La campagne EvaSy d'évaluation de la synthèse de la parole à partir du texte

Christophe d'Alessandro, Philippe Boula de Mareüil, Marie-Neige Garcia, Gérard Bailly, Michel Morel, Alexander Raake, Frédéric Béchet, Jean Véronis, Romain Prudon

Chapitre 1

La campagne Evasy d'évaluation de la synthèse de la parole à partir du texte

1. La campagne EVALDA/Evasy

1.1. Présentation

Ce chapitre est consacré à l'évaluation des systèmes de synthèse de la parole à partir du texte en Français, menée dans la campagne EVALDA/Evasy entre 2003 et 2005. Le projet Evasy fait suite, dans un autre cadre, au projet AUPELF ARC-B3 [dAL 98 ; YVO 98], le premier projet concerté au niveau de la francophonie sur l'évaluation de la parole de synthèse (1996-1999). Les systèmes de synthèse de parole à partir du texte, dont les premiers sont apparus il y a plus de trente ans, ne sont sortis des laboratoires pour des applications commerciales que depuis environ une quinzaine d'années. Pour un état de l'art sur la synthèse à partir du texte en Français au moment du début du projet Evasy, le lecteur pourra consulter [BOE 02, dAL 01]), et écouter le disque compact d'exemples sonore associé [dAL 01-2].

Avec la diffusion de l'ordinateur, la popularisation d'Internet et l'émergence de la Société de l'Information, la synthèse de parole prend une part de plus en plus importante dans la communication homme-machine pour répondre aux besoins des applications embarquées (automobile, traducteurs automatiques de poche), des télécommunications (services de consultation de courrier électronique par téléphone, serveurs vocaux interactifs, livres et journaux parlants) et du multimédia (jeux informatiques, aide aux handicapés, « machines à lire » avec scanner et OCR pouvant servir aux aveugles et malvoyants). La qualité de synthèse atteinte

---

Chapitre rédigé par Christophe d'Alessandro, Philippe Boula de Mareüil, Marie-Neige Garcia, Gérard Bailly, Michel Morel, Alexander Raake, Frédéric Béchet, Jean Véronis et Romain Prudon

actuellement est acceptable pour de nombreuses applications, mais en contrecoup l’exigence de qualité et de naturel s’accroît.

Ces cinq dernières années ont été marquées par le développement de la synthèse par sélection et concaténation. Cette technique de synthèse de la parole repose sur les progrès de l'étiquetage automatique et sur l'utilisation de grosses bases de données de parole enregistrée. Le signal est synthétisé par l’assemblage d’unités sonores de taille variable (diphones, syllabes, mots, voire incises) sélectionnées dans un grand échantillon de parole naturelle. Ces techniques posent des problèmes d’évaluation nouveaux car l’évaluation du “ naturel ” des voix devient prépondérante, et ceci en plus des exigences d’intelligibilité. Ainsi de nouveaux tests doivent être mis en œuvre puisque la distance entre parole naturelle et parole synthétique semble diminuer. Dans la génération précédente de systèmes, le timbre de la voix du locuteur était fortement distordu. Aujourd’hui, le timbre est mieux traité, mais il y a des problèmes de cohérence de la voix dus essentiellement à la longue durée de l’enregistrement original. Cette nouvelle situation appelle de nouveaux de tests, permettant en particulier de comparer objectivement les performances obtenues, en comparaison avec les techniques précédentes.

1.2. Organisation scientifique du projet

Tous les systèmes Français connus ont été invités à participer au projet. Une majorité a répondu positivement, et la campagne Evasy a réuni les partenaires suivants, que nous pensons tout à fait représentatifs de la situation actuelle en synthèse de la parole en Français :

- Organisateur des tests : ELDA ; L’organisateur a recruté les sujets, fait passer les tests, développé une partie des logiciels de test et d’analyse, préparé une partie des corpus de test, participé à l’analyse et la publication des résultats, organisé les réunions et l’administration générale du projet [MAP 04].

- Comité d’organisation scientifique : LIMSI-CNRS (responsable), DELIC, Université de Provence: (fournisseur de corpus, graphème-phonème), CRISCO, université de Caen, ICP Grenoble. Le comité a défini les paradigmes de test, développé une partie des logiciels de test et d’analyse, préparé les corpus de test avec l’organisateur, analysé et publié les résultats.

- Participants aux tests : Acapela Group – ELAN (2 systèmes complets), CRISCO (système complet), ICP Grenoble (système complet), LIA, université d’Avignon (système graphème–phonème), LIMSI-CNRS (2 systèmes complets), MULTITEL ASLB (système complet). Les

participants ont fourni les échantillons de synthèse demandés par l'organisateur, participé à l'analyse des résultats.

La notion de qualité est une notion complexe, liée à la fois aux objets étudiés, mais aussi à leur fonction, et à leur représentation cognitive. Ainsi plusieurs niveaux d'évaluation coexistent, du niveau hédonique (l'agrément, la beauté de la voix), jusqu'aux niveaux fonctionnel et ergonomique (l'utilité en contexte d'application, la fatigue induite) en passant par des niveaux intermédiaires sur la performance technique (comme l'intelligibilité, ou le taux d'erreur). Nous n'allons pas répéter ici la description générale des techniques d'évaluation de la synthèse de la parole que le lecteur trouvera ailleurs [dAL 04].

Après une phase de réflexion, il nous a semblé nécessaire à la fois de proposer des méthodologies nouvelles pour aborder des problèmes difficiles, et également de donner des mesures plus classiques pour évaluer les performances des systèmes, ceci afin d'établir le niveau de performance actuel. Ainsi notre travail a porté sur quatre aspects différents et complémentaires de l'évaluation. Le premier aspect, à la suite du projet ARC-B3, s'est attaché à évaluer la qualité de conversion graphème-phonème sur la tâche la plus difficile, la conversion des noms propres [BOU 06]. La seconde évaluation a porté sur la prosodie [GAR 06]. La troisième partie est consacrée à l'‘intelligibilité, avec un nouvel ensemble de phrases sémantiquement imprédictibles [BOU 06-2]. Enfin la dernière partie est un test d'opinion directe par catégories absolues de jugement [BOU 06-2].

2. Évaluation de la prononciation des noms propres

2.1. Introduction

Cette partie est consacrée à l'évaluation du module de transcription graphème-phonème (GP) de quatre systèmes : ceux du CRISCO, de l'ICP, du LIA et du LIMSI (systèmes tous décrits dans [dAL 01]) : ils reposent tous sur des approches à base de règles, éventuellement complétées de lexiques d'exceptions (jusqu'à plusieurs milliers d'entrées), sauf un système (Labo4), qui implémente l'algorithme ID3 à l'instar de [BLA 98].Un cinquième laboratoire, le DELIC, était chargé de produire le corpus. Il a été montré que la majorité des erreurs de transcription GP, pour les meilleurs systèmes, provenait des noms propres. Lors de la campagne ARC-B3, au cours de laquelle d'importantes ressources avaient été fournies — plusieurs dizaines de milliers de mots [YVO 98], le taux de 99,6 % de phonèmes corrects (99,1 % de mots corrects), obtenu par le meilleur système sur des textes de journaux, laisse penser que fournir manuellement une transcription phonétique de référence de tels corpus est très coûteuse, pour ne mettre en lumière finalement et laborieusement que peu d'erreurs. Dans une tâche d'évaluation limitée à une liste de noms propres, on

peut s’attendre à des scores bien différents (80–90 %). Ainsi nous nous sommes concentrés sur ces noms propres, items lexicaux souvent d’une grande importance sémantique, dont la prononciation incorrecte est susceptible de dégrader grandement le jugement global porté sur la qualité du système.

Les noms propres posent un problème épineux pour la synthèse vocale à partir du texte, car leur prononciation dépend fortement de leur origine et de l’usage. Dans ceux d’origine étrangère, la prononciation résulte d’un compromis entre respect de la graphie originale et approximation de la prononciation originale tout en suivant les conventions françaises. Ceci vaut surtout lorsque les langues source et cible partagent le même matériau graphique (l’alphabet latin) : il n’y a dans ce cas pas de réaménagement orthographique. Mais même la translittération (du cyrillique ou de l’arabe, par exemple) ne résout pas tous les problèmes. Les spécificités ou idiosyncrasies orthographiques sont frappantes, reflétant le caractère mono-référentiel du nom propre (dénotant une entité unique), la difficulté de le définir, d’en donner une typologie et un traitement lexicographique cohérent [GAR 94 ; LER 04]. Si un texte parle d’une personne ou d’une ville dont on soupçonne qu’elle est allemande, le nom, même d’apparence française (typiquement une célébrité telle que Berger), sera probablement lu d’une façon qui viole les principes élémentaires de notre orthoépie. Notre compétence géographique, nos connaissances des langues étrangères peuvent donc également avoir une influence. Cette diversité des usages, cette mosaïque linguistique sont illustrées dans de grandes bases de données de noms propres, comme, pour 11 langues européennes [ONO 95], dont le contenu a été pour partie collecté par enquêtes. Pour les phonétiser automatiquement, différentes solutions ont été proposées depuis: systèmes experts à base de règles [DIV 97 ; BOU 97], et modèles d’apprentissage automatique par analogie [YVO 96 ; BAG 98 ; BLA 98 ; DAM 01, notamment). Ces deux familles de techniques, toutes deux représentées au sein de nos laboratoires, méritent d’être évaluées et comparées.

À des fins de diagnostic et de développements ultérieurs, il est intéressant qu’une base soit étiquetée en origines linguistiques. Dans une optique de synthèse [CHU 86 ; VIT 91 ; BÉC 97 ; BÉC 00 ; LIT 01] et de reconnaissance de la parole [BAR 99], des origines telles que néerlandais, allemand, anglais, italien, espagnol, polonais, arabe, japonais ont été définies pour déterminer en conséquence la prononciation des noms propres. Cette voie pourrait être élargie à des groupes, branches ou familles de langues.

2.2. Méthode

2.2.1. *Construction du corpus*

Comme il est difficile de définir ce qu'est un nom propre –notamment pour les noms de sociétés, de marques et de produits [LER 04]– nous nous sommes cantonnés aux noms de personnes. Une liste de 4 115 couples prénom-nom a été extraite du journal *Le Monde* des années 1992–2000 (plus de 200 millions de mots). Cet échantillon a été obtenu en considérant les couples de mots qui apparaissent avec une majuscule initiale entre 100 et 200 fois dans le corpus. Cette gamme de fréquence a été retenue car les mots les plus fréquents risqueraient d'avoir été prévus dans les différents systèmes, et parce que des noms sélectionnés aléatoirement auraient abouti à beaucoup de fautes de frappes et de hapax difficiles à transcrire. Les noms propres retenus sont donc d'une difficulté moyenne. Ce travail a été mené au DELIC. Le matériel sélectionné a ensuite été transcrit manuellement dans l'alphabet phonétique SAMPA pour le français [GIB 97]. Des variantes ont été introduites, comme illustré dans les exemples suivants. La barre oblique sépare les prononciations alternatives, éventuellement réduites à zéro :

Kissinger	kisin{dZ/g}{E/9}R
Griotteray	gRi{j/}{O/o}t{@/}R{E/e}

Une première transcription a été produite, laquelle a ensuite été vérifiée par un deuxième expert. Des directives de transcription étaient livrées aux experts, portant sur le schwa (ou *e* muet) optionnel, les oppositions {e/E} et {o/O}, les voyelles nasales, les glides (semi-voyelles ou semi-consonnes) et la gémination facultative notamment. Il était également conseillé d'approximer la *jota* espagnole ([x]) par [R], et les interdentales anglaises [T] et [D] par {s/t} et {d/z} respectivement. D'autres prononciations sont sujettes à la variation [BOU 00]. En définitive, 80 % des noms propres de notre corpus comportent des variantes. Nous avons pensé que la surgénération de variantes et leur éventuelle incohérence ne représentaient pas un problème trop grave, dans la mesure où les systèmes de conversion GP évalués sont déterministes (une seule prononciation est prévue).

De plus, les transcripteurs avaient accès à 10 extraits où chaque couple prénom-nom apparaissait, avec 100 mots à gauche et à droite. Ils pouvaient également lancer une recherche Google pour les noms en question en cliquant simplement sur un hyperlien. Leur situation se rapprochait donc de celle de journalistes de la radio, confrontés à des noms propres à prononcer. De cette façon aussi, notre base de données est plus qu'une simple liste de mots.

2.2.2. *Annotation linguistique*

Notre liste a été enrichie d'informations concernant l'origine des noms de famille. À cet effet, un ensemble de 20 étiquettes linguistiques a été défini, montrant des comportements communs en regard de la stratégie à laquelle nous faisons appel pour prononcer ces noms propres. La compétence géographique des Français, liée

aux connaissances linguistiques que l'on possède, a été prise en considération. Par exemple, un Français doit être en mesure de reconnaître un nom d'apparence espagnole ou italienne, qui appartient à son entourage. En revanche, il n'est pas toujours évident, ni même faisable, de distinguer parmi les noms russes, ukrainiens, bulgares et macédoniens, ou parmi les noms allemands, alsaciens, yiddish et néerlandais.

Afin de fournir une étiquette linguistique à un nom propre, nous avons tiré profit du contexte : au-delà du prénom, par exemple, la phrase dans laquelle un nom de famille apparaît donne quelque information sur la nationalité de la personne. Cette dernière indication peut être utile dans certains cas, même si elle ne va pas nécessairement de pair avec une origine linguistique. Par exemple, les anciens chefs d'état latino-américains Fujimori, Pinochet et Stroesner sont notoirement et respectivement d'origine japonaise, française et allemande. La liste d'étiquettes linguistiques retenues est consignée dans le Tableau 1. Des langues non apparentées génétiquement peuvent être regroupées (ex. albanais et turc, coréen et chinois), si nous sommes incapables de les distinguer en surface. Eu égard au grand nombre de noms anglais, et en raison de la spécificité de leur phonétisme, il nous a semblé nécessaire de distinguer l'anglais des autres langues germaniques.

Étiquette	Signification	%	Étiquette	Signification	%
fre	français	51	ind	indien	1
eng	anglais	15	chi	chinois	1
ger	germanique	10	tur	turc	1
ita	italien	5	heb	hébreu	1
sla	slave	4	prt	portugais	1
spa	espagnol	3	jpn	japonais	1
ara	arabe	3			
afr	africain	2		autre	1

Tableau 1 ***Étiquettes linguistiques avec la proportion du corpus qu'elles représentent***

De nombreux dictionnaires étymologiques et livres dédiés aux noms propres existent, tels que [MAE 93], qui présente plus de 9 000 noms propres de personnalités célèbres ou contemporaines, classés par origine, avec leurs prononciations. Quand les experts transcrivent les noms propres, ils appliquent des stratégies (en s'aidant éventuellement du contexte des articles, pour lever certaines ambiguïtés) ; ils opèrent un transfert phonologique en faisant des hypothèses sur les origines des mots. Que celles-ci soient explicitées par le biais d'étiquettes, tel était le but de cette annotation. Comme un test préliminaire l'a démontré, cette tâche n'est pas nécessairement plus délicate ni plus coûteuse en temps que la transcription phonétique elle-même, avec toutes les variantes potentielles concernant le schwa ou

les voyelles moyennes notamment. Il est même plus aisé de détecter l'origine de noms tels que Chavez, Angelopoulos, Browning ou Ruggero que de les transcrire.

2.3. Résultats

2.3.1. *Résultats globaux*

Les participants devaient adapter leurs systèmes pour fournir en sortie des transcriptions en SAMPA. Pour chaque participant, la tâche consistait à phonétiser la liste de noms propres en 3 heures. Une fois les résultats calculés et envoyés, 3 semaines d'*adjudication* étaient ensuite prévues, pour donner aux participants la possibilité de contester certaines de leurs erreurs. La phase d'adjudication a conduit à apporter des corrections ou des ajouts de variantes à environ 200 des 8 230 noms, et n'a pas changé pas la performance globale des systèmes. Une nouvelle version de la référence a été produite, corrigée ou enrichie de variantes supplémentaires. Après chaque phase, on avait une passe d'alignement entre les sorties phonémiques et la référence, fondée sur l'algorithme de programmation dynamique `sclite` (http://www.nist.gov/speech/tools/).

Le Tableau 2 présente les résultats bruts. Les taux d'erreur de 12-20 % sur des noms propres sont comparables à ceux qui avaient été obtenus sur les 1 500 noms propres du texte de l'AUPELF [YVO 98] ; ils sont sensiblement meilleurs que ceux que rapporte [LIT 01] pour l'anglais — langue pour laquelle, il est vrai, se posent en plus des problèmes liés à la place de l'accent lexical. Aucun des systèmes par règles n'a été détrôné par l'approche dirigée par les données du Labo4. Cependant, il faut noter que ce système par auto-apprentissage avait été entraîné sur des mots français d'un lexique général et que son développement est moins long que la maintenance de règles, à supposer bien sûr que l'on dispose des ressources adéquates.

%Erreur	Labo1	Labo2	Labo3	Labo4
Prénoms	8,4	10,5	12,7	13,6
Noms	17,4	23,8	21,7	25,0
Total	12,9	17,1	17,2	19,3

Tableau 2 ***Pourcentage de prénoms et de noms de famille incorrectement phonétisés***

Une tendance commune que l'on observe entre les différents systèmes est que les prénoms sont généralement mieux phonétisés que les noms de famille. Une explication est que les participants peuvent avoir veillé à la prononciation des prénoms, qui sont plus fréquents que les noms de famille. Une autre explication est que la référence peut être plus tolérante sur les prénoms que sur les noms de famille. Quand un prénom anglais existe également en français, en effet, il peut être prononcé à la française : Michael (respectivement Thomas), par exemple, est plus

enclin à être prononcé [mikaEl] (respectivement [t{O/o}ma]) en tant que prénom qu'en tant que nom de famille.

2.3.2. *Analyse par étiquette linguistique*

Les étiquettes linguistiques n'étaient pas utilisées par les convertisseurs GP qui étaient évalués ici, mais permettent d'analyser les résultats par origine (*cf.* Tableaux 3 et 4 pour les étiquettes linguistiques les plus fréquentes) et d'envisager de futures techniques qui tireraient profit de ces informations.

%Erreur	Labo1	Labo2	Labo3	Labo4
fre	5	9	13	10
eng	32	43	36	46
ger	36	52	44	48
ita	18	22	21	23
sla	27	34	17	44
spa	30	41	22	36
ara	21	20	11	24
afr	29	34	24	34

Tableau 3 ***Taux d'erreurs sur les noms de famille pour les étiquettes linguistiques les plus fréquentes (%Erreur/Étiquette)***

Une certaine hiérarchie est respectée, entre les lignes et les colonnes du Tableau 3. Dans l'ensemble, les noms français se révèlent être les mieux transcrits, les noms anglais et autres noms germaniques les moins bien transcrits. Après ces cas extrêmes, on a les noms espagnols (plutôt mal transcrits) et les noms italiens (plutôt bien transcrits). Une telle différence, qui n'était pas attendue entre langues romanes, est importante à noter, et justifie a posteriori la distinction espagnol/italien.

Si les Tableaux 2 et 3 fournissent des taux d'erreur pour un type de nom propre donné, il est également intéressant de pointer du doigt les cas les plus problématiques et leurs distributions, ce qui est l'objet des Tableaux 4 et 5. Pour tous les systèmes, les pourcentages sont plus élevés pour les noms français dans le Tableau 4, par rapport au Tableau 3. Ceci se comprend facilement puisque les noms français couvrent la majorité du corpus. Inversement, par rapport au Tableau 3, les pourcentages sont plus bas pour les noms anglais et autres noms germaniques dans le Tableau 4. Mais ces noms représentent la principale source d'erreur pour tous les systèmes. En comparaison, les noms slaves rendent compte de peu d'erreurs : leur nombre est trop peu élevé pour tirer des conclusions, même si un examen détaillé des Tableaux 3 et 4 révèle des comportements très différents pour ces noms, en particulier entre le Labo3 et le Labo4.

%Étiquette	Labo1	Labo2	Labo3	Labo4
fre	16	20	32	21
eng	28	27	25	28
ger	29	22	21	20
ita	5	4	5	4
sla	7	7	3	8
spa	5	5	3	4
ara	4	3	2	3
afr	4	3	2	3

Tableau 4 ***Pourcentage d'erreurs sur les noms de famille pour telle ou telle étiquette linguistique (%Étiquette/Erreur)***

2.3.3. *Analyse par graphème*

Une classification des erreurs par graphème est également possible. Trois types de problème, en particulier, rendent compte d'un grand nombre d'erreurs : la voyelle 'e', qui est élidée ou prononcée comme un schwa ; les digrammes 'an', 'en', 'in', 'on' et 'un' qui sont improprement nasalisés ; les consonnes finales *–d, -g, -r, -s, t, -x, -z* qui ne sont pas prononcées. Le cas de la terminaison *–er* est particulier, dans la mesure où cette chaîne de caractère aboutit souvent à la prononciation [e] au lieu de [ER] ou [9R] (par exemple dans Schwarzenegger).

%Graphème	Labo1	Labo2	Labo3	Labo4
'e'	16,7	12,5	5,6	9,8
'Vn'	19,5	22,0	10,7	16,4
C finale	23,6	17,3	8,6	12,0

Tableau 5*Pourcentage d'erreurs pour tel ou tel grapheme (%Graphème/Erreur) : 'e : substitution/délétion e-{@/}; 'Vn' : nasalisation erronée des digrammes 'an', 'en', 'in', 'on' et 'un' ; C : délétion des consonnes* **-d, -g, -r, -s, -t, -x, -z**

Dans la majorité des cas, les erreurs liées au 'e' correspondent à des délétions de [e] (ex. Corea), plutôt qu'à des réalisations en schwa (ex. Boccanegra). Elles sont moins nombreuses dans le système du Labo3, qui en retour insère beaucoup de schwas superfétatoires (18,2 % des erreurs). Parmi les erreurs de type 'Vn' reportées dans le Tableau 5, les configurations les plus fréquentes sont les prononciations [a~] au lieu de [an] et [e~] au lieu de [in], provenant en partie de prénoms comme Juan ou Martin (quand le nom de famille est anglais). Ce dernier nom est un bon exemple de la dépendance contextuelle de la conversion GP. Parmi les consonnes finales qui sont le plus souvent muettes en français, l'omission d'un [s] est de loin la plus fréquente. Il y a 951 noms terminés par un *–s* ou un *–x* dans le corpus (ex. Coencas [k{O/o}Enkas]). Mais dans la majorité des cas (ex. Dumas [dyma]), on ne doit pas prononcer de [s] final.

2.4. Conclusion

Nous avons présenté un corpus et une méthodologie objective pour l’évaluation de la conversion GP des noms propres en français. Ce problème pratique et théorique s’est montré important, surtout pour les noms anglais et autres noms germaniques. Dans une optique applicative, ce sont donc sur eux que devraient se concentrer les efforts, quelle que soit la méthode employée pour détecter automatiquement l’origine linguistique des noms et pour déterminer en conséquence leur prononciation. La conversion graphème-phonème peut également être vue comme un modèle de la performance humaine : une étude de l’impact perceptif en synthèse est envisagée.

Un autre apport de ce travail est qu’il a permis d’examiner automatiquement des types d’erreur (par exemple liés au ‘e’). Les ressources qui ont permis d’établir cette grille d’analyse seront mises à la disposition de la communauté scientifique, pour servir de base pour d’autres domaines et pour d’autres langues. La construction de dictionnaires de prononciation pour la reconnaissance automatique de la parole et des applications de dictionnaire inverse serait concernée en premier lieu. Une liste dotée de prononciations de noms propres couvrant l’actualité pourrait également servir aux étrangers qui apprennent le français. Grâce aux ressources dont nous disposons pour la reconnaissance automatique de la parole, on peut rechercher comment les noms propres de notre base sont prononcés dans les journaux radio-télévisés qui aujourd’hui représentent une forme de norme de fait. Et le côté appliqué de ce travail ne dispense pas d’effectuer une recherche plus fondamentale sur la phonologie des emprunts et des noms propres.

3. Évaluation de la prosodie

3.1. Introduction

La prosodie des systèmes de synthèse est un paramètre important de la qualité perçue, et devrait idéalement ressembler à celle des voix naturelles. Pour évaluer cette qualité, un paradigme dérivé de l’étude [PRU 04] a été mis en place. Les contours intonatifs de la parole de synthèse sont transplantés, c’est à dire recopiés, sur un contenu segmental commun. Cette méthode devrait permettre l’évaluation de la prosodie indépendamment des autres modules de chaque système (comme les traitements textuels et la synthèse acoustique). Les cinq systèmes testés pour cette partie de la campagne sont 3 systèmes par diphones (nommés D1, D2 er D3) et deux systèmes par sélection/concaténation (nommés S1 et S2). Les systèmes sont issus (dans le désordre) de Acapela Group-ELAN (2 systèmes), CRISCO, ICP, LIMSI (2 systèmes). Une voix naturelle (nommée NR) sert de référence. La prosodie des systèmes par diphones est généralement calculée par un ensemble élaboré de règles

syntactico-prosodiques alors que la plupart des systèmes par sélection utilisent un ensemble réduit de règles, la prosodie étant plutôt apprise de façon implicite par les propriétés du corpus de sélection utilisé. Le lecteur intéressé est invité à consulter les références bibliographiques pour comprendre le détail du calcul prosodique.

3.2. Méthode de transplantation prosodique

Le corpus d'évaluation comprend sept phrases phonétiquement équilibrées extraites du corpus BREF [LAM 91]. La référence naturelle dure entre 4 et 11 secondes, mais la durée des signaux synthétiques est assez variable. Le corpus de phrases à tester comprend six versions de chaque phrase, 5 produites par les systèmes de synthèse et une référence naturelle. Les contours intonatifs de ces phrases (c'est-à-dire l'évolution de la fréquence fondamentale au cours du temps) sont appliqués à un contenu segmental commun.

Afin d'évaluer l'influence éventuelle de ce contenu segmental sur les jugements de qualité, deux conditions différentes ont été utilisées pour la transplantation prosodique. La première condition utilise la version française du système de synthèse par diphones MBROLA [DUT 96]. La seconde condition utilise la voix naturelle de référence, dont l'intonation est modifiée par un algorithme modification prosodique de très bonne qualité [BAI 92]. Les deux conditions sont donc nommées « voix diphone » et « voix naturelle modifiée». D'un point de vue technique, certains systèmes pouvaient produite directement la sortie prosodique sous le format utilise par MBROLA (format ".pho"), alors que d'autre systèmes ne le pouvaient pas. Dans ce second cas, le système d'alignement et d'analyse de la prosodie BROLign, [MAL 97] permettait de calculer les données prosodiques à partir du signal directement, et de les convertir au format « .pho ». Il est apparu des différences dans certains détails dans le contenu phonémique produit par les différents systèmes. Il est en effet inévitable que des différences de schwa, ou de coupes syllabiques par exemple apparaissent dès que l'on traite un contenu assez long. Une vérification manuelle du contenu phonémique des phrases a donc été nécessaire avant de pouvoir utiliser l'alignement automatique. Après ces analyses, les contours prosodiques obtenus (contour mélodique et durées) ont été appliquées sur la base de diphone MBROLA et sur la parole naturelle.

Une autre différence qui peut biaiser le jugement des sujets est la hauteur moyenne intrinsèque délivrée par les différents systèmes. En effet, chaque système utilise un registre différent (il y a en particulier des systèmes avec des voix masculines et féminines). Pour éviter ce biais, la fréquence fondamentale moyenne de tous les systèmes a été normalisée. Ainsi le registre propre à chaque système ne compte plus directement dans l'évaluation de la prosodie.

3.3. Protocole de test

Les tests de préférence par paire ont été conduits à Paris chez ELDA. 19 sujets (10 femmes et 9 hommes) âgés de 20 à 40 ans, sans problème auditif connu, ont participé à cette partie de la campagne d’évaluation. Les sujets étaient tous de langue maternelle française, sans familiarité particulière avec la parole de synthèse, et ils étaient rémunérés pour leur participation à l’expérience. Les tests se sont déroulés au casque, dans une pièce calme, en utilisant un logiciel d’évaluation automatique de test élaboré spécifiquement pour la campagne, et du matériel audio de haute-fidélité.

Pour chaque phrase, une paire de stimuli sont présentés au sujet: chaque membre de la paire est une même phrase, mais avec deux contours intonatifs différents. Les sujets sont invités à indiquer quelle version des phrases leur semble meilleure. Toutes les combinaisons possibles de paires de phrase sont évaluées. Afin d’éviter d’éventuels effets d’apprentissage ou d’ordre de présentation, les paires de phrases sont présentées dans un ordre aléatoire, et dans des ordres différents aux différents sujets. Les sujets étaient invités à écouter les phrases autant de fois qu’ils le souhaitaient, mais cependant à ne pas trop s’attarder pour prendre une décision et donc à répondre plutôt sur la base de leur impression initiale. Les sujets n’étaient pas informés du substrat segmental utilise (“voix diphone” ou “voix naturelle”). Les variables expérimentales de ce test sont donc au nombre de 4 : 1. Le système de synthèse testé (5 systèmes et référence naturelle) 2. Le substrat segmental (« voix diphone », « voix naturelle »). ; 3. Les phrases de test (7 phrases) ; 4. Les sujets (19 sujets).

3.4. Résultats

3.4.1. *Taux de préférence globaux*

Les taux de préférence globaux pour chaque système (pour les deux conditions de substrat segmental) sont donnés sur la Figure 1. Ces taux sont calculés en comptant le nombre de fois qu’un système est préféré dans une paire, divisé par le nombre de paires. Chaque système est représenté dans 665 paires. Les résultats globaux donnent l’avantage à la référence naturelle, préférée dans 80% des paires où elle est présente, suivie par S1 (71%) et D1 (58%). Des taux de préférence inférieurs à 50% sont obtenus pour les systèmes S2, D2 and D3 (respectivement 40%, 31% et 18%).

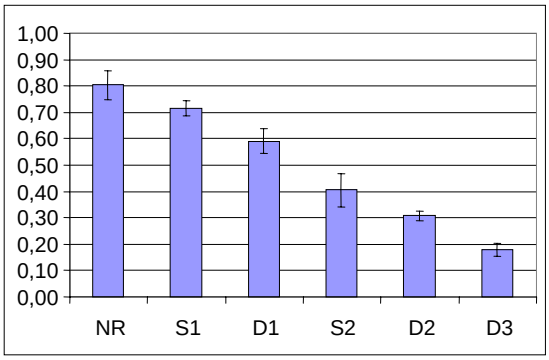
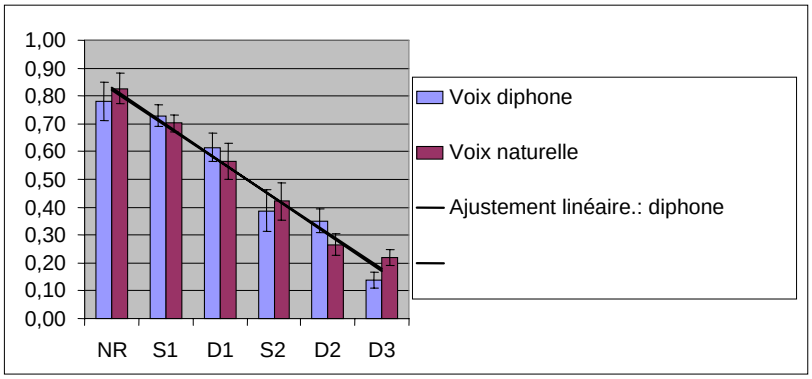


Figure 1 *Taux de préférence global pour chaque système. (Les barres indiquent les intervalles de confiance)*

Ce taux global montre donc une nette préférence pour S1, un système par sélection d'unités longues. Cependant, le second système étant basé sur des diphones, cela n'indique pas que les systèmes par sélection soient systématiquement meilleurs que les systèmes par diphones. Il semble au contraire que l'ajustement fin des systèmes soit un facteur plus déterminant dans la qualité prosodique obtenue que la méthodologie proprement dite. Dans tous les cas, le taux de préférence obtenu par le meilleur des systèmes de synthèse est inférieur à celui de NR. Ce résultat va à l'encontre de celui décrit dans [PRU 04], une étude où la référence naturelle avait obtenu un score légèrement plus faible que le meilleur système par diphones et règles prosodiques. Une explication possible est tout simplement la monotonie et le manque de conviction du locuteur naturel pris comme référence, dont l'intonation peut de fait sembler moins intéressante que celle générée par de bonnes règles de synthèse, du moins sur un petit ensemble de phrases isolées.

3.4.2. *Effet du substrat segmental*

Il serait souhaitable que les résultats ne dépendent pas du substrat segmental utilisé pour la transplantation prosodique. Nous avons introduit dans nos tests deux conditions segmentales différentes, afin d'évaluer cet aspect. Les résultats de préférence globale pour les deux conditions sont résumés dans la figure 2. Les résultats montrent que le même ordre est obtenu pour les systèmes quelque soit la méthode de transplantation prosodique utilisée. Les droites de régression linéaire des moyennes sont très voisines pour les deux conditions: $y = -0.1309x + 0.9582$ pour la condition « diphone » et $y = -0.1285x + 0.9498$ pour la condition « voix naturelle ». On peut en conclure que les différences entre les deux conditions ne sont pas significatives. La procédure de transplantation prosodique est donc viable, dans la mesure où elle ne semble pas affectée par le support de transplantation utilisée.



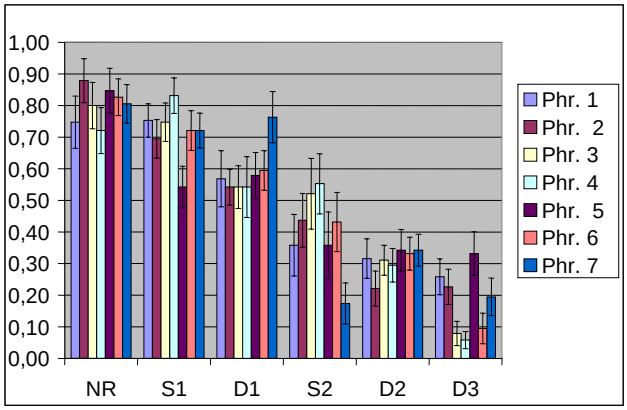


Figure 2 *En haut :taux de préférence globaux pour chaque système, en fonction de la méthode de transplantation prosodique utilisée (les barres indiquent les intervalles de confiance) En bas : taux de préférence par système et par phrase (les deux conditions segmentales confondues, les barres indiquent les intervalles de confiance)*

3.4.3. *Effet du facteur « phrase »*

À cause de la lourdeur de la procédure expérimentale, nos expériences sont limitées à un nombre relativement réduit de 7 phrases. L'effet éventuel du facteur phrase doit donc être évaluée. Les taux de préférence par système et par phrase sont affichés dans la Figure 2, pour les deux conditions segmentales confondues. Les phrases sont variées sous plusieurs aspects: durée, structure syntaxique et domaine sémantique. Il apparaît clairement que l'ordre de préférence des systèmes dépend de la phrase considérée. Par exemple, bien que l'ordre global de préférence soit NR >

S1 > D1 > S2 > D2 > D3, on peut noter que pour la phrase 4, cet ordre devient S1 > NR > S2 > D1 > D2 > D3 alors que pour la phrase 7 on obtient RN > D1 > S1 > D2 > D3 > S2. Ainsi même si pour la majorité des phrases (les 1, 2, 3 et 6) l'ordre global semble respecté, il faut reconnaître que l'effet du facteur « phrase » semble significatif.

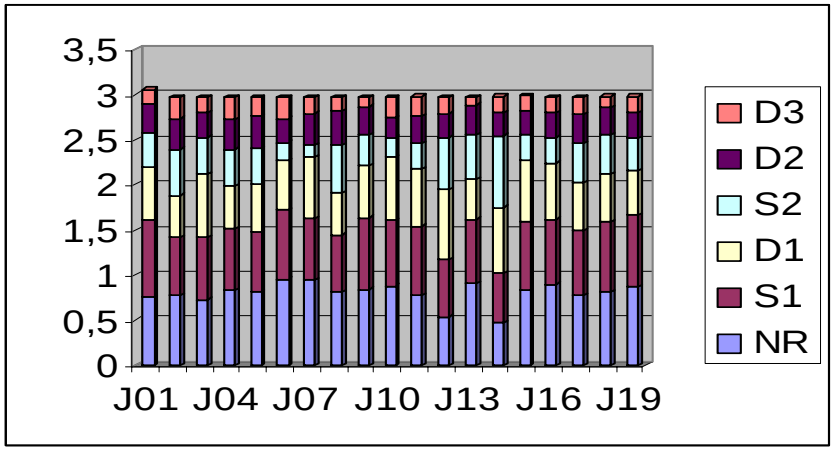


Figure 3. *Taux de préférence par sujet et par système (conditions segmentales confondues)*

3.4.4. *Différences interindividuelles*

Un nombre significatif de sujets a été recruté pour cette expérience (19). D'une façon générale, les sujets sont en accord entre eux quand à l'ordre de préférence des systèmes. Cependant, une analyse plus détaillée des résultats (cf. figure 3) montre des différences interindividuelles assez importantes.

Comme on pouvait l'espérer, ces taux sont globalement en accord avec la forme générale de la figure 2. Cependant certains sujets dévient de la moyenne : par exemple le sujet 14 donne comme ordre de préférence : S2 > D1 > S1 > NR. Le sujet 12 donne : D1 > S1 > S2>NR. Globalement, NR est le "système" préféré par 16 sujet sur 19 alors que S1, D1 et S2 le sont chacun par un seul sujet. Il semble également qu'un effet d'interaction entre la méthode de transplantation apparaisse : si l'ordre de préférence des systèmes est identique pour 3 sujets (J03, J06, J07), il y a une inversion de préférence entre deux systèmes pour 8 sujets. On peut donc conclure globalement que des variations importantes existent entre les sujets. Cependant, il faut tempérer cette conclusion, car l'essentiel de la variabilité provient d'un ou deux sujets. Si l'on ne considère pas ces sujets un peu atypiques, l'effet du facteur sujet reste modéré.

3.5. Discussion

3.5.1. *Aspects méthodologiques*

La première chose à noter est le coût élevé de telles expérimentations, à la fois en termes de temps pour les sujets (comme tout test perceptif) mais aussi en termes de temps de préparation des stimuli. Le contenu segmental des échantillons issus des différents systèmes de synthèse doit être soigneusement vérifié. Il est bien connu qu’en Français, le même texte orthographique peut donner lieu à des transcriptions phonétiques différentes, à cause en particulier du ‘e’ muet et des liaisons. Pour le test prosodique, ces effets doivent être neutralisés, afin de tester exactement le même contenu segmental. Un effort est aussi demandé aux participants, pour qu’ils fournissent une sortie prosodique standardisée, au format “.pho”. Sinon un alignement phonémique et le calcul de la prosodie doivent être effectués à posteriori. Une difficulté plus sérieuse provident des différences moyennes en fréquence fondamentale et en durée entre les systèmes. Des tests préliminaires ont montré que transplanter sur un substrat segmental directement une voix aigue sur des diphones a priori plus graves pouvait donner des distorsions très audibles, malgré un bon contour prosodique. Tous les systèmes ont donc été adaptés à un registre commun avant la transplantation.

En dépit d’un coût élevé de mise en œuvre, nous pensons que la transplantation prosodique est une méthode fiable pour comparer la prosodie calculée par un système avec celle d’autres systèmes ou une voix naturelle. Cette étude semble indiquer qu’il n’y a pas de différence notable entre les conditions « voix diphone » et « voix naturelle ». Ainsi ce type de méthodologie pourrait facilement être adapté à d’autres langues, grâce en particulier à MBROLA, un logiciel libre multilingue de synthèse par diphones. Notons finalement que le test de préférence par paires atteint rapidement ses limites lorsque le nombre de systèmes ou de phrases est élevé. Une stratégie possible est d’utiliser le test effectué comme une grille de référence et de sélectionner un ensemble de systèmes représentatifs. Une méthodologie alternative a été proposée et testée sur des systèmes de synthèse de l’allemand [BAI 06]. Elle consiste à placer à l’aide d’une interface graphique tous les stimuli générés pour une phrase sur un plan dont l’abscisse est une échelle de préférence.

3.5.2. *Conclusions*

Il faut remarquer la qualité prosodique remarquable obtenue par les deux meilleurs systèmes, qui sont préférés à la prosodie naturelle pour plusieurs phrases. Ce phénomène avait été observé de façon encore plus nette dans une précédente étude [PRU 04]: un locuteur réel peut produire une prosodie moins vivante et séduisante que celle d’un système de synthèse utilisant des règles prosodique soigneusement élaborées. Il faut bien sur tempérer notre observation qui ne porte

que sur des phrases isolées. Dans des textes plus longs ou plus construits, il est probable que la voix de synthèse serait identifiée au bout de quelques phrases.

Malgré un résultat global assez clair, si l'on observe l'évaluation des différentes phrases, d'importantes différences apparaissent. Il faut avouer que les bases perceptives et cognitives de l'évaluation prosodique sont quasiment totalement inconnues à ce jour. Il nous est impossible de comprendre pourquoi tel ou tel contour prosodique est préféré sur telle ou telle phrase. Ainsi une assez grande variabilité entre les phrases est prévisible pour ce genre de test, dont la seule façon de s'affranchir pour le moment est d'utiliser un nombre aussi grand que possible de phrases variées et longues.

4. Test d'intelligibilité : phrases sémantiquement imprédictibles

4.1. Introduction

Un des objectifs de la campagne d'évaluation était de comparer la qualité des systèmes à base de diphones et des nouveaux systèmes à base de concaténation d'unités longues. Si la qualité obtenue par ces derniers systèmes paraît plus fluide et naturelle, aucun test formel d'intelligibilité n'a été rendu public à notre connaissance, du moins pour le Français. Comme pour certaines applications (par exemple la lecture pour les malvoyants) l'intelligibilité, même à vitesse rapide, est plus importante que le « naturel », il est important d'évaluer si cette nouvelle technologie se révèle meilleure que les diphones en termes d'intelligibilité.

Le test d'intelligibilité mis en œuvre utilise des phrases sémantiquement imprédictibles (semantically unpredictable sentences, SUS), un paradigme expérimental permettant l'évaluation objective de l'intelligibilité au niveau des mots. Six systèmes de synthèse à partir du texte en Français, représentatifs du niveau actuel de qualité atteint, ont été mis à l'épreuve. Trois systèmes utilisent des diphones (systèmes D1, D2, D3) et trois sont des systèmes de sélection/concaténation (S1, S2, S3). Les résultats sont anonymes, mais les systèmes suivants ont participé au test: CRISCO, ICP, LIMSI-CNRS, Multitel, Acapela group (2 systèmes).

4.2. Corpus et protocole

Une nouvelle liste de 288 SUS a été construite [RAA 06]. Cette liste contient 4 types de structures syntaxiques (et non 5 comme dans [BEN 90]).

(1) adverbe déterminant. Nom_1 Verbe-*t*-pronom déterminant. $Noms_2$ Adjectif ?
(2) déterminant Nom_1 Adjectif Verbe déterminant Nom_2

(3) déterminant. Nom_1 $Verbe_1$ déterminant Nom_2 *qui* $Verbe_2$
(4) déterminant Nom_1 Verbe préposition déterminant Nom_2

La 3ème structure proposée dans [BEN 96] n'a pas été retenue, car elle ne contient que trois mots cibles au lieu de quatre (noms, verbes or adjectifs). Chaque bloc de test contient 12 phrases (un exemple est donné Tableau 6).

Pour obtenir des phrases et des blocs comparables, tous les mots pleins sont des monosyllabiques au singulier (au schwa final près), avec une fréquence d'usage élevée (d'après le lexique Brulex, [CON 90]). Les prépositions sont aussi des monosyllabes (ex. *sur*), et les déterminants les articles définis : *le, la* ou *l'* devant une voyelle. Les homographes hétérophones ont été soigneusement évités.

La loi brille par la chance creuse.	(4)
La classe gaie montre le frein.	(2)
Quand le lien signe-t-il l'onde pleine ?	(1)
Le test clair mange la haine.	(2)
L'or jaune porte le dôme.	(2)
Comment la soif lance-t-elle le bol proche ?	(1)
Le mur siffle la buée qui vole.	(3)
La banque dit la dinde qui plaît.	(3)
La terre dresse la boîte qui rage.	(3)
Où l'œuf cite-t-il le thé doué ?	(1)
Le nom luit sur le bras nu.	(2)
Le choix tape dans la queue close.	(2)

Tableau 6. *Exemples de SUS, avec leur type de structure syntaxique*

Une fois le matériau lexical défini, les listes de SUS sont générées aléatoirement, en retenant celles qui donnent la distribution de phonème la plus équilibrée par bloc. Cette répartition phonétique a été comparée (au sens du chi 2) à celle des lexiques Brulex [CON 90] et Lexique [NEW 04]. Toutes les listes ont été vérifiées et ajustées, afin d'éviter des séquences de mot sensées.

Les 288 phrases ont été prononcées par les 6 systèmes, ainsi que par un locuteur professionnel afin de donner une référence naturelle. Les équipes participantes devaient rendre les phrases de synthèse dans les heures suivant leur réception. Pour le test proprement dit, 3 blocs de 12 phrases ont été pris en compte par l'organisateur du test (ELDA) pour chaque système et la référence naturelle. De cette façon 22 des 24 blocs ont été utilisés. Les blocs étaient différents pour les différents systèmes, mais sans introduire de biais, puisque les blocs ont été construits pour être

comparables. Ainsi pour chaque auditeur participant au test n'avait à juger qu'une seule occurrence d'une phrase, et toutes les phrases ont été présentées (de façon aléatoire et à des niveaux égalisés). La durée du test était de 2 à 3 heures par sujet.

Une interface de test spécifique a été réalisée pour le test, afin de recueillir et de dépouiller automatiquement les réponses des sujets. Les sujets devaient transcrire orthographiquement les phrases entendues. Afin de s'affranchir des problèmes d'orthographe, les phrases écrites étaient phonétisées (par le système de [BOU 97] dont le taux d'erreur est inférieur à 1%) et comparée à la référence phonétique des phrases. Par exemple les réponses *voix* et *voie* sont équivalentes, puisque leur prononciation est (/vwa/). La campagne a été menée à ELSA (Paris) avec 20 sujets de langue maternelle française, âgés de 20 à 35 ans; sans problème auditif connu. Les sujets étaient payés pour effectuer la tâche, et sans familiarité particulière avec la synthèse de parole.

4.3. Résultats

Plusieurs façons de classer les résultats existent. Une première façon consiste a compter les erreurs par phrase. Une seconde façon, plus fine, consiste à compter les mots cibles mal compris. Un algorithme d'alignement dynamique des phonèmes est utilisé pour les comptages (`sclite`). Enfin un troisième niveau de finesse est obtenu en comptant le taux d'erreur au niveau des phonèmes pour les 4 mots cibles de chaque phrase.

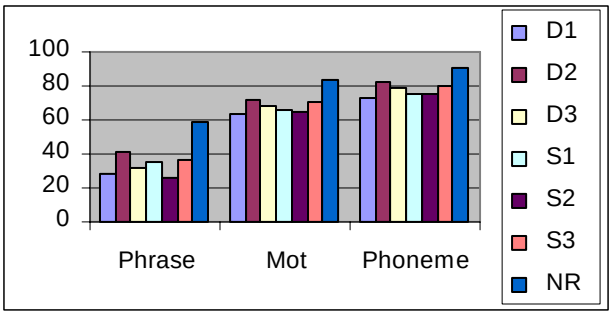


Figure 4 : *Pourcentages de transcriptions correctes pour le test SUS, par phrases, mots cibles et phonèmes, pour les 6 systèmes et la référence naturelle (NR)*

La Figure 4 montre que les résultats par phrase, mots cibles et phonèmes sont bien corrélés. Les meilleurs résultats dont obtenus pour la référence naturelle, puis les systèmes D2 et S3. Pour les phrases, ensuite S1 est meilleur que D3 et D1 est meilleur que S2, mais la tendance opposée est observée sur les mots et phonèmes.

Dans l'ensemble, les systèmes à diphones sont plutôt plus intelligibles que les systèmes par sélection/concaténation, mais l'ensemble des systèmes est nettement inférieur à la voix naturelle. On remarque également que nos pourcentages (entre 26.2% et 58.7%).sont très proches de ceux obtenus par [BEN 90], de 28.6% à 58.1%. Cela donne l'impression d'une faible évolution de l'intelligibilité des systèmes depuis une quinzaine d'années.

5. Essai d'opinion directe

5.1. Méthodologie

Le test d'intelligibilité par les SUS a été confronté à un test plus global utilisant des échelles de catégorie absolue (ACR) sur 5 points. Un point supplémentaire est ajouté à chaque extrémité de l'échelle pour éviter un effet de saturation [MÖL 00]. Six catégories ont été retenues, en plus de l'opinion moyenne (MOS) : compréhension, agreement, (non) monotonie, naturel, fluidité et prononciation. Ces catégories sont issues d'une adaptation en Français des critères d'évaluation proposes dans le projet Vermobil [KRA 95]. Le Tableau 7 montre les consignes données aux sujets. Afin d'encourager l'utilisation de toute la gamme de jugements, l'échelle effectivement présentée aux sujets était continue.

MOS (très mauvais — très bon)	Comment appréciez-vous globalement ce que vous venez d'entendre ?
Compréhension (très difficile— très facile)	Comment décririez-vous la facilité à comprendre le message ?
Agrément (très désagréable — très agréable)	Comment décririez-vous cette voix ?
Monotonie (très monotone — très varié)	Évaluez le caractère monotone ou varié de ce que vous venez d'entendre
Naturel (très artificiel — très naturel)	Comment apprécieriez-vous le naturel de ce que vous venez d'entendre ?
Fluidité (très haché — très fluide)	Comment appréciez-vous le côté haché ou fluide de l'élocution ?
Prononciation (sérieux problèmes — aucun problème)	Avez-vous remarqué des problèmes de prononciation ?

Tableau 7. *Questions présentées aux sujets pour le test ACR.*

Comme pour le test SUS, les participants devaient synthétiser des centaines de phrases dans un délai rapide. Le corpus de phrases est extrait d'EUROM 1 [CAM 98], qui a été collecté dans le cadre des projets Multext (Multilingual Text Tools and

Corpora) et Esprit 2589/SAM (Multilingual Speech Input/Output Assessment Methodology and Standardisation). Un ensemble de 20 passages d'environ 20 secondes (5 phrases cohérentes sur un thème donné) a été synthétisé par les 6 systèmes, en plus de la référence naturelle d'EUROM1 (locuteur FB).

Les sujets devaient évaluer des paragraphes lus par les 6 systèmes et la voix naturelle. L'ordre de passage des systèmes était aléatoire. Le test, qui a également pris place à ELDA durait environ 3 heures, avec des pauses toutes les 20 minutes. Une interface spécifique a été mise au point pour rendre le test plus agréable. 17 sujets, de profils identiques à celui du test SUS ont participé aux tests.

5.2. Résultats

Les résultats montrent (Figure 5) que la référence naturelle obtient les plus hauts scores (au dessus de 4), devant S3, puis D3. S3 passe le seuil de 4/5 pour la "compréhension" et "prononciation". C'est aussi le seul système à passer le seuil des 3/5 dans les autres catégories. C'est donc de loin le système le mieux apprécié dans ce test, en moyenne un point au dessus du second système, D3. D3 est le seul système qui obtient des notes entre 2 et 4 dans toutes les catégories, y compris le « naturel ». Il semble que les systèmes par concaténation/sélection ne sont pas toujours mieux appréciés que les systèmes par diphones. Ainsi D1 et S1 d'un côté, D2 et S2 de l'autre, sont très comparables, pour le MOS et le "naturel". Le pire score revient à D1 pour cette dernière catégorie. Cependant, S1 et S2 sont les pires systèmes pour respectivement « l'agrément » et la « prononciation ». D2 est plus compréhensible (mais aussi plus monotone) que S1 et S2.

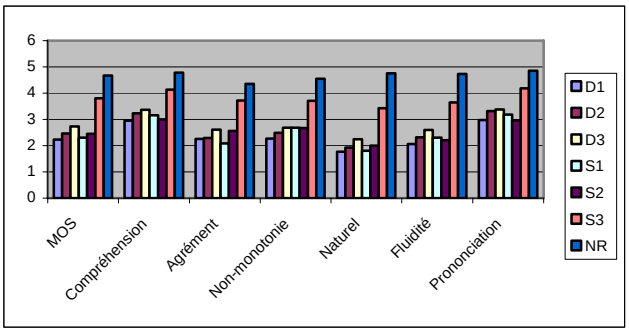


Figure 5 *Résultat du test ACR pour les 6 systèmes et la référence naturelle (NR). Les scores sont échelonnés entre 1 et 5, 5 étant la meilleure note.*

5.4. Conclusion

L'intelligibilité mesurée par le test SUS et la compréhension mesurée par le test ACR ne concordent pas, bien que ces deux concepts puissent paraître proches. Un

système par diphone (D2) l'emporte dans nos tests pour l'intelligibilité SUS, alors que c'est un système par sélection (S3) qui est nettement préféré dans le test ACR. Malgré un abord plus séduisant, il semble donc que les systèmes par sélection ne soient pas systématiquement meilleurs globalement que les systèmes par diphones au moins d'un point de vue fonctionnel, qui est celui de l'intelligibilité. D'un point de vue hédonique, celui de l'agrément, ils semblent l'emporter.

6. Discussion et conclusion

Une question intéressante est bien sûr la corrélation entre les différentes mesures de qualité obtenues. Par exemple une analyse statistique serait nécessaire pour déterminer la corrélation entre intelligibilité et qualité globale perçue. Cependant, on peut déjà avancer que les différents aspects de la qualité correspondent à différents usages des systèmes de synthèse. Un système hautement intelligible, même avec une qualité de voix plutôt faible serait un meilleur candidat pour un système de lecture pour les malvoyants. Au contraire, pour un système de dialogue intégré, la qualité de voix représente l'identité sonore, la carte de visite vocale en quelque sorte, du système. Cette qualité, dans des échanges assez prévisibles, peut l'emporter sur la pure intelligibilité dans ce cas.

Les résultats globaux pour l'évaluation prosodique donne le classement suivant: NR > S1 > D1 > S2 > D2 > D3. Ce classement est en accord avec celui obtenu par les tests d'opinion et les autres jugements catégoriel absolus. Comme les tests ACR sont essentiellement une évaluation de la qualité globale perçue, il semble donc que la qualité prosodique perçue soit fortement corrélée avec la qualité globale perçue.

A l'inverse, il semble que les aspects plus linguistiques de la qualité perçue, comme la transcription graphème-phonème, soient relativement peu corrélés avec la qualité prosodique. On peut le comprendre en première approximation, car les interactions prosodie/phonétisation, si elles existent certainement (par exemple les coupes syllabiques) ont été neutralisées dans nos tests. On peut penser que ce sont des effets de second ordre, mais de nouvelles recherches et des tests spécifiques sont nécessaires, nos tests étant insuffisants à cet égard. De même la qualité de transcription graphème phonème et la qualité globale ou intelligibilité semblent dans nos tests relativement décorrélées. Ceci est probablement dû à la différence entre les protocoles de tests : la conversion graphème phonème a été testée avec un grand échantillon (8000 mots ici) pour faire apparaître les effets statistiques. Au contraire, les tests perceptifs ont porté sur peu de phrases. L'influence de la phonétisation dans ce cas était forcement affaiblie, et largement masquée par l'influence de la qualité d'articulation, de la prosodie, de la qualité vocale. Cette apparente contradiction s'explique donc par les paradigmes d'évaluation utilisés et apparaît probablement comme un artefact.

Une autre question, pour des travaux futurs, question qui fait l'objet de recherches et d'investissement importants, est celle de l'expression vocale, pour passer d'un style de lecture de textes à une parole plus intime et expressive, avec un style vocal plus conversationnel (par exemple le système Laughter [CAM 05]). Nous n'avons pas considéré dans cette étude ce changement de paradigme, des "machines à lire" aux véritables "machines parlantes". En effet, cette évolution passionnante devra s'accompagner d'une réflexion très sérieuse sur l'environnement, l'usage et le profil des utilisateurs, afin d'éviter un divorce entre applications et évaluation.

Notons finalement que nous avons tenu dans cette campagne à conserver autant que se peut la diversité des approches, techniques et ressources employées. Cette approche doit compléter le travail d'évaluation des techniques de synthèse où les ressources sonores sont imposées aux participants [BLA 06]. Si ces campagnes nécessitent un effort de développement beaucoup plus important, elles permettent aux développeurs de systèmes d'avoir accès à un nombre important de ressources validées et de confronter les divers modules opérant par apprentissage ou stockage de références à des couples d'entrées-sorties comparables (cf. l'évaluation de systèmes d'apprentissage automatique de la prosodie par [RAI 04]).

En plus des résultats d'évaluation proprement dits, la campagne EVASY a produit un ensemble de ressources qui sont actuellement distribuées par ELDA, et qui, nous l'espérons, sera utile pour de futures campagnes d'évaluation de la synthèse de parole en Français.

Bibliographie

[dAL 01] D'ALESSANDRO C. & TZOUKERMANN E., *Synthèse de la parole à partir du texte, Traitement Automatique des Langues,* 42(1), Hermès, Paris, 2001.

[dAL 04] D'ALESSANDRO C., « L'évaluation des systèmes de synthèse de la parole », in S. Chaudiron (Dir.) *L'évaluation des systèmes de traitement de l'information*. 9,. Hermès, Paris, p. 215-239, 2004.

[dAL 98] D'ALESSANDRO C. AND B3 PARTNERS, « Joint evaluation of Text-To-Speech synthesis in French within the AUPELF ARC-B3 project », *Proceedings 3rd International Workshop on Speech Synthesis*, Jenolan Caves, Australie, p. 11-16, 1998.

[dAL 01-2] D'ALESSANDRO C. . « 33 ans de synthèse de la parole à partir du texte: une promenade sonore (1968-2001) ». *Traitement Automatique des Langues (TAL),* 42-1, Hermès, Paris, p. 297-321, avec un disque compact audio de 62 mn, 2001.

[BAG 98] BAGSHAW, P. « Phonetic transcription by analogy in text-to-speech synthesis: Novel word pronunciation and lexicon compression », *Computer Speech and Language*, 12(2), p. 119-142., 1998

[BAG 92] BAILLY, G., T. BARBE & H. WANG « Automatic labelling of large prosodic databases: tools, methodology and links with a text-to-speech system », *Talking Machines: Theories, Models and Designs*. G. Bailly and C. Benoît. Amsterdam, North-Holland, p. 323-333, 1992.

[BAI 06] BAILLY, G. & I. GORISCH « Generating German intonation with a trainable prosodic model », InterSpeech, Pittsburgh, p. 2366-2369, 2006.

[BLA 98] BLACK, A., LENZO, K. & PAGEL, V. 1998. « Issues in building general letter-to-sound rules »", *3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, p. 77-80, 1998.

[BLA 06] BLACK, A., K. TOKUDA, S. KING, T. HIRAI, M. PICHENY & S. NAKAMURA « Blizzard Challenge ». satellite workshop of Interspeech, Pittsburgh, 2006.

[BAR 99] BARTKOVA, K. & JOUVET, D. « Language based phone model combination for ASR adaptation to foreign accent », *ICPhS*, San Francisco, p. 1725-1728, 1999.

[BEC 97] BÉCHET, F. & EL-BÈZE, M. « Automatic assignment of part-of-speech to out-of-vocabulary words for text-to-speech processing », *Eurospeech*, Rhodes, p. 983-986, 1997.

[BEC 00] BECHET, F. & YVON, F. « Les Noms Propres en Traitement Automatique de la Parole », *Traitement Automatique des Langues* 41(3), p. 671-707, 2000.

[BEN 90] BENOÎT, C. «An intelligibility test using semantically unpredictable sentences: towards the quantification of linguistic complexity », *Speech Communication*, 9(4), p. 293–304, 1990.

[BEN 96] BENOÎT, C., GRICE, M., HAZAN, V. , «The SUS test: a method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences ». *Speech Communication*, 18(4), p. 381–392, 1996.

[BOE 02] BOEFFARD O., D'ALESSANDRO C. (2002) « Synthèse de la parole » *In:J. Mariani (Dir.) Traitement automatique du langage parlé, Vol.1:Analyse, synthèse et codage de la parole".. Hermès Science Publications;Lavoisier, Pari*s, p. 115-154, 2002.

[BOU 97] BOULA DE MAREÜIL, P. « *ÉTUDE LINGUISTIQUE APPLIQUEE A LA SYNTHESE DE LA PAROLE A PARTIR DU TEXTE* », THESE DE DOCTORAT, UNIVERSITE PARIS XI, ORSAY, 1997.

[BOU 00] BOULA DE MAREÜIL, P., YVON, F., D'ALESSANDRO, C., AUBERGÉ, V., VAISSIÈRE J., & AMELOT, A. « A French phonetic lexicon with variants for speech and language processing », *LREC*, Athens, p. 273-276, 2000.

[BOU 05] BOULA DE MAREÜIL, P. D'ALESSANDRO, C., BAILLY, G., BÉCHET, F., GARCIA, M.-N., MOREL, M., PRUDON, R., VÉRONIS, J. "Evaluating the pronuncation of proper names by four French grapheme-to-phoneme converters." In *Proc. Eurospeech'05, (Interspeech)*, Lisbon, p. 1521–1524, 2005.

[BOU 06-2] BOULA DE MAREUIL, P., D'ALESSANDRO, C., RAAKE, A., BAILLY, G, GARCIA, M.N., MOREL' M. (2006), "A Joint intelligibility evaluation of French text-to-speech systems: the EvaSy SUS/ACR campaign" In *Proc of LREC*, Genoa. 2006.

[CAM 05] CAMPBELL, N., KASHIOKA, H., OHARA, R., "No Laughing Matter." In *Proceedings of the Ninth European Conference on Speech Communication and Technology (Interspeech)*.Lisbon, p. 465–468, 2005.

[CAM 98] CAMPIONE, E. & VERONIS, J. A multilingual prosodic database. In *Proceedings of the Fifth International Conference on Spoken Language Processing*, Sydney, p. 3163–3166, 1998.

[CON 90] CONTENT, A., MOUSTY, P., RADEAU, M. (1990), BRULEX : Une base de données lexicales informatisée pour le Français écrit et parlé. *L'Année Psychologique*, 90, p. 551–566, 1990.

[CHU 86] CHURCH, K., « Stress Assignment in Letter to Sound Rules for Speech Synthesis », *IEEE-ICASSP*, Tokyo, p. 2423-2426, 1986.

[DAM 01] DAMPER, R.I., STANBRIDGE, C.Z. & MARCHAND, Y. « A Pronunaciation-by-Analogy Module for the Festival Text-to-Speech Synthesiser », *4th ISCA Workshop on Speech Synthesis*, Pitlochry, p. 97-102, 2001.

[DIV 97] DIVAY, M. & VITALE, A.J. « Algorithms for Grapheme-Phoneme Translation for English and French: Applica-tions », *Computational linguistics*, 23(4), p. 495-524, 1997.

[DUT 96] DUTOIT, T., PAGEL, V., PIERRET, N., BATAILLE, F., VAN DER VRECKEN, O. "The MBROLA project: towards a set of high quality speech synthesizers free of use for non commercial purposes" In *Proc. of ICSLP*, Philadelphia, p. 1393–1396, 1996.

[GAR 06] GARCIA M-N, D'ALESSANDRO C., BAILLY G, BOULA DE MAREUIL P., MOREL M. "*A joint prosody evaluation of French text-to-speech systems: the EvaSy Prosody campaign*", *LREC'06*, Genoa, 2006.

[GAR 94] GARY-PRIEUR, M.-N. *Grammaire du nom propre*, Presses Universitaires de France, Paris, 1994.

[GIB 98] GIBBON, D., MOORE, R. & WINSKI, R. (Eds). *Handbook of Standards and Resources for Spoken Language Systems*, Mouton de Gruyter, Berlin, 1998.

[KRA 95] KRAFT, V. & PORTELE, T. (1995), Quality Evaluation of Five German Speech Synthesis Systems, *Acta acustica*, 3, p. 351–365, 1995.

[LAM 91] LAMEL,, L.F., GAUVAIN, J.-L., ESKÉNAZI, M., BREF, "a Large Vocabulary Spoken Corpus for French" in *Proc of Eurospeech*. Genova, p. 505–508, 1991.

[LER 04] LEROY, S. *Le Nom propre en français*, Ophrys, Gap/Paris, 2004.

[LLI 01] LLITJOS, A.F. & BLACK, A.W. « Knowledge of Language Origin Improves Pronunciation Accuracy of Proper Names », *Eurospeech*, Aalborg, p. 1919-1923, 2001.

[MAE 93] MAES, P. *La prononciation des langues européennes*, Éditions du centre de formation et de perfectionnement des journalistes, Paris, 1993.

[MAP 04] MAPELLI V., NAVA M., SURCIN S., MOSTEFA D., CHOUKRI K., "*Technolangue: A Permanent Evaluation and Information Infrastructure*", *LREC'04*, Lisbonne, mai 2004.

[MAL 97] MALFRÈRE, F, DUTOIT, T "High Quality Speech Synthesis for Phonetic Speech Segmentation", Proc. Eurospeech '97, p. 2631-2634, 1997.

[MÖL 00] MÖLLER, S., *Assessment and Prediction of Speech Quality in Telecommunications*, Kluwer Academic Publishers, Boston. New, B., Pallier, C., Brysbaert, M.,, Ferrand, L.

(20004), Lexique 2: A New French Lexical Database. *Behavior Research Methods, Instruments, & Computers*, 36(4), p. 516–524, 2000.

[NEW 04] NEW, B., PALLIER, C., BRYSBAERT, M., FERRAND, L. "Lexique 2 : A New French Lexical Database" *Behavior Research Methods, Instruments, & Computers*, 36 (3), p. 516-524, 2004

[ONO 95] THE ONOMASTICA CONSORTIUM. « The ONOMASTICA interlanguage pronunciation lexicon », *Eurospeech*, Madrid, p. 829-832, 1995.

[PRU 04] PRUDON R., D'ALESSANDRO C., BOULA DE MAREUIL P., « Unit selection synthesis of prosody : evaluation using diphone transplantation », in S. Narayanan, A. Alwan "*Text to speech synthesis : new paradigms and advances*". Chap.10, p. 203-217. Prentice Hall PTR, New Jersey, 2004.

[RAI 04] RAIDT, S., BAILLY, G., HOLM, B., AND MIXDORFF, H. (2004) « *Automatic generation of prosody: comparing two superpositional systems* », *Speech Prosody*, Nara, Japan,. p. 417-420, 2004.

[RAA 06] RAAKE, A.& KATZ, B.F., « SUS-based Method for Speech Reception Threshold Measurement in French » In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa, 2006.

[VIT 91] VITALE, T. « An Algorithm for High Accuracy Name Pronunciation by Parametric Speech Synthesizer », *Computational Linguistics*, 17(3), p. 237-276, 1991.

[YVO 97] YVON, F. « *Prononcer par analogie : motivations, formalisation et évaluation* », Thèse de doctorat, ENST, Paris, 1997.

[YVO 98]YVON F.' BOULA DE MAREÜIL P.,' D'ALESSANDRO C.' AUBERGE V.' BAGEIN M.' BAILLY G.' BECHET F.' FOUKIA S.' GOLDMAN J.F.' KELLER E.; O'SHAUGHNESSY D., PAGEL, V. SANNIER F., VERONIS J.; ZELLNER B. « Objective evaluation of grapheme-to-phoneme conversion for text-to-speech synthesis in French », *Computer Speech and Language*, 12(4), p. 393-410, 1998.