CILL 30 (2004) n° 1-3, 153-180

SYNTHÈSE PAR SÉLECTION : PROSODIE, DIALOGUE ET QUALITÉ VOCALE

Christophe D'ALESSANDRO, Philippe BOULA DE MAREÜIL & Romain PRUDON
LIMSI-CNRS

1. INTRODUCTION

Le domaine de la prosodie et du dialogue offre un bon exemple des difficultés que peut rencontrer la collaboration entre informatique et linguistique, du fait du statut même de ces deux disciplines : la linguistique vise à décrire la richesse de la langue, alors que les réalisations informatiques tendent inévitablement à des simplifications réductrices. Ceci vaut également pour la synthèse de la parole à partir du texte, qui est devenue un élément clef des interfaces de communication homme-machine.

Le dialogue est souvent ressenti, en linguistique, comme une application de la pragmatique (i.e. l'étude de l'influence de la situation sur le sens du discours des intervenants, voire des informations qui ne sont pas dans le discours mais qu'il faut connaître pour que celui-ci soit compréhensible). Quant à la prosodie, elle est un des grands enjeux en synthèse de la parole à partir du texte, qui a atteint aujourd'hui un degré d'acceptabilité globale très satisfaisant, mais pèche encore par manque de naturel, d'intelligibilité ou d'agrément. La synthèse classique par diphones avec un module syntactico-prosodique par règles produit souvent une mélodie stéréotypée, monotone, répétitive et génératrice de fatigue. La synthèse par sélection dynamique permet en introduisant une certaine variabilité de pallier cet inconvénient, pourvu que l'on dispose d'une base acoustique suffisante et adéquate, et que les paramètres du système soient correctement ajustés (PRUDON & D'ALESSANDRO 2001). Dans cette nouvelle génération de systèmes, en effet, les unités concaténées ne sont plus

extraites d'une base de diphones mono-représentés, mais de corpus de parole beaucoup plus volumineux, à l'aide de techniques dérivées de la reconnaissance de la parole (CAMPBELL 1998 ; BALESTRI *et al.* 1999 ; BREEN 2000 ; COORMAN *et al.* 2000).

Mise au défi de synthétiser un extrait du dialogue proposé dans le cadre du Colloque Prosodie de Genève, il faut pourtant bien reconnaître que la synthèse de la parole est incapable d'atteindre un objectif de ce genre, c'est-à-dire la synthèse réaliste d'un dialogue véritable. D'ailleurs, comment pourrait-on atteindre un but qui n'est pas véritablement visé dans l'état actuel des techniques ? Synthétiser de la parole lue, éventuellement des annonces ou des réponses dans le cadre de systèmes d'information, tel est le cadre d'étude et d'évaluation des systèmes de synthèse de parole actuels. Cependant, la confrontation avec cette cible hors de portée, loin d'être absurde, permet de susciter un effort d'analyse et de comparaison, entre les résultats de la synthèse et l'ambition d'une synthèse plus « réaliste », « spontanée » ou « naturelle ».

Cette analyse s'organise en trois temps. Dans une première partie, on s'intéressera à la prosodie obtenue en synthèse de parole par sélection et concaténation, comparée à la synthèse par règles de la prosodie et à la prosodie recopiée de parole naturelle. Une seconde partie analyse d'un point de vue linguistique quelques aspects du dialogue radiophonique mal ou pas du tout rendus par la synthèse, qui mériteraient un traitement automatique spécifique : des pistes sont proposées. La troisième partie analyse les effets expressifs liés à la prosodie au sens large du terme, en incluant le jeu de la source et de la qualité vocales. Au-delà de la prosodie au sens étroit (mélodie, durée, intensité), ces aspects sont souvent négligés en synthèse de la parole, ce qui pourrait bien être en partie la cause du manque de naturel que l'on constate à l'audition des différents systèmes.

2. SYNTHÈSE PAR SÉLECTION

2.1. Principe de la synthèse

Le principe des systèmes de synthèse de la parole actuels est de sélectionner, dans une importante base de données de parole lue, des unités acoustiques de tailles variables, et de les concaténer pour reconstruire un signal de parole le plus naturel possible. Le but est d'utiliser au maximum les informations présentes dans la base. L'unité minimale utilisée dans le système développé au LIMSI (SeLimsi) est le diphone (PRUDON & D'ALESSANDRO 2001). Cependant dans la pratique, nous cherchons à maximiser cette taille ; c'est justement la longueur des chaînes qui

assure en partie la qualité de la parole synthétisée. L'architecture générale de notre système est donnée en figure 1 ; elle se décompose en trois parties principales.

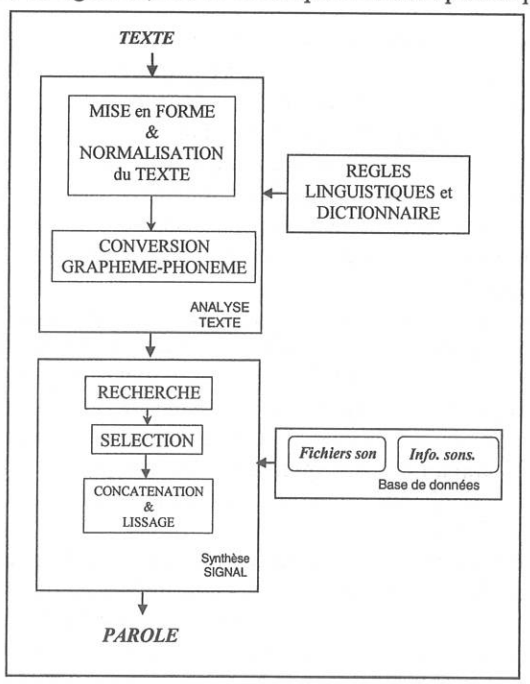

Figure 1. Architecture du système de synthèse par sélection du LIMSI

- **L'analyseur de texte** : cette partie est pour l'essentiel commune aux précédents systèmes de synthèse ; elle réalise la normalisation (mise en forme) du texte et la conversion graphème-phonème. L'entrée est la phrase que l'on veut synthétiser, la sortie sa transcription phonétique, enrichie des frontières entre les mots et les syllabes.
- **La base de données** : elle contient les données sonores (LAMEL *et al.* 1991) ainsi que toutes les informations utiles pour réaliser la sélection soit, pour chacun des phonèmes, la position dans le mot et la syllabe d'origine, la fréquence fondamentale, la durée, l'énergie et le contexte phonétiques. Elle est composée en moyenne de 700 phrases phonétiquement équilibrées, ce qui représente une durée d'environ 1 heure 30.
- **Le synthétiseur de signal** : à partir de la chaîne phonétique, nous recherchons dans la base toutes les combinaisons possibles pour les traiter, puis

nous sélectionnons celles qui correspondent le plus à ce que nous voulons ; enfin les unités sont concaténées.

Dans ce type de système, la qualité finale de la parole synthétisée est très dépendante de la sélection ; cette partie représente le cœur du système. Pour la réaliser, nous utilisons deux fonctions de coût : le coût de cible et le coût de concaténation. Le premier établit une distance entre la phrase cible (texte d'entrée) et ce qui est disponible dans la base ; le second quantifie la qualité de la jonction entre deux unités étudiées. À l'aide de ces deux fonctions, nous parcourons toutes les combinaisons possibles, et nous déterminons quelle est la meilleure. Chacune des fonctions de coût est décomposée en plusieurs critères pondérés, qui permettent de bien réaliser la sélection.

Le coût de cible est composé de trois critères : la place dans le mot, la place dans la syllabe et la durée. Avec les deux premiers, le but est de sélectionner des diphones ayant la même place dans la syllabe et le mot d'origine que dans la phrase cible. Nous cherchons aussi à respecter la place dans la phrase. Ainsi, pour respecter l'allongement de fin de phrase, nous forçons le système pour qu'il sélectionne prioritairement des diphones correspondants. Le critère de durée est quant à lui utilisé pour assurer la qualité du rythme. Le coût de cible est donc utilisé pour garantir la prosodie de la phrase.

Le coût de concaténation est composé de cinq critères : la maximisation des chaînes, le respect du contexte phonétique et la minimisation des écarts de F_0, de durée et d'énergie entre les phonèmes à droite et à gauche de la jonction. Le plus important de ces critères est de maximiser les chaînes, en considérant la jonction entre deux diphones consécutifs dans la base de données comme parfaite. Ainsi les unités sélectionnées sont en moyenne des triphones.

L'algorithme de sélection est une adaptation de celui de Viterbi. Celui-ci a été développé dans le contexte des grammaires stochastiques, et est couramment utilisé dans les systèmes de reconnaissance vocale. Il est utilisé pour calculer le meilleur chemin au sein d'une importante combinatoire en calculant localement les coûts. Ainsi, à chaque étape, on calcule le coût de concaténation d'un diphone avec les précédents et sa distance avec la cible. Pour mieux expliquer le principe, nous donnons un exemple simplifié de sélection. La séquence à synthétiser est le mot *livre*, précédé et suivi d'un silence, sa transcription dans l'alphabet phonétique utilisé au LIMSI est la suivante /.livr./. Pour plus de clarté, nous simplifions les données en limitant le nombre d'unités candidates. Supposons qu'il y ait trois exemplaires des diphones /.l/, /li/, /vr /, et deux des diphones /iv/ et /r./. Dans un cas réel, il y a beaucoup plus de possibilités – les diphones en caractères gras sont les cibles, les autres sont les résultats de la recherche. Le premier graphique de la figure 2 représente le calcul des coûts pour le premier diphone /li/, C_t pour celui de cible, C_c pour celui de concaténation. Tous les chemins sont ainsi calculés localement, en

mémorisant les étapes précédentes (voir le deuxième graphique de la figure 2). La dernière étape consiste à chercher l'unité qui a le plus petit coût final et le plus petit chemin utilisé pour l'obtenir – celui-ci est grisé sur le dernier graphique.

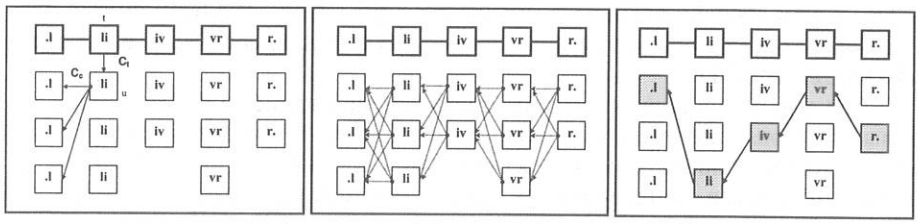

Figure 2. Exemple de sélection pour le mot livre

La sélection réalisée, il ne reste plus qu'à concaténer les diphones. Cette partie est réalisée en deux étapes : dans un premier temps, on recherche le point de concaténation optimum dans le signal (par intercorrélation) ; et dans un deuxième temps, on lisse ce point dans le domaine temporel, pour éliminer les discontinuités du signal.

2.2. Synthèse par sélection et prosodie

La synthèse peut aussi être vue comme un outil de calcul de la prosodie, qui peut ensuite être recopiée et synthétisée par concaténation de diphones. Elle peut même contribuer à valider certaines hypothèses et à modéliser l'*attente* en matière de prosodie : la perception d'une discontinuité en un point donné de la chaîne parlée peut en effet être due au fait que l'on escomptait un segment plus haut ou plus bas, plus long ou plus bref, créant un autre mouvement. Et si l'on constate un trop grand nombre d'erreurs imputées à des hétérogénéités de rythme, celui-ci peut être prédit par règles et croisé avec des valeurs de hauteur sélectionnées.

Une évaluation récente (PRUDON *et al.* 2002) a été menée au LIMSI, recopiant sur une même base de diphones utilisant MBROLA (DUTOIT *et al.* 1996) la prosodie d'une voix naturelle, celle que génère notre système par sélection (à partir de la même voix masculine), et celles que prédisent deux systèmes fondés sur des règles syntactico-prosodiques : l'un mis au point au LIMSI (BOULA DE MAREÜIL *et al.* 2001a), l'autre au CNET puis à Elan (BOULA DE MAREÜIL *et al.* 2001b). Des tests perceptifs par paires ont montré qu'aucune préférence claire ne se dégage en faveur de telles ou telles durées (croisées avec une même mélodie, pour un débit comparable). Dans l'ensemble en revanche, le système par sélection se situe entre les deux systèmes à base de règles, et la prosodie du meilleur système à base de règles (CNET/Elan) est jugée aussi bonne, voire légèrement meilleure, que la prosodie naturelle.

Bien que la tendance à la baisse du coût de la mémoire des ordinateurs permette une approche reposant sur des bases de données toujours plus volumineuses, on peut en déduire que les règles de calcul de la prosodie ont encore de beaux jours devant elles, et qu'il reste du travail en la matière. Mais par ailleurs, le LIMSI dispose d'un important vivier de voix à travers un gros corpus de parole enregistrée et étiquetée (LAMEL *et al.* 1991).

Cet aspect multilocuteur est particulièrement bien adapté au dialogue, que nous nous proposons ici d'examiner – avec ses phénomènes d'hésitations, de répétitions de mots-outils et autres faux départs. Alors qu'il est long et coûteux d'élaborer un modèle d'interface syntaxe-prosodie (lequel demande une certaine expertise en matière de regroupements et de hiérarchisations syntaxiques), il est facile et rapide par sélection de changer de style, dans la mesure où les tours de parole peuvent être repérés automatiquement dans les textes. D'où aussi une étude de la prosodie qui déborderait du cadre de la phrase (objet de la syntaxe) et qui irait jusqu'au niveau du paragraphe – unité d'analyse de la parole spontanée MOREL & DANON-BOILEAU (1998). C'est à ce stade qu'une étude du dialogue s'impose[1].

3. ANALYSE DU DIALOGUE

Cette partie analyse en 3.1 le schéma du dialogue et les actes illocutoires réalisés. Pour l'analyse de la mélodie, nous nous sommes abondamment servis de la stylisation issue des recherches de MERTENS (1987) et D'ALESSANDRO & MERTENS (1995). La section 3.2 étudie un certain nombre de phénomènes typiques de la parole non lue, d'opérations d'extraction et de focalisation, ainsi que d'effets de listes, qui sont autant de problèmes pour la synthèse vocale. Enfin la section 3.3 revient sur ces accents d'insistance qui sont liés plus généralement à la fonction expressive de la prosodie. La discussion qui s'ensuit plaide en faveur d'une approche à base de balises, pour appliquer différents types de prosodie.

3.1. Schéma de l'interview et actes illocutoires réalisés

Il s'agit d'un dialogue « promotionnel », entre deux interlocutrices qui ont un message à faire passer, et qui vraisemblablement sont d'accord sur celui-ci avant de démarrer l'échange. On a le schéma classique :

1. introduction – jusqu'à BG3 inclus[2] ;
2. mise en situation, explication du pourquoi du livre ;

---

[1] On peut retrouver la lecture du corpus proposé par notre synthétiseur sur Internet, à l'adresse : http://www.limsi.fr/Individu/cda/colgen02.html et sur le CD-ROM joint à ce volume.

[2] Les références au corpus oral renvoie à la transcription orthographique de celui-ci (sur le CD-ROM) et aux tours de parole respectifs des interlocutrices (BG1, RF2, etc.).

3. conclusion (?) – le dialogue est coupé.

L'animatrice Roselyne Fayard (RF) dirige le dialogue et l'oriente : les *vous* dominent, tandis que ce sont les *je* qui sont les plus nombreux chez l'écrivaine Benoîte Groult (BG). BG répond par davantage de *oui* que de *non* ; la connivence entre les deux interlocutrices est marquée par la complétion, en tant qu'acte de dialogue, notamment aux énoncés BG1-RF2 et RF3-BG3 (rephrasés avec une surenchère marquée par *même*).

Bien que l'interview soit relativement consensuelle, une première réorientation du dialogue est initiée par le *mais* de RF4, qui, renforcé par le connecteur *quand même*, arrive de façon quelque peu incongrue. Cassure du rythme et surtout du thème (sans progression) – si l'on suit le point de vue des deux axes du modèle genevois (ROULET *et al.* 1985) : le nouveau thème (l'interrogatif *qu'est-ce qui*) est d'ailleurs très haut sur l'échelle mélodique. Une deuxième transition brutale est initiée par RF7, délimitant des paquets à l'intérieur desquels il n'y a qu'une petite progression thématique – ce qui pourrait être mis en évidence par les fréquences lexicales. Tout se passe comme si RF avait en tête un dialogue précis et faisait tout pour que l'interaction se déroule à son idée. C'est elle qui décide des thèmes abordés et du temps à y consacrer, ce qui implique des changements très nets.

Le modèle genevois, rappelons-le, a permis de dégager que tout dialogue ne peut se dérouler que dans deux directions complémentaires : on a affaire soit à un dialogue régissant, soit à un dialogue incident. Le premier vise à réaliser les buts successifs des locuteurs ; le second résout les difficultés personnelles qui gênent le déroulement du premier (incompréhension, manque d'informations, etc.). Ainsi, deux contraintes, issues des maximes de Grice, déterminent dans une large mesure la structure des discours en situation : la *complétude interactionnelle* tend à faire progresser le dialogue vers la satisfaction des deux intervenants ; et la *complétude interactive* tend, lorsqu'il y a désaccord, à la résolution des conflits pour revenir au dialogue principal et envisager de satisfaire la complétude interactionnelle. Le problème est de rendre compte d'un dialogue au cours de son déroulement, de façon dynamique.

L'école de Genève incarne un courant majeur dans l'étude encore récente du dialogue en tant que tel ; elle a rendu des services considérables au traitement automatique (LUZZATI 1995, SABAH *et al.* 1997 ; MINKER & BENACEF 2001). Il s'agit de construire un système descriptif organisé, aussi structuré que possible, tout en partant de conversations authentiques, avec leurs interruptions, leurs reprises (de souffle) et autres régulateurs du type *hm*. Il en ressort un modèle fondé à la fois sur le courant conversationnel (le dialogue est conçu comme une négociation), sur la pragmatique anglo-saxonne et sur la théorie de l'énonciation de Ducrot (ROULET *et al.* 1985). L'énonciation est vue comme un acte destiné à communiquer, produit dans une situation qu'il modifie. L'intervention comporte différents actes de langage

hiérarchisés ; elle s'intègre dans un échange, car une intervention initiative appelle une intervention réactive.

Si l'on s'intéresse aux actes illocutoires réalisés, le présent dialogue n'est pas caractérisé par les demandes et les injonctions, comme un dialogue finalisé. Le jeu de questions-réponses implique que ces dernières soient des assertions, ayant pour but de rendre le contenu propositionnel des mots conforme au monde. Il en va tout autrement des questions, qui prosodiquement génèrent une lutte entre deux forces : l'interrogation qui tend à élever la mélodie et la fin de phrase qui tend à l'abaisser. Il est fait grand cas en pragmatique de ce « *high rising terminal tune* ». Ici, l'animatrice demande une coopération, un retour de son interlocutrice, ce qu'illustre l'absence de descente finale dans l'énoncé de RF1, qui est pourtant grammaticalement déclaratif (« c'est à peine croyable… », où les points de suspension transcrivent l'attente de quelque chose, d'une suite ou d'une confirmation). RF prend une voix de proximité, au risque de perdre la position haute que pourrait lui conférer celle d'initiatrice de l'échange. On remarque, et c'est frappant, que la réponse, c'est-à-dire la première intervention de BG commence par un *mais*, un *mais* qui peut être interprété comme passablement inamical, voire agressif (SIMON & AUCHLIN, ici-même). Ce connecteur pragmatique situé en position initiale a été analysé en détail, récemment, par SIMON & GROBET (2002), de même que le connecteur argumentatif *parce que* : tantôt il apparaît intégré prosodiquement au segment qu'il introduit, tantôt il est prononcé de façon autonome, d'où son statut ambigu. Nous n'en dirons pas plus, car BG1 est interrompue par RF2 – avant de reprendre par *alors que*. Notons simplement que ce dernier marqueur commence et termine le tour de parole BG2, encadrant deux connecteurs argumentatifs (*parce que* et *d'ailleurs*), pour reprendre la terminologie de ROULET *et al.* (1985). Ce sont ces oppositions introduites par *alors que* qui s'accompagnent d'une forte montée mélodique : « alors que là », « alors que ça devrait être objectif, la science ». Nous aurons l'occasion d'y revenir.

3.2. Petite étude de « texte »

3.2.1. Phénomènes typiques de la parole non lue

On peut voir un continuum entre parole spontanée et lecture, entre la conversation libre et le discours préparé. Le corpus étudié est un dialogue radio-diffusé, donc public, même s'il s'agit d'une entrevue de type face à face. Étant guidée, l'interaction est asymétrique ; la situation est plutôt formelle, et le registre de langue plus soutenu que relâché. Certes, sur le plan segmental, quelques schwas sont élidés, par exemple dans « un petit peu » (posant des problèmes à la recopie de prosodie) ; et deux *ne* de négation sont omis. Ces phénomènes sont connus, et nous les avons revisités récemment, dans de gros corpus oraux alignés automatiquement, d'où il

ressort bien qu'ils apparaissent davantage en oral spontané qu'en lecture (BOULA DE MAREÜIL & ADDA-DECKER 2002). Mais le nombre relativement faible d'hésitations témoigne d'un style plutôt soigné, même si les *euh* sont un peu plus nombreux, surtout chez BG, que ce qui a été transcrit orthographiquement : nous en avons relevé une dizaine, tous situées à des pauses à une ou deux exception(s) près, sur 4 minutes de parole. Placés sur un registre très bas, souvent en voix craquée, ces *euh* peuvent se distribuer n'importe où à une frontière de mot, mais sont stigmatisés comme peu prestigieux, voire choquants s'il y en a trop. L'allongement final est préféré pour remplir la pause, pour gagner du temps, surtout si la dernière syllabe est ouverte.

Proches acoustiquement de ces *euh* d'hésitation, les *e* d'appui, non étymologique, terminant un contour montant-descendant typique du parler des jeunes parisiens – et surtout des jeunes parisiennes (WALTER 1988 ; CANDEA 2002) – semblent absents de ce corpus. L'âge et l'origine géographique des locutrices peuvent l'expliquer, même si l'observation de ces *clichés* mélodiques dépasse la population « branchée » de notre capitale. Absence également de ces « 'petits mots' qui habitent l'oral », que MOREL & DANON-BOILEAU (1998) nomment « ligateurs » (plus heureux que « remplisseurs », terme péjoratif souvent utilisé) :

- ces *ben, quoi*, qui règlent la co-énonciation (un seul *eh bien* est à noter) ;
- ces *genre, style,* qui ont valeur de préposition restreignant le champ référentiel.

Quelques *donc*, *alors, enfin* scandent néanmoins le discours, par exemple dans une même phrase de RF1 (en 1), où la conjonction de coordination ne joue aucun rôle de démonstration ou de lien consécutif, mais assure une fonction de réactualisation.

(1) un livre qui a pour titre Cette mâle assurance donc m-a accent heu circonflexe l-e, Mâle assurance éditée donc chez Albin Michel

Il est un fait établi en syntaxe que *donc* a un comportement distinct des (autres) conjonctions de coordination, dans la mesure où, entre autres, il peut se combiner avec *et* : on peut dire *et donc* alors que **et mais* est incorrect. Ici, le premier *donc* introduit et le second *donc* clôt, pour ainsi dire, une incise qui mérite bien les appellations de « décrochement vers le bas » (MOREL & DANON-BOILEAU 1998) ou de « parenthèse basse » (DELATTRE 1966), en ce qu'elle est placée sur un registre plateau, peu modulée jusqu'à la syllabe finale, plus bas que le membre de phrase qui la précède. En synthèse de la parole à partir du texte, on pourrait appliquer une diminution de 20 % de la hauteur moyenne et du registre, ainsi qu'une diminution de l'énergie de 2 dB, comme BOULA DE MAREÜIL & MAILLEBUAU (2002) l'ont suggéré. L'intonation reprend ensuite au niveau où elle a été laissée, ce qui permet de souligner l'autonomie de l'élément enchâssé, même si le mot *mâle* est repris (sans

démonstratif). Cette répétition réinitialise le fil du discours, après la précision levant l'ambiguïté sur l'orthographe – ou le jeu de mot voulu.

3.2.2. Opérations d'extraction et de focalisation

Les hésitations et les répétitions (lointaines ou non, parfois difficiles à distinguer des faux départs, comme chez BG2 (en 2) ne sont habituellement pas manipulées en synthèse de la parole à partir du texte, qui est conçue pour une tâche de lecture.

(2) j'ai divisé en catégories ce ce cette *Mâle assurance*

Mais les opérations d'extraction à gauche (topicalisation) ou à droite (thématisation) et de focalisation existent aussi bien à l'écrit qu'à l'oral (LACHERET-DUJOUR & BEAUGENDRE 1999). On a chez BG2 (en 3) un bel exemple de cliticisation (sujet pronominal antéposé), où l'élément thématisé, ce dont on parle, est rejeté après l'apport d'information.

(3) ça devrait être objectif, la science

À la fin de l'extrait, on a la configuration symétrique (en 4), avec un pic mélodique sur la syllabe finale du mot *unique*.

(4) leur vocation unique c'était pas d'être mères et ménagères

L'emphase a reçu un certain nombre de dénominations, comme « focalisation » (TOUATI 1987) ou « accent d'insistance » (FOUCHÉ 1961). L'échange suivant (en 5) en est symptomatique, aussi bien dans la première occurrence de *on* que dans la dernière, où l'usage qui en est fait est métalinguistique.

(5) BG4 à vrai dire **on** m'a poussé : on est venu me proposer heu

RF5 c'est qui « on » ? [petit rire]

Dans la question partielle, l'insistance est renforcée par le procédé grammatical de mise en relief « c'est qui » – dispositif d'extraction souvent appelé *dislocation* ou *phrase clivée*. Dans la réponse (en 6), le *c'* peut être considéré comme un simple déictique.

(6) c'est l'homme qui s'occupe de cette collection qui a… qui publie souvent des dictionnaires

Mais dans la question totale, avec inversion, un peu plus loin (en 7), le focus (*ça*) porte un pic mélodique.

(7) est-ce un petit peu ça qui ressort de votre étude?

On note enfin un accent initial sur *péjoratif* ou sur *tellement pratique*, où l'adverbe a une forte charge sémantique, etc. Dans « tout le monde sait que la culture est misogyne », l'accent d'insistance souligné par nous, absent dans « je crois que »

(un peu plus loin), permet d'opposer un univers de connaissances supposé partagé par tous à l'univers de croyance de BG, exposé dans le préambule de BG7 (en 8).

(8) oui, parce que je crois que tout ce qui conforte le pouvoir est utilisé

Enfin, l'opposition marquée par les accents d'insistance entre les « hommes » et « l'invasion comme ils disent des femmes » se combine avec le fait de citer un mot avec une certaine distance, empreinte d'indignation.

3.2.3. Effets de liste

Ces phénomènes, de même que l'effet de liste, ne sont pas traités dans notre système. Dans l'exemple suivant le *hein…* (en 9), un hyperonyme tel que *intellectuels* ou *notables* peut pourtant aisément être trouvé. Le terme *grands esprits* le précède même, dans la bouche de RF1.

(9) Qu'il s'agisse de philosophes, d'écrivains, d'historiens, de scientifiques, […]

L'énumération se retrouve d'ailleurs à plusieurs reprises un peu plus loin chez BG2 (en 10), où l'on peut percevoir la reproduction d'un même contour intonatif, avec des accents sur l'article *les* ; et, sous forme de *ni… ni*, (en 11).

(10) les grands chefs religieux, les hommes de science, les hommes de loi, les hommes de lettres

(11) ni les poètes, ni surtout les savants !

L'énumération est explicite quand BG3 termine la liste de RF3 (en 12).

(12) RF3 donc le dénigrement de la femme n'est pas une question de niveau social, ni de religion, ni de caste

BG3 non – de rien ; ni même de civilisation ! [petit rire]

Ou bien (en 13), le début, qui est du discours rapporté (avant l'adjonction initiée par le connecteur réévaluatif *enfin*) est traînant, et trahit l'engagement, la condescendance de BG6. Ensuite, le quantificateur *tous/tout* répond positivement au *rien* de (24).

(13) de Molière, à Baudelaire, à Maupassant, à Montherlant, à Dutourd – enfin on peut les citer tous presque – tout le monde est misogyne.

D'une façon générale, le corpus est particulièrement riche en mises en parallèle de ce genre. Un indice formel tel que *etc.* dans la parenthèse (en 14), où l'on peut déceler une courbe mélodique en forme d'accent circonflexe, devrait permettre de les identifier automatiquement.

(14) dictionnaire de la bêtise, dictionnaire des erreurs politiques, etc.

Mais à vrai dire, on peut à la limite considérer les suites de citations « sont venus me dire », « j'ai dit », « on m'a dit : 'Mais pas du tout […]' » (avec un *mais*

très expressif) comme des énumérations, d'où la difficulté de trancher, et *a fortiori* de détecter automatiquement les énumérations.

3.3. Discussion

3.3.1. Fonction expressive et accents d'insistance

La fonction expressive est fondamentale dans la prosodie : quand on parle, on communique beaucoup d'informations sur soi-même, informations que des distorsions faites à une configuration spécifique de la courbe mélodique sont chargées de transmettre. Cette fonction expressive recouvre aussi l'accent emphatique. Dans notre système, seul l'accent interne au système de la langue, « normal », est pris en considération ; l'accent externe, dit « d'insistance » ou « de focalisation », qui est de plus en plus fréquent en français contemporain, n'est capturé par la synthèse de la parole qu'au hasard de la sélection. Avec TROUBETZKOY (1986), qui oppose le plan intellectuel de la langue au plan stylistique (auquel renvoient les fonctions psychologiques et émotionnelles), nous distinguons le niveau de la prosodie dite « grammaticale », qui est une donnée purement linguistique, de celui de la prosodie spontanée, qui renseigne essentiellement sur l'état d'esprit du locuteur, et est régi par l'intention de communication de celui-ci.

Cette prosodie spontanée, d'ordre extra-linguistique, dépend de la situation et de ce que l'on veut faire entendre. Elle met en jeu la sensibilité ; elle peut accompagner un désir de persuasion, comme chez les hommes politiques avant une élection (TOUATI 1995), traduire un sentiment de surprise, signifier la joie (BERGHE 1976), exprimer dans diverses langues une attitude de rejet, de mépris, de peur ou de colère envers quelqu'un ou quelque chose (MEJVALDOVÁ 2001 ; ROSENFELDER 2001).

Pour de telles raisons relevant de la stylistique, il arrive qu'on ajoute un élément accentuel, ayant une fonction identificatrice (involontaire de la part du sujet parlant) ou impressive (pour produire un effet sur l'auditeur). Nous en distinguons trois types : l'accent rhétorique (comme sur *péjoratif* ou *l'invasion* dans le corpus), l'accent contrastif (*sait* vs *crois*) et l'accent émotif (*tellement pratique*).

L'accent rhétorique permet de faire remarquer ou comprendre un mot particulier, en frappant la première syllabe de ce qui devient le *centre d'attention* du groupe. Il s'emploie pour traduire des effets littéraires, dans un souci de définir ou caractériser une notion, pour annoncer le commencement d'une idée importante et pour citer un mot – un degré de plus dans l'insistance peut amener une pause assez appréciable entre les deux mots. Cet accent « gagne lentement du terrain » (MARTINET 1969). Et il est fréquent de l'entendre dans une élocution pédante, chez

les orateurs qui ont l'habitude de s'adresser à un auditoire, chez les présentateurs et journalistes de radio ou de télévision, dans les salles de classe ou de conférence, dans les allocutions publiques et les exposés scientifiques, dans un but argumentatif, pédagogique ou didactique. Nous ne nous hasarderons pas à des suppositions sociolinguistiques sur le rythme saccadé de notre monde moderne, mais, comme l'a évoqué A. SÉGUINOT, la grande place prise par les actualités audio-visuelles peut favoriser un certain style (CARTON *et al.* 1977) ; si bien que cette prolifération a fait dire que le français est en période de transition, avec un *recul de l'accent*, qui serait plutôt sur la première syllabe des mots. P. PASSY avait, dès 1890, pressenti cette mutation avec perspicacité. Mais avancer la thèse d'un déplacement d'accent est sans doute exagéré : l'accent rhétorique se fait précisément remarquer parce qu'il n'est pas la règle. En outre, un autre *changement linguistique* à l'oeuvre dans le « parler des banlieues » aurait plutôt tendance à faire porter la proéminence accentuelle sur l'avant-dernière syllabe (HINTZE *et al.* 2000).

L'accent contrastif répond au souci de lever une ambiguïté ou d'opposer **explicitement** deux mots. La bipartition traditionnelle de MAROUZEAU en *accent intellectuel* et *accent affectif* ne rend pas compte de ce que cet accent peut frapper n'importe quelle syllabe : l'accent *intellectuel*, en effet – le terme paraît de plus curieux par rapport à la terminologie de TROUBETZKOY – est supposé tomber sur la première syllabe des mots, à la différence également de l'accent suivant.

L'accent émotif est attaché à des mots possédant une certaine charge sémantique – surtout des adjectifs, mais pas seulement. Cet accent emphatique peut aller jusqu'à modifier totalement le contenu notionnel du message (RIGAULT 1971). Il peut avoir pour effet de modifier la courbe intonative de base : les phénomènes mettant en jeu accentuation et intonation interagissent. Et le problème est compliqué par le fait que ces particularités se mélangent, et sont mesurées en bloc par les enregistreurs (BLANCHE-BENVENISTE *et al.* 1990).

3.3.2. Pour une approche de la prosodie à base de balises

Faut-il donc générer des variations phonostylistiques en synthèse de la parole à partir du texte ? Le faire systématiquement serait dangereux, par rapport à une prosodie « neutre », qui représente un cadre de référence sûr, un terrain mieux exploré (FAGYAL 1995). Pour éviter que la synthèse vocale ne lise un conte de fée à la manière d'un commentateur sportif, notre synthèse se contente de générer une prosodie dépourvue d'affection. La langue est étudiée comme système, laissant de côté la fonction impressive de la prosodie, bien que celle-ci soit, sans nul doute, cruciale dans la conversation de tous les jours.

En faire autrement serait de plus difficile voire impossible, les analyses sémantiques (*a fortiori* pragmatiques) n'étant à l'heure actuelle efficaces que dans des domaines restreints. Une décomposition automatique en thème-rhème, en

information « donnée » ou « nouvelle » est aujourd'hui hors de notre portée dans le tout-venant des textes, pour prévoir et synthétiser un focus « étroit » ou « large » (i.e. s'étendant au-delà d'un mot plein).

Mais des balises de contrôle permettent d'envisager une lecture plus « intelligente » du texte, tout en restant fondée sur les formes, en modifiant le type de prosodie « neutre » que produit automatiquement un système de synthèse de la parole à partir du texte – plutôt par diphones (SPROAT *et al.* 1998 ; MERTENS *et al.* 2003). Ces balises peuvent être insérées directement dans les textes pour communiquer des émotions (BOULA DE MAREÜIL *et al.* 2002) ou posées automatiquement sur la base de critères lexico-syntaxiques, typographiques et de ponctuation, pour développer un modèle d'emphase ou appliquer une prosodie de type parenthétique. En ce qui concerne les émotions, que nous avons étudiées dans le cadre d'un projet européen, de grandes tendances peuvent être dégagées, comme l'inversion des pentes de F_0 pour le dégoût, l'écrêtage des mouvements mélodiques pour la tristesse et la remontée en fin de phrase pour la surprise. Ce travail mériterait d'être complété par des investigations sur la qualité de voix, pour être appliqué à un système de synthèse par sélection.

4. PROSODIE ET QUALITÉ VOCALE

Les aspects « physiques » ou « acoustiques » de la prosodie sont plus difficiles à définir qu'il n'y paraît. La voix, en effet, est un instrument fort complexe : « instrument » et « instrumentiste » sont confondus. Depuis la théorie acoustique source–filtre de production de la parole, cet instrument est traditionnellement séparé en « phonation » et « articulation », une distinction fonctionnelle qui est en bonne entente mutuelle avec la division qui oppose « segmental » et « supra-segmental » dans le champ linguistique.

Cependant, ces différentes divisions de la langue et de la parole négligent un autre aspect de l'oralisation (pour reprendre un terme de LÉON 1993), qui est la qualité vocale. Négliger ce terme ne résiste guère à des analyses plus serrées ou à l'épreuve de la synthèse. Mais où se trouve la qualité vocale ? La « phonation » en particulier est responsable à la fois de la mélodie et du rythme, mais aussi d'une bonne partie du *timbre*. Le timbre est cependant plus volontiers associé à « l'articulation », qui elle-même influence la phonation, ne fût-ce qu'à travers la micro-prosodie. Ainsi, on accepte (sans généralement en tirer les conséquences) le fait que la « prosodie » au sens large du terme ne se limite pas à F_0 et aux durées, mais doit inclure aussi l'articulation et la qualité vocale. C'est ce que l'on va tenter d'expliciter sur quelques exemples, grâce au corpus étudié dans ce volume.

4.1. Les paramètres de la qualité vocale

4.1.1. Paramètres de la source glottique

Il est difficile de définir d'une façon générale ce qu'est la « qualité vocale », ce qui caractérise le timbre, la qualité d'une voix. Pour les besoins de l'analyse, on peut reprendre la décomposition de la production de parole en deux ensembles fonctionnels : la phonation et l'articulation. La phonation, ou action de la source vocale, porte une grande part de responsabilité dans la « qualité vocale ». Au niveau de la source, les paramètres physiques peuvent être décrits dans le domaine temporel par (FANT *et al.* 1985 ; CHILDER & LEE 1991 ; FLATT & KLATT 1990) :

- T_0 ou période fondamentale, inverse de la fréquence fondamentale (sans doute le paramètre le plus important du point de vue perceptif, pour cette raison le plus étudié et le mieux connu) ;
- *AV* ou amplitude de voisement, un paramètre du signal dont l'origine physique peut être diverse, selon que l'on considère que *AV* représente l'amplitude du débit glottique ou celle de sa dérivée ;
- *Oq* ou quotient ouvert. C'est le rapport de durée entre la phase ouverte de la glotte et le cycle glottique. Ce paramètre rend compte en particulier du caractère « serré » voire « étranglé » de la source ;
- *SQ* ou vitesse de fermeture. Ce paramètre rend compte de la dissymétrie de l'onde glottique, donc aussi du côté tendu de la voix ;
- *Ta* ou temps de fermeture, paramètre qui contrôle l'effort (une fermeture abrupte correspond à une voix forte, une fermeture lente à une voix douce) ;
- *PAP* ou rapport entre énergie périodique et apériodique. Ce paramètre représente le taux de bruit dans la source glottique.

Il est intéressant de considérer l'effet de ces paramètres de la source glottique dans le domaine spectral : ils sont alors plus faciles à relier aux aspects perceptifs que dans le domaine temporel. On peut considérer le jeu de paramètres spectraux suivant (DOVAL & D'ALESSANDRO 1997 ; HENRICH 2001) :

- F_0 ou fréquence fondamentale, inverse de T_0 ;
- *AG* l'amplitude spectrale globale, qui a le même sens que l'amplitude temporelle ;
- *ST* ou pente spectrale. C'est l'atténuation spectrale de la source en hautes fréquences – mesurée par exemple à 3 kHz. Ce paramètre est lié à la vitesse de fermeture et au temps de fermeture ;
- *FG* ou fréquence du « formant glottique ». L'amplitude du spectre de la source est en forme de pic, montant puis descendant. *FG* est le sommet de ce pic, dans la région de 0,5 à 3 F_0 environ. En général, *FG* est sous, ou bien

dans, la région du premier formant. Ce paramètre est lié au quotient ouvert et à la dissymétrie de la source (en plus de F_0, bien sûr) ;

- *FB* ou largeur de bande du « formant glottique ». Ce paramètre dépend de la dissymétrie et du quotient ouvert. Les paramètres du formant glottique vont régler le spectre en basse et moyenne fréquences, par exemple l'amplitude relative des premiers harmoniques de la source (HANSON 1997) ;
- *PAP* a le même sens de rapport d'énergie que dans le domaine temporel.

Ces paramètres de la source n'évoluent pas de façon absolument indépendante : par exemple, une voix « douce » ou « forte » impliquera des changements de *ST, TA, PAP, FG, FB, OQ, SQ* etc. Il est donc intéressant de citer des configurations typiques de paramètres.

4.1.2. Configurations de la source glottique

Les variations expressives de qualité vocale sont perçues par des combinaisons complexes de variation des paramètres acoustiques (BANSE & SCHERER 1996). En effet, il n'y a pas de lien direct et univoque entre tel ou tel paramètre, le système musculaire de la phonation et l'effet produit. Par exemple, le quotient ouvert et la dissymétrie peuvent jouer un rôle similaire, le *PAP* et la pente spectrale co-varient, etc. Par conséquent, pour décrire le signal produit, il faut plutôt considérer des types de qualités vocales qui correspondent à des configurations de la source. Les configurations typiques de la source glottique s'expriment mieux en considérant des paires, dont voici les plus communes sous forme résumée (GAUFFIN & SUNDBERG 1989).

- **Voix forte/voix faible**. La force de voix est une source importante de variation intra-individuelle. Les variations d'effort vocal vont impliquer des changements de plusieurs paramètres (SLUIJTER *et al.* 1997). Il faut distinguer la force de voix de la tension : un bon orateur parle fort sans tension de la voix, même si l'orateur peu entraîné aura tendance à « forcer » sa voix pour parler fort, c'est-à-dire à la tendre excessivement. La voix forte est associée à une pente spectrale *ST* faible (de l'ordre de –6dB/octave), un grand rapport *PAP*, une *AG* élevée. Au contraire, une voix faible aura une pente spectrale *ST* considérable (à la limite seul le premier harmonique sera vraiment présent), un rapport *PAP* faible, une *AG* faible.
- **Mécanismes laryngés**. Les mécanismes laryngés correspondent à des changements de l'activation musculaire du larynx, en fonction principalement de la fréquence fondamentale. Reprenant les travaux de ROUBEAU (1993), nous décrirons les registres de la source, c'est-à-dire les changements de timbre de la source en fonction de F_0 (MILLER 2000) en fonction des mécanismes laryngés.
  1. Mécanisme 0. Le registre ou mode de friture vocale correspond à une F_0 très grave (quelques dizaines de Hz), un *OQ* très petit, une forte asymétrie

SQ de l'onde glottique. Ce registre est commun en fin de phrase, hésitation, parenthèse, et pour les voix plus âgées. Si l'on monte la fréquence fondamentale à partir du registre de friture, il y a une brusque rupture vocale, et l'on abouti au mécanisme I.

2. Mécanisme I. Le mécanisme I correspond au « registre de poitrine », ou à la « voix modale » : les cordes vibrent sur toute leur longueur ; elles sont relativement épaisses. F_0 est grave ou moyenne, le quotient ouvert *OQ* variable, de même que les autres paramètres glottiques. Si F_0 continue à monter, il y a une seconde brusque rupture vocale, qui abouti au mécanisme II.
3. Mécanisme II. Le mécanisme II correspond au « registre de tête » ou à la « voix de fausset ». Les cordes ne vibrent que sur une partie de leur longueur. F_0 est élevée ; le quotient ouvert *OQ* est souvent plus grand, avec parfois une fermeture incomplète des cordes vocales. Les voix de femmes font souvent usage des mécanismes I et II, comme nous le verrons dans l'analyse du corpus.
4. Mécanisme III. Ce mécanisme correspond au registre de sifflet, registre suraigu proche du cri, qui n'apparaît pas ici.

- **Voix serrée/voix détendue**. Le serrage vocal est marqué principalement par un changement de quotient ouvert *OQ*. Chez le locuteur entraîné, cela se fait indépendamment de l'effort vocal. L'effet spectral est un mouvement du formant glottique *FG* vers l'aigu, et donc une diminution de l'amplitude du premier harmonique par rapport aux harmoniques suivants.
- **Voix rauque/voix lisse**. La raucité de la voix est liée aux apériodicités de la source, donc à un rapport *PAP* faible. La voix rauque montre une F_0 très irrégulière (gigue), accompagnée ou non de bruit d'aspiration.
- **Voisement/dévoisement**. Le voisement ou le dévoisement de la source est lié au tonus vocal – ainsi, la voix faible aura tendance à se dévoiser. Ce dévoisement est souvent utilisé, en particulier dans le contexte radiophonique qui nous intéresse ici, pour marquer la proximité avec l'interlocuteur. Cependant, une voix faible n'est pas forcément dévoisée, et le locuteur entraîné jouera indépendamment de l'effort vocal et du voisement. Le paramètre principal est ici le rapport *PAP*, mais aussi le quotient ouvert *OQ*, qui est proche ou égal à un lorsque l'on dévoise.
- **Voix timbrée/voix détimbrée**. Indépendamment de l'effort vocal, et du serrage glottique, une certaine ampleur peut être donnée en « plaçant » la voix. Si l'on pense au formant du chanteur, présent souvent aussi chez l'orateur, l'effet serait à attribuer au conduit vocal plutôt qu'à la source vocale.

4.1.3. Les paramètres de la qualité vocale : le conduit vocal

Le conduit vocal joue aussi un rôle important dans la qualité vocale : c'est lui qui définit en grande partie le « timbre » vocalique. Bien sûr, le conduit vocal doit avant tout garantir l'intelligibilité de la parole, réaliser les bons phonèmes. Il reste néanmoins un espace de variabilité suffisant pour que le conduit vocal contribue aussi à la fonction expressive, au-delà du strict contenu phonémique. Les principaux effets du conduit vocal pour la variation expressive sont les suivants (LAVER 1993).

- **Étirement des lèvres**, sourire. L'effet acoustique du sourire semble se situer aux alentours du 3^{e} formant vocalique (cette région spectrale est amplifiée).
- **Allongement** du conduit vocal. Les variations de longueur du conduit vocal, par descente (allongement) ou montée (rétrécissement) du larynx ont un effet de grossissement de la voix. Un conduit vocal long donne une « grosse » voix, la voix d'un individu plus imposant. Un conduit vocal court donne au contraire une « petite » voix, la voix d'un individu plus petit (OHALA 1983).
- **Antériorisation/postériorisation**. Ici, c'est l'articulation supraglottique qui va changer la longueur apparente du conduit vocal. Une voix plus antérieure sera jugée plus « petite », et un voix plus postérieure plus « grosse ».
- **Effort** d'articulation. Particulièrement marqué sur les fricatives par exemple (plus ou moins sifflantes), la tension, l'effort d'articulation peut marquer la tension du locuteur.
- **Vitesse** d'articulation. Cette vitesse ou netteté d'articulation peut également marquer la tension ou la détente du discours et/ou du locuteur.

Cette liste ne prétend pas à l'exhaustivité, mais donne quelques exemples du rôle important du conduit dans le jeu de la qualité vocale, en plus des effets de la source.

4.2. Petite analyse phonostylistique

Dans ce paragraphe, le corpus sonore Fayard-Groult est analysé en mettant en lumière des exemples d'utilisation expressive de la qualité vocale, ainsi que des exemples de phénomènes à la frontière entre prosodie et qualité vocale. Les notations des extraits sonores en secondes correspondent aux temps relevés dans les fichiers sonores du corpus. Il est particulièrement indiqué dans ce contexte d'écouter le fichier sonore fayarda, qui contient une recopie de la prosodie en regard du fichier sonore FayardGroult22k, qui est le signal original (fichiers sonores disponibles sur le CD-ROM).

Les locutrices dont nous analysons les enregistrements peuvent être considérées comme des locutrices « professionnelles » : elles ont une grande habitude de la prise de parole en public. À la fois leur discours, leur oralisation et

leur expression vocale sont remarquablement structurés, et en général bien contrôlés. Ce corpus est riche en effet expressifs, et se prête bien à une analyse des détails phonostylistiques.

4.2.1. Qualité de voix globale

À l'audition de la recopie prosodique, on est immédiatement frappé par la différence globale de qualité vocale de la recopie prosodique obtenue par resynthèse. Bien sûr, il ne s'agissait pas de reproduire les voix des locutrices ; cependant, l'écoute des voix de synthèse par sélection est également instructive à cet égard. Dans le cas de la synthèse par sélection, la qualité de voix varie (certes de façon non véritablement contrôlée), alors que dans la recopie prosodique la qualité de voix semble comme figée, fixée.

Cette qualité de voix invariable est due à la façon dont sont enregistrés les diphones. Les défauts de la recopie sont dès lors de deux ordres :

- D'une part, la synthèse par diphones donne une qualité vocale que l'on perçoit comme serrée. Cela s'explique par le traitement du signal appliqué aux diphones : le fait de manipuler des périodes du signal a tendance à réduire le quotient ouvert. De plus, on peut faire l'hypothèse que les voix qui par diphones se synthétisent bien doivent avoir des périodes bien marquées, donc un quotient ouvert intrinsèquement faible, ayant pour effet une voix qui semble serrée.
- D'autre part, ce sont les mêmes segments acoustiques, les mêmes diphones, qui sont utilisés pour réaliser toutes les hauteurs mélodiques. L'effet de l'instrument vocal qui possède de façon intrinsèque des registres (laryngés et résonanciels), c'est-à-dire des changements de timbre avec la hauteur mélodique, est donc perdu.

Au contraire en synthèse par sélection, l'effet des registres est plus ou moins préservé, puisque les segments de parole naturelle utilisés contiennent ces effets, au moins localement. De même, la voix de synthèse n'est pas spécialement plus serrée que la voix naturelle correspondante. Ainsi, les effets de covariations des paramètres glottiques sont localement corrects. En revanche, il n'y a pas la cohérence expressive de la qualité de voix, qui varie au hasard de la sélection.

Si l'on compare maintenant la qualité de voix globale des deux locutrices, on constate que RF garde un timbre égal, mesuré, avec peu d'effets non contrôlés. C'est une « belle » voix, bien placée, régulière, détendue, sans effort, égale et contrôlée. La voix de BG est moins contrôlée, plus relâchée, avec des effets parfois peu radiophoniques, comme des passages biphoniques (deux hauteurs fondamentales simultanées), beaucoup de friture vocale, une voix parfois très serrée.

4.2.2. Changements de mécanismes laryngés

Les deux locutrices utilisent en majorité le mécanisme I. En conséquence, ce sont les changements vers les mécanismes 0 et II qui peuvent apporter des effets contrastifs.

Les changements du mécanisme I (modal) vers le mécanisme II (fausset) sont présents et assez fréquents pour les deux locutrices. Ces changements accompagnent une montée mélodique : c'est souvent une mimique vocale (au sens de FÓNAGY 1983) qui exprime le doute ou l'indignation, donc un incrément de tension dans le propos et son expression. Quelques exemples en sont (15-17) :

(15) RF de grands esprits [0.391-0.397]

(16) BG gravité [0.512-0.516]

(17) BG avons fait [2.034-2.041]

Remarquons que ces deux derniers exemples, avec des variations brusques de F_0, ne sont pas rendus par la recopie prosodique.

Les changements du mécanisme I (modal) vers le mécanisme 0 (friture) sont aussi fréquents dans l'enregistrement. Ils sont plus présents chez BG que chez RF. Ce type de transition est une marque de phonation plutôt relâchée (de fait, il s'agit d'une détente vocale), qui est en général évitée dans une élocution soignée. RF, plus professionnelle en la matière, en use moins que BG. Elle est dans une situation « d'hôtesse radiophonique » et ne s'autorise pas un style plus familier. Au contraire pour BG, l'utilisation de la friture vocale peut sembler une marque d'aise, d'assurance, de franchise et de discours « naturel », attitudes que lui permet sa situation d'invitée. En témoignent les exemples 18-20 :

(18) BG catégories professionnelles [1.07-1.08]

(19) RF leur cause [0.14-0.15]

(20) RF circonflexe elle euh [0.296-0.231]

Un dernier exemple remarquable de BG (en 21) montre un passage direct du mécanisme II au mécanisme 0, de la voix de fausset à la friture vocale. De fait, le discours est à cet endroit tendu ; et la voix, plutôt étranglée, se détend subitement en friture vocale.

(21) BG parce que [2.1526 2.159]

La friture vocale échappe presque systématiquement à la recopie prosodique, ainsi d'ailleurs qu'aux analyseurs de F_0. Dès lors, les passages utilisant ce mécanisme laryngé sont systématiquement mal rendus à la synthèse, et on risquerait de commettre des erreurs d'interprétation si l'on se fiait trop à la resynthèse pour les étudier.

4.2.3. Jeu expressif de la glotte

Un autre effet très commun de qualité vocale, qui influe sur la prosodie (au moins sur la mélodie), est le dévoisement : BG dévoise plus que RF, encore une fois à cause d'un contrôle moins précis de son élocution. Le dévoisement est souvent associé aux fins de phrases ou d'incises, à la baisse de tonus de la voix, comme dans l'exemple (22) : dans ce cas, le dévoisement est accompagné d'un mouvement mélodique vers le registre grave.

(22) BG est utilisé [2.192-2.201]

Mais le dévoisement peut aussi intervenir, comme dans l'exemple (23), en cas de mimique d'étranglement : la voix se serre jusqu'au moment où les cordes vocales ne vibrent plus. Ici, c'est une marque de l'indignation de la locutrice, dont le dévoisement s'accompagne d'un mouvement mélodique vers le registre aigu :

(23) BG faites pour ça 2.431-2.437]

BG présente parfois une voix légèrement biphonique, avec une vibration bimodale des cordes vocales. Cela arrive en particulier en 24, où la voix se serre – mimique vocale qui marque une réprobation par rapport à ce qui est dit.

(24) BG leur vocation unique [2.514-2.528]

L'instrument vocal n'a pas un timbre homogène sur tout son ambitus. Ainsi, des changements de timbre accompagnent forcément les changements de hauteur mélodique : changements de quotient ouvert *Oq*, de *PAP*, de *ST*, etc. Dans l'exemple (25), l'indignation dubitative que RF veut marquer est rendue par une voix subitement adoucie, comme si la locutrice se parlait à elle-même pour se convaincre de l'incroyable. Il s'agit de changements de *Oq*, F_0, *PAP* et *ST*, etc.

(25) RF à peine croyable [0.439-0.451

Un effet inverse, de renforcement du tonus vocal est donné par BG : la répétition de l'exemple (26) est l'occasion d'un nouveau départ de la qualité vocale. Après une prise de souffle, la répétition du « j'ai » a retrouvé un timbre tonique, alors que le premier « j'ai » perdait de sa force. L'hésitation du discours donne lieu à une reprise en main de la voix, si l'on peut dire :

(26) BG d'ailleurs j'ai, j'ai [0.589-1.009]

Des différences locales d'effort vocal sont utilisées pour renforcer un mot, le mettre en valeur, l'accentuer. L'exemple (27) de BG met en valeur le « on », en introduisant une occlusion glottique, et en serrant la voix (changements de quotient ouvert *Oq* et de pente spectrale *ST*), ce qui donne une voix plus tonique, un accent d'insistance (voir aussi § 3.2.2 et 3.3.1).

(27) BG a vrai dire on m'a poussée [1.265-1.277]

L'utilisation de l'aigu du mécanisme I donne une voix peu assurée, un peu faible, avec un peu de souffle et une augmentation de pente spectrale. Cette voix affaiblie est inconfortable pour la locutrice, qui du coup l'utilise pour introduire une nuance de doute, de questionnement, de diminution. La diminution du volume vocal est analogue à celle de l'assurance du discours. Un exemple en est donné par RF dans l'extrait suivant (voir également l'exemple 7) :

(28) RF un petit peu ça qui [2.119-2.125]

La voix rauque exprime une gêne ; c'est encore une mimique d'étranglement, de serrage glottique. Il y a comme une rétention de l'air expiré, un frottement glottique. BG l'utilise (en 29) pour appuyer dans son discours sa réprobation, son rejet, presque son dégoût pour la « réaction » :

(29) BG espèce de réaction [2.566-2.572]

Enfin, un dernier effet expressif de la source consiste à rendre audibles les prises de souffle : il s'agit ici d'une mimique d'étouffement, de manque d'air. La prise de souffle est nécessaire car le discours a été très soutenu, dense, épuisant, ou bien parce qu'on s'apprête au contraire à lancer une longue tirade, comme RF dans l'exemple (30).

(30) RF leur cause (aspiration) [0.147-0.152]

BG reprend son souffle en même temps qu'elle hésite (en 31). À la fois la pensée, le discours et l'appareil vocal profitent de cette pause pour se renouveler.

(31) BG les hommes de loi (aspiration) euh les hommes [0.559-0.564]

4.2.4. Jeu expressif du conduit vocal

Un des effets expressifs du conduit vocal les plus communs (et un des mieux perçus) est celui de l'étirement des lèvres, qui accompagne le sourire. Dans cette émission sérieuse, sur un sujet qui ne porte guère au rire, le sourire est rare : RF en donne néanmoins un bon exemple (en 32). Ce sourire (qui va d'ailleurs vers le rire) est lourd de sens : il marque la gêne de RF devant l'indiscrétion qu'elle montre. Il est une mimique vocale dont le sens serait « vous en avez dit trop ou pas assez, révélez à nos auditeurs la mystérieuse personne que vous évoquez ». Pour cette demande indiscrète, RF doit faire montre d'humour, user du sourire, afin d'obtenir une information de son interlocutrice sans abuser de son autorité. Ce sourire s'accompagne d'une détente glottique :

(32) RF c'est qui on ? (rires) [1.295-1.308]

Un autre sourire, de BG, a semble-t-il un sens tout différent (en 33) : il marque une sorte de dépit devant l'universalité de la misogynie. Ici, le sourire est comme une

marque d'humour désespéré, que l'on pourrait paraphraser ainsi : « puisqu'il n'y a plus rien à faire, sourions-en ! ».

(33) BG ni même de civilisation [1.200-1.210]

Le sourire est typique de l'utilisation de la qualité vocale pour apporter des nuances au discours, nuances que l'on pense d'ordinaire réalisées par la prosodie, mais qui sont ici tout à fait indépendantes de F_0 ou de la durée : c'est véritablement un autre moyen de nuancer le discours par la mimique expressive du conduit vocal. Au niveau labial, il serait également possible d'user de l'arrondissement, afin de marquer le doute – même si on n'en rencontre pas d'exemple dans le court extrait analysé ici.

Le « grossissement » de la voix peut quant à lui s'obtenir par l'allongements du conduit vocal : celui-ci peut s'allonger vers l'arrière, en abaissant le larynx, ou vers l'avant, en avançant les lèvres. Une autre façon de simuler un allongement du conduit vocal est de postérioriser l'articulation : on « grossit » ainsi sa voix à peu de frais. RF prend (en 34 et en 35) une voix assurée, détendue, bien timbrée, dans le registre grave du mécanisme I, grâce à un léger allongement postérieur du conduit vocal, qui fait bien sonner sa voix et lui donne de l'ampleur. C'est encore une marque de professionnalisme vocal.

(34) RF patriarcale [2.106-2.115]

(35) RF castes [1.174-1.186]

Le raccourcissement du conduit vocal au contraire va donner une « petite » voix. Dans l'exemple (36), BG explore l'aigu de son registre de poitrine, avec un conduit vocal raccourci, et une voix de ce fait un peu tendue, qui demande un effort vocal. Cela aide à marquer l'accent, à mettre en relief « les savants ».

(36) BG ni surtout les savants [1.111-1.120]

4.3. Discussion : qualité vocale, instrument expressif et stylistique de l'expression

Après une étude de quelques phénomènes rencontrés dans ce court extrait de dialogue radiophonique, il est clair que la recopie de la prosodie au sens physique étroit du terme ne peut aboutir à une synthèse qui semble naturelle : il faudrait également prendre en compte toutes les variations de qualité de la voix. Cependant, cet attribut de « naturel » n'est pas simple à définir : bien des variations de la qualité vocale sont un jeu subtil avec l'instrument vocal pour faire passer des nuances de signification. Ces nuances pourraient bien aller au-delà de l'aspect linguistique proprement dit, et s'inscrire davantage dans une dimension expressive, para- ou extra-linguistique. Ainsi, il ne serait pas question de l'inadéquation des modèles de

la prosodie linguistique, mais plutôt de créer des modèles phonostylistique opérationnels pour véhiculer de l'expressivité.

Cette étude stylistique de l'expression ouvre de vastes perspectives. D'un côté, ce domaine peut être abordé du point de vue de l'instrument vocal : en effet, l'analyse de la prosodie en termes de paramètres physiques comme F_0 ou la durée est probablement une façon inadéquate de poser le problème. Le locuteur (contrairement au chanteur par exemple) ne se préoccupe pas avec précision de la mélodie, n'en a pas une conscience directe : ce sont probablement des couples de paramètres comme tension/détente ou douceur/force qui pilotent l'instrument. Alors, un accroissement de tension par exemple va se traduire certes par une augmentation de F_0, mais également par un changement de quotient ouvert, voire par un changement de vitesse des mouvements articulatoires. Une piste pour gérer ensemble tous ces paramètres qui co-varient serait l'utilisation de modèles mécaniques, physiques, de la production vocale, modèles directement pilotés par des paramètres physiques comme la tension de la glotte par exemple. L'auditeur est parfaitement rompu à l'interprétation des moindres variations de qualité vocale. Il détecte donc les incohérences de la source sonore : une fréquence fondamentale qui augmente sans co-variation des paramètres glottiques viole les règles d'un appareil vocal « naturel ». Ce type de d'incohérence sur le plan physique n'échappe donc pas à l'auditeur : il s'agit ici des effets pour ainsi dire « instrumentaux » de l'expression.

Un autre point de vue sur la qualité vocale utilisée par la stylistique de l'expression est le point de vue sémiologique : ici interviennent les phénomènes expressifs significatifs, dans le cadre conventionnel de l'échange, mais qui ne ressortissent pas à proprement parler d'une analyse linguistique. Par exemple, le sourire qui peut avoir fonction d'approbation du discours, ou bien le resserrement glottique (la mimique d'étranglement) qui marque l'indignation. Il est difficile dans un dialogue radiophonique de faire la part de l'émotion véritable et de l'émotion simulée des interlocutrices, par ailleurs intellectuelles rompues à l'exercice de la « mise en scène » vocale. Toujours est-il que les variations stylistiques de qualité vocale font usage d'un code, de convenances qui font sens pour les auditeurs. Une étude « sémiologique » des effets de qualité vocale serait donc également à entreprendre, en plus de l'étude « instrumentale ». En l'absence de tels travaux, la prosodie de synthèse ne peut que rester pauvre en regard de la parole naturelle.

5. CONCLUSION

Le « texte » du corpus proposé a été resynthétisé par le système SeLimsi que nous avons décrit. Ce système possède de nombreuses voix synthétiques et nous avons, pour les exemples sonores associés, choisi un couple de voix féminines ainsi qu'une voix masculine. Le lecteur pourra apprécier le résultat de la synthèse : il est évident

que celle-ci manque la plupart des effets expressifs que les locutrices avaient réalisés. De ce fait, la synthèse semble artificielle ; elle joue faux, comme un acteur joue faux ou peut sembler artificiel s'il n'a pas le « ton juste ». Ceci est particulièrement vrai lorsque le texte et la parole coopèrent pour un effet particulier, par exemple les hésitations.

En confrontant la synthèse et un dialogue oral dans toute sa richesse, on mesure l'effort de recherche nécessaire pour que la machine passe avec succès le test de Turing, c'est-à-dire arrive à faire croire que c'est un humain qui parle. Cet objectif reste manifestement utopique pour les quelques années, voire les quelques décennies à venir.

Christophe D'ALESSANDRO, Philippe BOULA DE MAREÜIL
& Romain PRUDON
LIMSI-CNRS
BP 133 91403 Orsay CEDEX
France
{cda,mareuil,prudon}@limsi.fr

RÉFÉRENCES BIBLIOGRAPHIQUES

D'ALESSANDRO, C. & P. MERTENS. 1995. « Automatic pitch contour stylization using a model of tonal perception », *Computer Speech and Language* 9, 257-288.

BALESTRI M., A. PACCHIOTTI, S. QUAZZA, P. L. SALZA & S. SANDRI. 1999. « Choose the best to modify the least: a new generation concatenative synthesis system », *Eurospeech*, Budapest, 2291-2294.

BANSE, R. & K. SCHERER. 1996. « Acoustic profiles in vocal emotion expression », *Journal of Personality and Social Psychology* 70/3, 614-636.

BERGHE, C. & L. VAN DEN. 1976. *La phonostylistique du français*, The Hague-Paris : Mouton & Co.

BLANCHE-BENVENISTE, C., M. BILGER, C. ROUGET & K. VAN DEN EYNDE. 1990. *Le français parlé, études grammaticales*, Paris : Éditions du CNRS.

BOULA DE MAREÜIL, P., C. D'ALESSANDRO, F. BEAUGENDRE & A. LACHERET-DUJOUR. 2001a. « Une grammaire en tronçons appliquée à la génération de la prosodie », *Traitement Automatique des Langues* 42, 115-143.

BOULA DE MAREÜIL, P., P. CÉLÉRIER, T. CESSES, S. FABRE, C. JOBIN, P.-Y. LE MEUR, D. OBADIA, B. SOULAGE & J. TOEN. 2001b. « Elan Text-To-Speech :

un système multilingue de synthèse de la parole à partir du texte », *Traitement Automatique des Langues* 42, 223-252.

BOULA DE MAREÜIL, P. & M. ADDA-DECKER. 2002. « Studying pronunciation variants in French by using alignment techniques », *ICSLP*, Denver, 2274-2277.

BOULA DE MAREÜIL, P., P. CÉLÉRIER & J. TOEN. 2002. « Generation of Emotions by a Morphing Technique in English, French and Spanish », in B. BEL & I. MARLIEN (eds.), *Proceedings of the Speech Prosody 2002 conference*, 11-13 avril 2002, Aix-en-Provence : Laboratoire Parole et Langage, 187-190.

BOULA DE MAREÜIL P. & E. MAILLEBUAU 2002. « Traitement des incises en français : capture automatique et modèle prosodique », *JEP*, Nancy.

BREEN, A. 2000. « Issues in the development of the next generation of concatenative speech synthesis systems », *IEE Seminar on the State of the art in speech synthesis*, London, 1-4.

CAMPBELL, N. 1998. « Multi-lingual concatenative speech synthesis », *ICSLP*, Sydney.

CANDEA, M. 2002. « Le *e* d'appui parisien : statut actuel et progression », *JEP*, Nancy, 185-188.

CARTON, F., A. MARCHAL, D. HIRST & A. SÉGUINOT. 1977. *L'accent d'insistance (Emphatic stress)*, Montréal : Didier.

CHILDERS, D.G. & C.K. LEE. 1991. « Vocal quality factors: Analysis, synthesis and perception ». *J. Acoust. Soc. Am.* 90/5, 2394-2410.

COORMAN, G., J. FACKRELL, P. RUTTEN & B. VAN COILE. 2000. « Segment selection in the L&H RealSpeak Laboratory TTS system », *ICSLP*, Beijing.

DELATTRE, P. 1966. « Les dix intonations de base du français », *French Review* 40, 1-14.

DOVAL, B. & C. D'ALESSANDRO. 1997. « Spectral correlates of glottal waveform models: an analytic study », *ICASSP*, 446-452.

DUTOIT, T., V. PAGEL, N. PIERRET, F. BATAILLE & O. VAN DER VRECKEN. 1996. « The MBROLA project: towards a set of high quality speech synthesizers free of use for non commercial purposes », *ICSLP*, Philadelphia, 1393-1396.

FANT, G., J. LILJENCRANTS & Q. LIN. 1985. « A four-parameter model of glottal flow », *STL-QPSR* 85/2, 1-13.

FAGYAL, Z. 1995. *Aspects phonostylistiques de la parole médiatisée lue et spontanée: âge, prestige, situation, style et rythme de parole de l'écrivain M. Duras*, Thèse de doctorat de l'université Paris III. Sorbonne Nouvelle.

FÓNAGY, I. 1983. *La vive voix. Essais de psycho-phonétique*, Paris : Payot.

FOUCHÉ, P. 1961. *Phonétique historique du français*, Paris : Klincksieck.

GAUFFIN, J. & J. SUNDBERG. 1989. « Spectral correlates of glottal voice source waveform characteristics », *Journal of Speech and Hearing Research* 32, 556-565.

HANSON, H. M. 1997. « Glottal characteristics of female speakers: Acoustic correlates », *J. Acoust. Soc. Am.* 101, 466-481.

HENRICH, N. 2001. *Étude de la source glottique en voix parlée et chantée*, Thèse de doctorat de l'université Paris VI. Paris.

HINTZE, M.-A., T. POOLEY & A. JUDGE (eds). 2000. *French accents : phonological and sociolinguistic perspectives*, London : AFLS/CiLT.

KLATT, D. & L. KLATT. 1990. « Analysis, synthesis, and perception of voice quality variations among female and male talkers », *J. Acoust. Soc. Am.* 87, 820-857.

LACHERET-DUJOUR, A. & F. BEAUGENDRE. 1999. *La prosodie du français*, Paris : Édition du CNRS.

LAMEL, L. F., J.-L. GAUVAIN & M. ESKÉNAZI. 1991. « BREF, a Large Vocabulary Spoken Corpus for French », *Eurospeech*, Genova, 505-508.

LAVER, J. 1993. *Principles of phonetics*, Cambridge : Cambridge University Press.

LÉON, P. 1994. *Précis de phonostylistique. Parole et expressivité*, Paris : Nathan.

LUZZATI, D. 1995. *Le dialogue verbal homme-machine. Étude de cas*, Paris: Masson.

MARTINET, A. 1969. *Le français sans fard*, Paris : PUF.

MEJVALDOVÁ, J. 2001. *Expressions prosodiques de certaines attitudes en tchèque et en français : étude comparative*, Thèse de doctorat de l'université Paris VII et de l'université Charles de Prague.

MERTENS, P. 1987. *L'intonation du français, De la description linguistique à la reconnaissance automatique*, Doctorale dissertatie. K. Universiteit Leuven.

MERTENS, P., A. AUCHLIN, J.-P. GOLDMAN & A. GROBET. 2003. « L'intonation du discours : une implémentation par balises ; motifs et premiers résultats », in V. AUBERGÉ, A. LACHERET & H. LOEVENBRUCK (éds), *Journées Prosodie 2001,* Actes du colloque, 10-11 octobre 2001, Grenoble : Université de Grenoble, 93-98.

MILLER, D.G. 2000. *Registers in singing*, Thèse de doctorat de l'université Royale de Groningen, Groningen.

MINKER, W. & S. BENNACEF. 2001. *Parole et dialogue homme-machine*, Paris : Eyrolles/CNRS Éditions.

MOREL, M.-A. & L. DANON-BOILEAU. 1998. *Grammaire de l'intonation. L'exemple du français*, Paris : Ophrys.

OHALA, J. 1983. « Cross-language use of pitch: an ethological view », *Phonetica* 40, 1-18.

PRUDON, R. & C. D'ALESSANDRO. 2001. « A selection/concatenation TTS synthesis system : databases development, system design, comparative evaluation », *4th ISCA Workshop on Speech Synthesis*, Pitlochry, 137-142.

PRUDON, R., C. D'ALESSANDRO & P. BOULA DE MAREÜIL. 2002. « Prosody Synthesis by Unit Selection and Transplantation on Diphones », *IEEE Workshop on Speech Synthesis*, Santa Monica.

ROSENFELDER, M. 2001. *Une grammaire du syldave.* Klow : Verkhwen Klowaswa.

RIGAULT, A. (éd.). 1971. *La grammaire du français parlé*, Paris : Hachette.

ROUBEAU, B. 1993. *Mécanismes vibratoires laryngés et contrôle neuro-musculaire de la fréquence fondamentale*, Thèse de doctorat de l'université Paris XI. Orsay.

ROULET, E., A. AUCHLIN, J. MOESCHLER, C. RUBATTEL & M. SCHELLING. 1985. *L'articulation du discours en français contemporain*, Berne : Peter Lang, rééd. en 1991.

SABAH, G., J. VIVIER, A. VILNAT, J.-M. PIERREL, L. ROMARY & A. NICOLLE. 1997. *Machine, langage et dialogue*, Paris : L'Harmattan.

SIMON, A.C. & A. GROBET. 2002. « Intégration ou autonomisation prosodique des connecteurs », in B. BEL & I. MARLIEN (eds.), *Proceedings of the Speech Prosody 2002 Conference*, 11-13 avril 2002, Aix-en-Provence : Laboratoire Parole et Langage, 647-650.

SLUIJTER, A. VAN HEUVEN, V. J. & , J. J. A. PACILLY. 1997. « Spectral balance as a cue in the perception of linguistic stress », *J. Acoust. Soc. Am.* 101, 503-513.

SPROAT, R., A. HUNT, M. OSTENDORF, P. TAYLOR, A. BLACK, K. LENZO & M. EDGINGTON. 1998. « SABLE: a standard for TTS markup », *Third ESCA COCOSDA International Workshop on Speech Synthesis*, Jenolan Caves, 27-30.

TOUATI, P. 1987. *Structures prosodiques du suédois et du français. Profils temporels et configurations tonales*, Lund : Lund University Press.

TOUATI P. 1995. « Pitch range and register in French political speech », *ICPhS*, Stockholm, 244-248.