CILL 28 (2003) n°, ppp-ppp

LA VÉRIFICATION D'HYPOTHÈSES LINGUISTIQUES AU MOYEN DE LA SYNTHÈSE DE LA PAROLE

Eric KELLER
Université de Lausanne

1. INTRODUCTION

Ces dernières années, un nombre important de linguistes ont découvert la grande utilité de la synthèse de la parole pour la vérification d'hypothèses portant sur le langage et la parole. Par une simulation globale ou partielle de traitements linguistiques, il est devenu possible de vérifier rapidement un certain niveau d'adéquation entre hypothèses et réalisations. Par ce biais, on peut relever des erreurs ou des limites d'hypothèses, clarifier des interactions entre différents traitements, ou encore examiner le degré de pertinence de certains traitements dans différents contextes linguistiques et communicatifs (Keller 2001; Keller & Zellner 1998b).

De plus, la performance produite par une machine a le mérite de reposer la question de l'adéquation d'un modèle linguistique. Quand un modèle informatique de la performance linguistique humaine peut-il être jugé adéquat? Serait-ce quand la machine produit *exactement* ce qu'un locuteur humain aurait produit à cet instant communicatif? Et si oui, quel type de locuteur humain la machine devrait-elle alors simuler (jeune, vieux, locuteur hautement compétent ou naïf, etc.)? Ou encore, serait-ce quand la machine produit le *même ordre* de chose? Si oui, comment définir « même ordre » ? Le concept traditionnel et catégoriel de la *distinctivité* suffit-il pour effectuer de simples jugements d'acceptabilité dans des domaines aussi nuancés comme la prosodie, l'expression émotive ou un discours interactif? Ou aurions-nous dorénavant besoin d'une méthode d'évaluation plus souple?

Dans les quelques lignes qui suivent, nous évoquerons différents éléments de cette thématique. Après une discussion de quelques problèmes initiaux entourant l'utilisation de la synthèse, nous proposons quelques pistes de réflexion autour de problèmes posés par l'utilisation concrète de cette approche.

2. DISPONIBILITÉ, QUALITÉ, ACCESSIBILITITÉ

Admettons initialement que le processus de vérification par la simulation ne se déroule pas sans difficulté. L'adoption générale de cette technique se heurte principalement à deux formes de difficulté.

Premièrement, les moyens disponibles sont rarement à la hauteur de la simulation désirée. Même si les progrès des dix dernières années ont permis d'améliorer la qualité de systèmes effectuant une simple lecture à haute voix de textes, nul ne serait induit à croire que la performance linguistique de tels systèmes est comparable à celle d'êtres humains engagés dans des actes de communication orale. Le manque de flexibilité ou d'interactivité, ainsi que la trop grande régularité trahissent presque inévitablement la machine aux oreilles d'un interlocuteur humain. De plus, il y a absence d'outils pour simuler certaines langues ou certains aspects de l'activité communicative (p.ex. interactivité, nuances d'expressivité ou émotivité.)

Le deuxième problème concerne l'inaccessibilité et la difficulté d'utilisation de tels systèmes. Les systèmes commerciaux sont des « boîtes noires » résolument fermées aux expérimentateurs. La plupart des systèmes non commerciaux, en contrepartie, sont incomplets, trop compliqués à adapter aux besoins de l'expérimentateur, inadaptés aux besoins de linguistes non-programmeurs, et souvent limités à une seule plateforme informatique. De plus, sachant que leurs réputations peuvent être mises en jeu par une exploitation mal avertie, peu de chercheurs produisants de tels systèmes permettent un emploi sans limite de leurs systèmes.

Un réseau de modules. Une solution à ces problèmes consisterait idéalement dans le développement de systèmes spécifiquement conçus pour l'expérimentation linguistique, modularisés et relativement simples à l'emploi. Ceci demanderait non seulement une collaboration importante entre linguistes et producteurs de systèmes, mais également une mise en place de formats de communication entre modules, ainsi qu'un système de rétroaction efficace entre client et producteur. Un tel réseau de modules pourrait être créé par différentes équipes dont chacune assume la responsabilité pour un ou plusieurs modules dans le cadre d'un grand système

collaboratif, un peu selon le modèle Linux[1]. Aucune équipe n'exerçant le contrôle sur la totalité du système, et aucune équipe n'ayant l'exclusivité d'un certain module, chaque équipe pourrait choisir d'être constructeur et/ou client d'autant de modules souhaités. Globalement, cette approche collaborative et interactive promeut l'émergence de modules concurrentiels qui deviennent graduellement de mieux en mieux adaptés aux besoins d'expérimentation de chaque groupe.

Certaines parties de ce réseau existent déjà. Les projets Mbrola, Festival et FreeTTS représentent des précurseurs importants à un tel réseau. Parmi les trois, le système de génération sonore Mbrola[2] se rapproche le plus fortement de la notion de « module » tel que proposée ici, et de ce fait, a su attirer une clientèle variée et fidèle au fil des années. Le projet Festival et son descendant indirect FreeTTS, par contre, sont conçus comme systèmes relativement monolithiques, difficiles à subdiviser en modules. Nous avons vainement essayé dans notre laboratoire, par exemple, d'utiliser Festival pour vérifier la prédiction des durées en *ms* pour des segments ou syllabes anglaises. De plus, l'adaptation à certaines autres langues, ou à une parole interactive, s'avère difficile dans un système non modulaire. Pour un linguiste, l'utilité d'un système de vérification linguistique repose fortement sur la modularité, la transparence et l'accessibilité de chaque composante de traitement.

3. PHÉNOMÈNES PARASITES, QUANTIFICATION DE L'ERREUR, TRANSPARENCE, PÉRENNITÉ ET RÉVERSIBILITÉ

Une fois que l'on dispose d'un système de synthèse, l'usage concret d'un tel système soulève des problèmes supplémentaires. Notamment, l'interprétation des *différences* entre ce que produit la machine et ce qu'aurait prédit notre hypothèse pose problème. Serait-ce la machine qui a fait défaut ou serait-ce notre hypothèse? Ou encore, est-ce que le critère de comparaison a été bien choisi? Le jugement perceptif de nos auditeurs est-il approprié à la tâche? Et la déviation obtenue avec un système de synthèse donné serait-elle la même avec un autre système ou avec la prochaine génération de systèmes? Et finalement, le processus d'abstraction prosodique est-il entièrement réversible? Il est important que les chercheurs souhaitant utiliser la

---

[1] Un tel réseau peut fonctionner même si les modules sont codés ou chiffrés. La seule chose qui importe est la bonne documentation et l'accessibilité des entrées-sorties et traitements. Contrairement à Linux, où l'ensemble du code est accessible à tous, une approche plus souple à l'égard de la confidentialité de la source promouvrait la contribution de laboratoires qui ne voudraient ou ne pourraient pas participer à l'expérience sous une règle d'« open source », vu leurs contrats de confidentialité avec différents fournisseurs ou contrats de subventions.

[2] http://tcts.fpms.ac.be/synthesis/mbrola.html, Faculté Polytechnique de Mons, Belgique.

simulation linguistique soient mis en garde par rapport à l'ensemble des problèmes que cette méthode soulève.

Phénomènes parasites. Quelle que soit la qualité ou la structuration du système de synthèse, la parole synthétique produite déviera du moins partiellement d'une parole produite par un locuteur humain dans un contexte semblable. Ceci pose un problème plutôt grave à l'expérimentateur: non seulement ses propres hypothèses et modules pourraient faire défaut, mais les modules extérieurs utilisés pour la vérification synthétique pourraient introduire des distorsions supplémentaires. Il s'agit du problème bien connu des phénomènes parasites. De plus, les divers systèmes de synthèse étant en évolution constante, le degré et la forme du bruitage varie fréquemment. Dans un système à bruitage variable et souvent mal documenté, il devient particulièrement difficile de séparer les aberrations de nos propres hypothèses de celles induites par nos instruments de vérification.

Classification et quantification de l'erreur. Un problème supplémentaire surgit lors de l'identification, classification et quantification de la déviation. A quel point est-ce qu'une déviation ou divergence devient une *erreur*? L'oreille humaine montre une assez grande tolérance, par exemple, à l'égard d'écarts de la fréquence fondamentale. A quel point peut-on dire que la machine a produit « un monstre prosodique » (terme favori du colloque sous-jacent à ce volume) ou seulement « un écart mineur » à la norme prosodique? Et qui se fera juge de la qualité du modèle prosodique global ou des problèmes parasites du système? Les locuteurs naïfs émettent souvent des jugements moins sévères et moins ciblés que les auditeurs expérimentés. Devrait-on devrait-on appliquer les critères les plus sévères?

Interaction, stylisations, transparence et pérennité. De plus, il est probable que ces deux problèmes cumulent, c'est-à-dire, que les phénomènes parasites interagissent avec le problème de la quantification de l'erreur. Quand la voix de la sortie synthétique est diforme, les jugements des erreurs ou écarts changent de caractère par rapport à une situation sans ou avec moins de distorsion.

Cette dernière observation soulève des questions importantes par rapport aux *stylisations* de la f0 effectués par un grand nombre de chercheurs. Selon Campione et al (2000), une stylisation sert « à remplacer une courbe f0 par une fonction numérique plus simple préservant l'information macroprosodique ». Cette opération sert à identifier les paramètres déterminants de la prosodie. Il s'agit d'une tradition lancée par 't Hart et Cohen du laboratoire IPO (Pays-Bas) durant les années '60-80, et reprise par d'autres auteurs, dont P. Mertens (ce volume). Par rapport au problème des parasites reliés à la perception, la question se pose: une stylisation de f0, concrétisée par une méthode simple (comme la PSOLA) et ne semblant pas introduire de modification significative lorsque jugée par des auditeurs naïfs, ne serait-elle pas

jugée significative quand elle est produite par une méthode plus performante (p.ex., la HNM[3]) et lorsqu'elle est jugée par des auditeurs avec une oreille particulièrement critique pour les paramètres prosodiques de la langue?

Si cela était le cas, une stylisation qui semblerait *transparente* (apparemment sans différences audibles) avec l'emploi d'une technologie x, pourrait bien devenir non transparente (introduire des différences audibles) lorsque transférée à la technologie x+1, spécialement si elle est jugée par des auditeurs sévères. Quelques expériences pilotes avec le recodage HNM nous a davantage convaincus de la vraisemblance de cet argument. Au cours de nos expérimentations avec certaines phrases du corpus Groult, nous avons entendu de nettes différences entre l'original et les copie-synthèses HNM effectuées sur la base d'une extraction par le logiciel Praat. Par contre, ces différences étaient à peine audibles via une reproduction Mbrola.

De plus, le maintien de la variabilité de la durée inter-cycle (du « jitter ») dans le signal est devenu incontournable avec HNM. Le jitter est de l'ordre d'un demi pour-cent dans les sections voisées du signal de parole d'une personne normale, alors que dans une stylisation simple de la courbe f0, il sera réduit à zéro. La suppression du jitter est audible en reproduction HNM, tandis que sa préservation sert à augmenter la qualité naturelle du signal. Peu importe si la majorité des stylisations obtenues dans les années '80 et '90 préservaient le jitter (nous soupçonnons que non), la prochaine génération de synthétiseurs imposera sans question qu'une courbe stylisée respecte le jitter naturel dans les parties voisées du signal.

Confronté à ce type de logique, on entend souvent deux contre-arguments: d'une part, ce ne seraient que les auditeurs exigeants qui entendraient ces phénomènes, et d'autre part, la majorité des auditeurs préfèreraient qu'une synthèse soit claire, plutôt que naturelle. Le domaine de la reproduction sonore fournit un précédent intéressant à considérer dans ce contexte. Les standards de qualité sonore ont subi des améliorations successives ces 50 dernières années, notamment avec l'adoption générale du disque à 33 tours/minute (après la 2e guerre mondiale), de la bande FM-stéréo (années 50, limite 9-15 kHz), du CD (années '80, 44 kHz), et de l'extension récente des standards d'enregistrement au-delà des 16 bits/44 kHz. A chaque point de cette évolution, ce n'étaient pas les « auditeurs naïfs » qui réclamaient l'amélioration du standard, c'était plutôt les « puristes », c.-à-d. les auditeurs avec entraînement musical et exerçant des jugements sévères. Cependant une fois généralement disponibles, les standards supérieurs se sont rapidement répandus dans le public général, au point qu'il aurait été impensable de retourner aux

---

[3] HNM: Harmonics and Noise Modelling, une approche de dé- et recomposition sinusoïdale du signal de la parole. Pour davantage d'explications et exemples, voir www.unil.ch/imm/docs/LAIP/LAIPTTS_signal_fr.htm et www.unil.ch/imm/docs/LAIP/LAIPTTS_signal_GE02_fr.htm.

standards antérieurs une décennie plus tard. De manière analogique, nous nous attendons à ce que toute avance vers une reproduction synthétique plus naturelle et de meilleure qualité s'imposera sans aucune question dans le long terme.

Ce processus évolutif des systèmes de synthèse illumine également la forte progression durant les années '90 de la *prédiction prosodique* par des systèmes stochastiques[4]. La logique à la base de la majorité des systèmes prédictifs des années '80 était *phonologique*, et pour une bonne partie d'entre eux, le fonctionnement était fondé sur un *codage symbolique* de la courbe f0. Un codage symbolique (p.ex. ToBI, Silverman et al 1992 ou INTSINT, Hirst & Di Cristo 1998) regénère une courbe qui serait, dans le cas idéal, considérée comme « linguistiquement équivalente » par les auditeurs (Campione et al, op. cit.). La prédiction prosodique va bien au delà d'une stylisation, car elle implique l'abstraction, la modélisation et la recréation en synthèse, de paramètres prosodiques et phonétiques.

Or si au milieu des années '90, nous avons insisté sur l'insuffisance des modèles phonologiques pour la prédiction du timing (Keller & Zellner 1996, 1998a, Zellner 1998), ainsi que sur les limites des symbolisations de type ToBI de la f0, c'était bien parce que la synthèse supérieure via Mbrola (Dutoit 1997) associée à l'excellente base originale F1 de Thierry Dutoit nous a mieux démontré que les systèmes précédents, les limites intrinsèques de tels modèles. Comme beaucoup d'autres, nous nous sommes rendus compte que les systèmes de prédiction phonologique pour le timing (largement inspirés par ailleurs par la phonologie américaine portant sur des langues à accentuation forte et distinctive) étaient inappropriés au français, et que les approches statistiques et neuromimétiques développées au début des années '90 pour l'anglais et l'allemand donnaient des résultats plus performants (Campbell 1992; Traber 1992). De plus, la dépendance de tels modèles sur des symbolisations f0 comme ToBI était remise en question, car celles-ci ne s'avéraient que très partiellement *réversibles*, en ce sens qu'il était difficile ou impossible de déduire des valeurs Hz systématiquement acceptables de l'inversion d'une transcription. Par contre, les approches statistiques ou

---

[4] Un système de prédiction prosodique à base stochastique emploie des méthodes statistiques ou neuromimétiques pour calculer les paramètres prosodiques (durées, f0, amplitude) d'un passage oral. En 1995, un bon système incorporant une prédiction stochastique pour les durées ainsi que la f0 était le système SVOX pour l'allemand (ETH, Zurich, Traber 1992, 1995), entretemps commercialisé (allemand, français, anglais; www.svox.ch). Les systèmes LAIPTTS pour le français (1995-97) et l'allemand (1998-2001, Keller & Zellner, 1998a, Siebenhaar-Rölli et al. 2002) étaient des systèmes hybrides, incorporant une prédiction statistique pour la durée et un système par règles pour la f0. La version française de LAIPTTS est actuellement en réfection et la nouvelle version utilisera des prédictions stochastiques pour la durée, la f0 et l'amplitude.

neuromimétiques (p.ex., Traber 1995) fournissaient des résultats étonnamment proches des modèles naturels.

Qualité du modèle et niveau d'abstraction. Ces développements posent plusieurs problèmes au linguiste. La première question concerne les meilleures performances des systèmes stochastiques. En théorie, une bonne grammaire phonologique, associée à une symbolisation adéquate, devrait fournir un tracé f0 convenable. Cependant, les systèmes de synthèse basés sur ce type de grammaire, p.ex. le système IMS de Stuttgart (Mayer 1995), ont souvent fourni des réalisations qui n'étaient clairement pas à la hauteur des attentes. A nos oreilles de langue maternelle allemande, par exemple, les exemples produits par le système de Stuttgart montraient des écarts majeurs entre la courbe f0 produite et celle qui aurait été acceptable du point de vue d'un auditeur de langue allemande[5]. L'objectif de « l'équivalence linguistique » n'était donc pas atteint par un tel système.

Est-ce que ce type de résultat remet en cause les notions de symbolisation linguistique ou de prédiction phonoloqique? Ce n'est clairement pas le cas, car le codage logique et numérique à la base des systèmes stochastiques n'est rien d'autre qu'une symbolisation plus fine et parfois un peu moins abstraite, et le jeu de prédiction stochastique ne diffère que par degré de raffinement d'un jeu de prédiction par règles phonologiques. La solution n'est pas la suppression, mais *l'amélioration* de la symbolisation et de la puissance prédictive des modèles phonologiques, au point où l'inversion de la chaîne de symboles permettrait de reconstituer parfaitement le signal. Cet effort est d'autant plus pressant que les paramètres et interactions mis en jeu par les modèles statistiques et particulièrement les systèmes neuromimétiques sont généralement moins explicites que les éléments constituant une théorie linguistique, rendant difficile la compréhension et la manipulation des différents prédicteurs prosodiques. Pour pouvoir gérer un ensemble de symboles et d'opérations prédictives de manière cohérente et transparente dans un système computationnel ou au sein d'une théorie linguistique, il faudra, au cours d'importants travaux à venir, identifier et réduire à un niveau symbolique abstrait et adéquat, les différents paramètres actifs et opérations prédictives d'une solution stochastique.

Réversibilité. Peu importe quand les systèmes à base symbolique finiront par « rattraper » la performance des modèles stochastiques, la simulation de la parole a déjà commencé par poser un défi de taille au linguiste: dorénavant il ne suffira plus qu'une hypothèse linguistique ou phonologique semble théoriquement plausible et qu'elle effectue des distinctions symboliques systématiques, elle a également intérêt à être *réversible,* en ce sens que les résultats de l'analyse devraient être reproductibles

---

[5] La plupart des exemples ont maintenant disparu du web, mais il reste un exemple sur la page www.kgw.tu-berlin.de/~felixbur/ttsDemos_ger.html.

(à un niveau de granularité encore à définir) par une synthèse de la parole. Quoiqu'il soit tout à fait légitime de continuer à considérer les modèles non réversibles comme de simples abstractions analytiques, du point de vue de l'histoire de la science, il ne fait pas de doute que les théories entièrement réversibles sont plus fortes que les théories partiellement ou non réversibles. Les meilleurs modèles sont clairement ceux qui ne décrivent pas seulement les phénomènes observés, mais qui de plus permettent une reconstitution (parfaitement) acceptable de la parole en synthèse.

Ces nouveaux standards d'acceptabilité théorique finiront par interpeller directement nos méthodes de recherche en linguistique, car l'observateur de l'histoire de la linguistique ne manquera pas à y voir un changement de paradigme pour notre science: après un siècle et demi d'insistance sur les critères de *différentiation*, *contrastivité*, et *cohérence interne* d'une théorie, la linguistique se trouve confrontée au nouveau critère nettement plus sévère de la *réversibilité* entre analyse et synthèse.

4. UNE APPROCHE MÉTHODOLOGIQUEMENT SAINE

Il est possible que les arguments précédents aient eu un effet dissuasif sur un certain nombre de lecteurs linguistes. La solution ne consiste pas à *éviter* la simulation, mais à l'utiliser de manière avertie, informée et prudente. Selon nous, l'utilisation de la synthèse s'allie avant tout à une méthodologie empirique impeccable. Dans ce contexte, voici quelques éléments qui nous semblent utiles à considérer.

Style de parole et locuteur. La première question concerne l'objet à simuler. Quelle parole faut-il reproduire? La parole lue? Conversationnelle? Interactive? Émotive? Et produite par qui? Un enfant? Un adulte âgé, homme, femme? La majorité des synthèses ne sont entraînées que sur la parole lue de phrases isolées par un adulte (souvent mâle), et ne s'adaptent que partiellement à d'autres styles de parole ou d'autres locuteurs. Jusqu'à l'arrivée de systèmes spécifiquement conçus pour d'autres objectifs, l'expérimentateur rencontrera nettement moins de problèmes parasites s'il se cantonne à la simulation de la parole lue par une voix adulte (mâle).

Jugements perceptifs. Le jugement des déviations prosodiques n'est pas facile, que ce soit sur la prosodie naturelle ou synthétisée. Sonntag (1999) indique que les difficultés entourant l'évaluation de la prosodie synthétique seraient dues à toute une série de facteurs, dont l'imprécision de définition des paramètres prosodiques, les origines souvent extrinsèques et pragmatiques d'une interprétation prosodique, la non-linéarité entre effets perceptifs pertinents et effets mesurables sur le signal, ainsi que les non-linéarités inhérentes aux systèmes de synthèse (p.ex., déviations aux extrêmes de manipulation de la f0). Face à ces questions, deux approches sont

pratiquées: les jugements perceptifs humains, et l'application de mesures objectives correspondant aux jugements perceptifs humains.

Avant de revenir sur les jugements perceptifs humains, illustrons brièvement la deuxième approche, très prometteuse. De bons exemples de cette approche sont les recherches par Klabbers et Veldhuis (1998) et Stylianou et Syrdal (2001) qui ont évalué différentes approches envers l'estimation de la différence perceptive entre deux segments de signal, en faisant intervenir plusieurs mesures de distance calculées sur des segments de signal, des spectres, des cepstres ou des transformées LPC. Dans les deux études, la distance Kullback-Leibler (une mesure d'entropie) semble le mieux approximer les jugements perceptifs humains, quand il s'agit de juger la gravité d'une perturbation dans le signal proche d'une concaténation entre deux segments de signal. Si ces résultats se confirment, ce type de correspondance permettra de plus en plus de remplacer les jugements perceptifs onéreux par des mesures calculées automatiquement sur le signal.

Si par contre l'expérimentateur choisit des jugements perceptifs humains, quelques considérations supplémentaires s'avèrent utiles.

Juges naïfs ou expérimentés? La psychologie expérimentale nous a appris à nous fier aux juges naïfs. Les réflexions de ce texte nous suggèrent que cela ne soit pas toujours la seule manière d'examiner les faits prosodiques, particulièrement si une amélioration d'un système de synthèse est en jeu. Parfois l'application de jugements aussi sévères que possibles prépare le chemin vers la prochaine génération de systèmes. Cette réflexion s'apparente à la précédente: il importe non seulement de spécifier le style de parole et le locuteur de la simulation, mais également l'autre partenaire de la transmission orale, c.-à-d. l'auditeur ainsi que sa qualité d'écoute.

Méthodes appropriées. Comme indiqué ci-haut, l'utilisation d'une méthodologie de modification du signal aussi « transparente » que possible favorisera la pérennité des résultats obtenus. Au besoin, il peut s'avérer préférable de modifier (légèrement) un signal existant que de le resynthétiser à partir de diphones, puisqu'on évite ainsi les déviations sonores induites par la concaténation de diphones, généralement obtenus dans différents contextes phonétiques et modifiés pour permettre leur concaténation.

En outre, les meilleures méthodes expérimentales sont toujours celles qui ciblent clairement un aspect spécifique du système linguistique. En ce qui concerne les jugements sur la prosodie ou sur l'émotivité en parole, il faudrait avoir recours à des méthodes capables d'évaluer des faits prosodiques indépendamment des faits segmentaux de la parole. Sonntag (op. cit.) a utilisé pour cela une *méthode délexicalisante*, la méthode PURR qui substitue des ondes sinusoïdales aux ondes de la parole. Cette méthode ne retient que la mélodie, le rythme et l'amplitude d'un

énoncé et élimine tout effet segmental. De plus, cette méthode ne semble pas déformer l'expérience perceptive, car il existe des indications provenant de mesures MRI que les stimulus délexicalisés activent des régions neurocérébrales semblables à celles activées par des stimulus normaux (Sonntag, op. cit., p. 177, Kotz et al. 2001). Vu la comparabilité directe entre les stimulus originaux et délexicalisés, tout effet secondaire dû à l'interprétation du message par l'auditeur serait supprimé. Ce type de méthode évite donc les problèmes parasites soulevés par la méthodologie de la lecture de phrases sans sens utilisée par certains auteurs.

Prudence et avertissement public. Finalement, la préoccupation de tout scientifique reste la prudence à l'égard des différents facteurs parasites. Un travail de simulation gagne en crédibilité s'il évoque toute source d'erreur et d'interprétation possible. Ceci se traduit par des indications explicites sur les conditions précises de la simulation, ou par l'affichage de prudence appropriée lors de l'interprétation de résultats. Une adhérence générale à cette pratique ne manquera pas d'avoir un effet bénéfique sur l'évolution de l'instrument de simulation. Quand les effets parasites sont publiquement affichés, leur quantification deviendra plus urgente et leur élimination successive sera encouragée. Par ce biais, le développement de systèmes synthétiques de plus en plus « transparents » sera sans doute facilité. Indirectement, ceci aura du moins autant d'effets bénéfiques pour le développement de l'expérimentation linguistique que pour celui de la synthèse de la parole.

Eric Keller[6]
Laboratoire d'analyse informatique de la parole (LAIP)
Informatique et Méthodes Mathématiques, Faculté des Lettres
Université de Lausanne, 1015 Lausanne, Suisse
eric.keller@imm.unil.ch

REFERENCES BIBLIOGRAPHIQUES

CAMPBELL, W.N. 1992. *Multi-level Timing in Speech*. PhD. Thesis, University of Sussex.

[6] Eric Keller dirige un laboratoire (LAIP, www.unil.ch/imm/docs/LAIP/LAIP_fr.html) qui a développé des synthèses du français et l'allemand standard. Des systèmes pour le suisse-allemand, le latin classique et l'anglais sont actuellement en élaboration. Ces efforts sont entrepris avant tout pour la vérification d'hypothèses portant sur le langage et la parole.

CAMPIONE, E., Hirst, D., & Véronis, J. 2000. « Stylisation and symbolic coding of f0: comparison of five models », dans A. BOTINIS (dir.), *Intonation: Models and Theories*, Dordrecht: Kluwer Academic Publishers, 185-208.

DUTOIT, T. 1997. *An Introduction to Text-to-Speech Synthesis.* Kluwer.

HIRST, D. & Di CRISTO, A. 1998. « A survey of intonation systems », dans Hirst & Di Cristo (dir), *Intonation Systems: A Survey of Twenty Languages,* Cambridge: Cambridge University Press, 1-44.

KELLER, E. 2001. « Towards Greater Naturalness: Future Directions of Research in Speech Synthesis », dans E. KELLER, G. BAILLY, A. MONAGHAN, J. TERKEN & M. HUCKVALE (dir.), *Improvements in Speech Synthesis*, Wiley & Sons, 3-17.

KELLER, E., & ZELLNER, B. 1996. A timing model for fast French. *York Papers in Linguistics,* 17, 53-75, University of York.

KELLER, E., & ZELLNER, B. 1998a. « Motivations for the prosodic predictive chain. *Proceedings of ESCA Symposium on Speech Synthesis* », Paper 76, 137-141, Jenolan Caves, Australia.

KELLER, E., & ZELLNER, B. 1998b. « Préface: Les défis actuels en synthèse de la parole », dans E. Keller, & B. Zellner (dir.), *Études des Lettres, vol 3.: Les défis actuels en synthèse de la parole*, Lausanne: Université de Lausanne, 3-7.

KLABBERS, E., & VELDHUIS, R. 1998. « On the reduction of concatenation artefacts in diphone synthesis », *ICSLP*, 1983-1986, Sydney, Australia,.

KOTZ, S.A., MEYER, M., ALTER, K., BESSON, M., von CRAMON, D.Y, & FRIEDERICI, A.D. 2001. « Differentiation of affective prosody: An event-related fMRI study », *Cognitive Neuroscience Society 2001 Proceedings*, New York.

MAYER, J. 1995. *Transcription of German Intonation: The Stuttgart System*, WWW version, May 15, 1995, www.ims.uni-stuttgart.de/phonetik/joerg/labman/STGTsystem.html.

SIEBENHAAR-RÖLLI, B, ZELLNER KELLER, B., & KELLER, E. 2001. « Phonetic and Timing Considerations in a Swiss High German TTS System », dans E. KELLER, G. BAILLY, A. MONAGHAN, J. TERKEN & M. HUCKVALE (dir.), *Improvements in Speech Synthesis*, Wiley & Sons, 165-175.

SILVERMAN, K., BECKMAN, M. E., PITRELLI, J., OSTENDORF, M., WIGHTMAN, C., PRICE, P., PIERREHUMBERT, J. & HIRSCHBERG, J. 1992. « ToBI: A standard for labelling English prosody », *ICSLP*, 2, 867-870.

SONNTAG, G.P. 1999. *Evaluation von Prosodie.* Thèse de doctorat, Friedrich-Wilhems Universität Bonn, Aachen, RFA: Shaker Verlag.

STYLIANOU, Y. & SYRDAL, A.K. 2001. « Perceptual and objective detection of discontinuities in concatenative speech synthesis », *ICASSP*, Salt Lake City, Utah.

TRABER, C. 1992. « f0 Generation with a data base of natural f0 patterns and with a neural network », dans G. BAILLY, C. BENOÎT, & T.R. SAWALLIS, dir., *Talking Machines: Theories, Models and Designs,* Elsevier, 287-304.

TRABER, C. 1995. SVOX: *The Implementation of a Text-to-Speech System for German*, Zürich: ETH-v/d|f-Hochschulverlag AG.

ZELLNER, B. 1998. *Caractérisation et prédiction du débit de parole en français. Une étude de cas.* Thèse de Doctorat. Faculté des Lettres, Université de Lausanne.

5. RÉACTIONS À L'ARTICLE D'ERIC KELLER

Brigitte ZELLNER KELLER
Université de Lausanne

5.1. Introduction

Eric Keller a développé dans ce volume une solide argumentation sur la vérification de nos hypothèses linguistiques par la synthèse de la parole. Je montrerai dans un exemple comment ce nouveau paradigme de la simulation peut faire surgir de nouvelles questions scientifiques et je conclurai en intégrant ce paradigme dans le contexte plus vaste du développement des sciences cognitives.

5.2. Y a t-il une *méthode* de simulation du rythme ?

La simulation du rythme de la parole dans une synthèse à partir du texte s'obtient au moyen d'une procédure qui se subdivise typiquement en deux temps (voir les systèmes de Genève, Leuven, Caen, Aix). Tout d'abord, la manière dont les mots sont agglutinés dans le temps, est prédite à partir de l'analyse des *structures intonatives* de l'énoncé. Ensuite, la simulation mélodique et temporelle (i.e., manipulation des paramètres de F0 et de durées) est générée grâce à une *inversion* semi-transparente *d'éléments phonologiques*, comme par exemple la présence d'une syllabe accentuée. Il y a donc, au delà de la diversité des synthèses une unanimité méthodologique concernant la simulation du rythme. (Est-ce cette méthode est optimale pour simuler le rythme?)

Cette méthode fonctionne assez correctement pour la lecture de phrases isolées ou de petits textes. Peut-on généraliser cette méthode de simulation au delà de la lecture d'énoncés de l'écrit? (La parole recouvre en effet bien d'autres modes que celui de la lecture de textes écrits.) Par exemple, un bon *test de simulation du rythme de la parole* consisterait à synthétiser des énoncés produits spontanément à l'oral

(donc acceptables), mais pas forcément conformes à la syntaxe de l'écrit. Par exemple :

> "Le problème nous ce qui nous fait peur c'est qu'elle se retrouve dans une structure inadaptée". (Exemple transcrit d'une émission de télévision)

Ce simple exemple pose des problèmes rythmiques notables à l'analyse phonosyntaxique habituelle. Premièrement, si la transcription de cet énoncé est introduite sans ponctuation (pour ne pas fausser le test en modifiant un énoncé oral pour l'adapter à l'écrit), beaucoup de systèmes ne sont plus capables de prédire correctement les césures majeures. De ce fait, la gestion des pauses et la courbe de déclinaison de F0 s'en trouvent affectés. Ensuite, l'ordre des mots à l'oral ne répond pas toujours à l'ordre syntaxique de l'écrit, et dans ce cas ce sont les agglutinations de mots avec les effets locaux d'accélération et de ralentissement de débit qui sont affectés, générant des effets de disfluences. Lorsqu'il s'agit de simuler un rythme de parole à partir de la transcription d'un énoncé produit spontanément par un locuteur, les systèmes de synthèses sont donc souvent mis en difficulté, et cela a été confirmé au colloque de Genève avec les expérimentations sur le corpus Groult. Est-ce à dire que les modèles prosodiques sous-jacents dans tous ces systèmes seraient "faux"?

On pourrait plutôt questionner cette *méthode* typique de génération prosodique qui est utilisée sans changement, quel que soit le type de parole simulée : 1/ analyse intonative; 2/ analyse phonologique. Cette procédure rigide ne permet pas une simulation de parole proche du style spontané car elle induit des relations figées entre intonation et structure temporelle. Or, la structure temporelle, (en particulier les groupes de mots), n'est *pas toujours* sous la dominance de la structure intonative. Fougeron *et al.* (1995) et Zellner (1998) ont montré qu'en français, le débit rapide est caractérisé par une diminution du nombre de groupes prosodiques par rapport au débit lent. La modification de l'agglutination des mots constitue un véritable changement de rythme et non pas seulement un simple changement de tempo. Cette modélisation du changement de rythme nécessiterait de pouvoir simuler que l'organisation temporelle affecte les structures intonatives, ce qui n'est pas possible avec les architectures précédemment décrites. J'ai montré dans mes travaux qu'il était possible de prédire correctement le rythme de la parole et ce dans différents débits, sans utiliser de connaissances sur la structure intonative (Zellner, 1998, Zellner Keller, 2002). En d'autres mots, la *méthode* de génération prosodique qui est privilégiée dans la plupart des systèmes, entrave par sa rigidité une simulation adéquate du changement de débit, fait fréquent de l'oral.

Une reconsidération de l'importance des facteurs temporels en prosodie s'impose d'autant plus qu'on souhaite s'éloigner de la simulation de la lecture de phrases isolées pour aller vers une simulation de parole plus vivante et plus

individualisée, en d'autres mots vers une parole plus "incorporée", où les phénomènes de respiration, d'accélération et de ralentissement seront mieux captés.

6. POUR UNE PAROLE "INCORPORÉE"

Deux réflexions peuvent être tirées du constat ci-avant. Premièrement, sans ce problème de méthode de génération prosodique en synthèse de parole, la question plus fondamentale des relations entre la composante temporelle et la composante intonative dans un module prosodique n'aurait sans doute jamais été posée, car difficile ou impossible à tester. C'est ainsi que la simulation permet de repousser les limites de nos explorations.

Une deuxième réflexion (que suggère ce problème de rigidité dans la génération prosodique) est celle de notre héritage de linguistes. Les sciences du langage se sont développées à une époque où l'on modélisait la *cognition*, et en particulier les processus linguistiques, avec des *modèles mécaniques*. Les processus cognitifs supposés universels étaient alors réduits à des ensembles d'opérations sur des symboles discrets identifiés, où la composante du temps n'était pas pertinente en dehors de l'application de la règle de séquentialité. Aujourd'hui, nos représentations de la cognition ont beaucoup évolué grâce aux techniques d'investigation sur le cerveau (voir Besson dans ce volume) et aux travaux de ces vingt dernières années en psychologie expérimentale, mais nos modèles en sciences de la parole ne reflètent pas encore résolument cette évolution. Ils n'ont par exemple pas encore véritablement intégré la complexité de la composante du temps. Il faut souhaiter que le paradigme de la simulation de la parole encouragera à investiguer davantage la dimension temporelle en prosodie car elle permettra d'ancrer les faits de parole dans notre spécificité d'être humain communiquant.

Brigitte Zellner Keller
Laboratoire d'analyse informatique de la parole (LAIP)
Informatique et Méthodes Mathématiques, Faculté des Lettres
Université de Lausanne
1015 Lausanne, Suisse
Brigitte.ZellnerKeller@imm.unil.ch

REFERENCES BIBLIOGRAPHIQUES

FOUGERON, C. & JUN, S-A. 1995. Properties of French intonation at fast speech rate. *XIIIème Congrès International des Sciences Phonétiques, 3* (pp. 488-491). Stockholm.

KELLER, E. 2001. « Towards Greater Naturalness: Future Directions of Research in Speech Synthesis », dans E. KELLER, G. BAILLY, A. MONAGHAN, J. TERKEN & M. HUCKVALE (dir.), *Improvements in Speech Synthesis*, Wiley & Sons, 3-17.

ZELLNER, B. 1998. *Caractérisation et prédiction du débit de parole en français. Une étude de cas.* Thèse de Doctorat. Faculté des Lettres, Université de Lausanne.

ZELLNER, B. 2002. « Revisiting the Status of Speech Rhythm », dans Bernard Bel & Isabelle Marlien (eds.). *Proceedings of the Speech Prosody 2002 conference*, 11-13 April 2002.(pp. 727-730) Aix-en-Provence: Laboratoire Parole et Langage. ISBN 2-9518233-0-4.