	00:12:05
Piste 77	Texte libre 43	ICP	00:12:65
Piste 78	Texte libre 44	ICP	00:25:68
Piste 79	**Texte référence 13**	ICP	00:22:05
Piste 80	*Présentation 22*	(locuteur naturel)	00:08:02
Piste 81	**Texte référence 14**	FipsVox	00:21:63
Piste 82	*Présentation 23*	(locuteur naturel)	00:08:16
Piste 83	Texte libre 45	Infovox	00:18:69
Piste 84	Texte libre 46	Infovox	00:35:14
Piste 85	*Présentation 24*	(locuteur naturel)	00:07:25
Piste 86	Texte libre 47	LAIP	00:31:21
Piste 87	*Présentation 25*	(locuteur naturel)	00:06:42
Piste 88	**Texte référence 15**	L&H - TTS3000	00:21:45
Piste 89	**Texte référence 16**	L&H - RealSpeakClassic	00:21:39
Piste 90	**Texte référence 17**	L&H - RealSpeakCompact	00:20:63
Piste 91	*Présentation 26*	(locuteur naturel)	00:11:25
Piste 92	Texte libre 48	LIMSI - SeLimsy	00:51:53
Piste 93	Texte libre 49	LIMSI - SeLimsy	00:49:03
Piste 94	Texte libre 50	LIMSI - SeLimsy	00:50:58
Piste 95	**Texte référence 18**	LIMSI - SeLimsy	00:21:04
Piste 96	**Texte référence 19**	LIMSI - SeLimsy	00:19:54
Piste 97	**Texte référence 20**	LIMSI - SeLimsy	00:19:04

Tableau 7. *Troisième partie: exemples sonores d'autres systèmes contemporains*