SYNTHÈSE DE LA PAROLE

Christophe d'ALESSANDRO

1. INTRODUCTION

1.1. Synthèse à partir du texte

L'objet de la synthèse de la parole est de calculer automatiquement le signal de parole qui correspond au mieux à un énoncé écrit et au contexte d'élocution. Les sources de l'énoncé écrit sont diverses. Le texte peut être plutôt littéraire et bien rédigé, tiré d'un livre ou d'un journal. Il peut être produit automatiquement, issu par exemple d'une base de données, d'un système d'information, d'un système de réponse vocale, de dialogue, d'une borne interactive, d'un assistant personnel ou d'un jeu vidéo. Le texte peut provenir de sources écrites mal structurées ou mal rédigées, comme des courriers électroniques, des SMS, des traductions automatiques, des documents butinés sur la toile, ou tout simplement être saisi au clavier d'un ordinateur. En plus de la variabilité des sources de texte, la parole produite par un système automatique, comme toute parole, est sujette à la variabilité de l'élocution. Cette parole est située dans un contexte d'élocution particulier, qui comprend la voix du locuteur virtuel, sa façon de parler, voire ses attitudes ou son état émotif supposé, ainsi que la situation d'élocution, par exemple téléphonique ou directe, en situation de dialogue ou d'information.

1.2. Architecture d'un système de synthèse

Tout système de synthèse de parole s'articule en deux phases, une phase d'analyse du texte et du contexte, puis une phase de calcul du signal sonore. Ainsi les deux pôles de la synthèse de la parole sont du côté de la linguistique (l'analyse et l'interprétation du texte écrit), et du côté de l'acoustique (la prédiction des paramètres acoustico-phonétiques du son et la synthèse du signal).

Sur le versant du texte, la qualité de synthèse dépend de la capacité d'interpréter le texte écrit, tout comme le fait un orateur. Cela implique de « comprendre » le texte, ses nuances et ses connotations, la situation du discours et l'acte de parole à effectuer. Il est difficile d'imaginer qu'un tel programme soit actuellement à la portée d'une machine. Néanmoins la synthèse de la parole à partir du texte a suscité la mise en œuvre des procédures automatiques pour plusieurs aspects de l'analyse linguistique. Ce besoin d'automatisation a par contrecoup mis en avant des problématiques jusqu'alors plus ou moins ignorées, et a fonctionné souvent comme une mise à l'épreuve, par le biais de l'écoute, de bien des propositions théoriques. Sur le versant du son, la qualité de synthèse dépend de la représentation du signal, à travers un modèle de synthèse, ou bien par des procédures de codage, si on assemble directement des échantillons. Cette partie de la synthèse de parole partage certaines méthodes avec la synthèse musicale, ou le traitement audionumérique d'une façon générale, mais peu de choses avec la linguistique.

1.3. Les unité du texte, les unités phoniques

Suivant l'axe syntagmatique, un texte se décompose en unités juxtaposées et emboîtées : le texte, les paragraphes, les phrases, les propositions, les mots, les morphèmes, les graphèmes, incluant lettres, chiffre, signes de ponctuation et signes divers. La structure d'un texte fait apparaître les plus petites unités graphiques, les caractères ou graphèmes, qui sont groupés en mots, séparés par des espaces ou autres séparateurs, qui sont eux-mêmes agencés en propositions et en phrases par les signes de ponctuations, et enfin en paragraphes par les sauts de lignes et marques de paragraphe. De même la structure syntagmatique d'un énoncé oral se décompose en unités phoniques juxtaposées et emboîtées : paragraphe, phases prosodiques, groupes de souffles ou groupe prosodiques, syllabes, et enfin phonèmes. La correspondance entre les unités du texte et les unités phoniques n'est ni simple ni univoque. C'est bien une des difficultés du traitement de la parole en général, et de la synthèse en particulier.

Au niveau phonique, la synthèse de parole considère en général deux échelles de temps principales, l'échelle du « segment » acoustique ou phone (réalisation acoustique de l'unité abstraite « phonème ») et l'échelle « suprasegmentale » ou prosodique. A l'échelle prosodique il faut moduler la mélodie, ou intonation, le rythme, le débit de parole et le groupement des accents, la qualité vocale. La « double articulation » du langage, une première articulation des sons en phonèmes et une seconde articulation des phonèmes en mots, rend bien compte du niveau segmental, mais ne dit rien du niveau prosodique. Un énoncé dépourvu de prosodie, peut à la rigueur transmettre une chaîne de phonèmes et donc de mots. Cependant, dans bien des cas certains mots seront difficiles à distinguer, et les relations entre mots difficiles à comprendre. De plus il y manquera les nuances et l'expression, qui dans bien des cas constituent une partie fondamentale, voire prépondérante,

de la communication parlée. En effet, parler n'est pas toujours transmettre une information à la façon d'un code de communication, mais le plus souvent communiquer de façon naturelle un mélange complexe d'informations linguistiques, expressives, voire comportementales, pour lequel la nuance prosodique est fondamentale. Le niveau segmental relève des théories phonologiques. Pour le niveau prosodique, c'est moins évident. Certaines approches tentent de définir une phonologie de l'intonation dans des termes calqués sur la phonologie segmentale, mais d'autres approches considèrent plutôt que la prosodie est d'une autre nature, comme un geste continu sans possibilité de décomposition en unités phonologiques discrètes.

1.4. L'instrument vocal

Le langage est longtemps resté exclusivement oral, parole portée par la voix. En fonction des possibilités sonores offertes par l'appareil vocal, chaque langue a adopté un ensemble particulier de sons distinctifs (ou phonèmes) qui permettent de construire tous les mots. En Français, on considère qu'il existe une trentaine de phonèmes. Au-delà ou en deçà du langage, l'appareil vocal est également utilisé comme moyen de communication non-linguistique, dans le chant ou les vocalisations expressives non articulées. En tant que système de production du son, la voix appartient à la famille des instruments à vent. Comme eux, la voix peut se décomposer en une source sonore et un corps sonore. Le rôle de la source sonore est de produire le son, par la vibration des plis vocaux (comme pour les voyelles), ou par divers bruits continus ou impulsionnels. Le corps sonore de l'appareil vocal, le conduit vocal, est court : une moyenne d'environ 17 cm chez l'homme adulte. Ce conduit vocal forme une cavité très plastique dont la forme est donnée par la position des articulateurs (mâchoires, lèvres, langue, uvule). Il peut être ainsi raccourci, rallongé, ouvert, fermé, divisé en plusieurs cavités. Son rôle est de former le spectre, de colorer le son produit par la source sonore, par ses positions et ses mouvements.

Le modèle acoustique de la voix est un modèle source/filtre. De façon simplifiée, la source donne au son sa hauteur, sa force, et un timbre initial : c'est le domaine de la prosodie. Le conduit vocal, filtre acoustique, colore le son de la source, les mouvements du conduit vocal permettent d'articuler les sons de la parole : c'est le domaine des phonèmes.

2. LES PARADIGMES DE SYNTHÈSE À PARTIR DU TEXTE

Supposons pour le moment que le texte à synthétiser soit sous la forme d'une suite de phonèmes, jointe à des consignes prosodiques. Plusieurs paradigmes ont été proposés pour calculer le signal vocal à partir de cette représentation du texte. Ces méthodes sont issues des recherches en informatique, croisées dans certains cas avec des théories linguistiques. Les principaux paradigmes sont les approches utilisant la concaténation d'échantillons, les approches par règles, les approches probabilistes.

2.1. Paradigme de concaténation

La première idée pour synthétiser de la parole a été tout simplement d'enregistrer des échantillons correspondant aux phonèmes, de les « concaténer » (les assembler) et de les restituer directement, sans modèle de l'appareil vocal. Dès les premiers essais dans les années 1950, on s'est aperçu que les théories phonologiques ne pouvaient pas être utilisées directement dans le domaine acoustique. Contrairement aux prédictions de la phonologie : concaténer des phonèmes ne produit pas de la parole intelligible. Il faut pour cela utiliser des unités phoniques qui comprennent la coarticulation entre phonèmes, donc au minimum les « diphones » (segments sonores comportant la transition entre phonèmes). Après avoir enregistré toutes les combinaisons de phonèmes, (en Français, environ 1 200 diphones, soit 2 à 3 minutes de son) la parole est reconstituée par concaténation de la chaîne de diphones correspondant à l'énoncé à synthétiser. Il n'y a pas de modèle acoustique pour la synthèse, mais un « vocodeur », un système de codage des échantillons de la voix. Après un début difficile à cause de la piètre qualité sonore des premiers vocodeurs, la synthèse par diphones a dominé le « marché » de la synthèse de la parole dans les années 1990. Dans un tel système, la prosodie doit être calculée et appliquée sur la séquence de diphones, par une méthode de modification prosodique.

Le principe de la synthèse par enregistrement de corpus et concaténation d'unités phoniques s'est généralisé pour aboutir dans les années 2000 aux systèmes actuels de synthèse par sélection et concaténation. Un corpus relativement long (une ou plusieurs heures de parole) est étiqueté en phonèmes et analysé en termes prosodiques. Lors de la synthèse, on va rechercher dans ce corpus les segments phoniques les plus longs qui correspondent à l'énoncé à synthétiser. Comme il n'y a en général pas une solution unique, une méthode d'optimisation est utilisée pour trouver la suite de segments qui maximise un critère de qualité de la synthèse. Il est remarquable que, dans de tels systèmes, la prosodie ne soit pas calculée explicitement. Si les segments sont suffisamment longs (par exemple un ou plusieurs mots), et si les contextes syntagmatiques sont bien spécifiés (place dans la phrase, ponctuation) la prosodie portée par la concaténation des segments est suffisamment réaliste.

Ce paradigme est celui de la reconnaissance des formes (« pattern matching »), dans lequel on cherche la meilleure correspondance entre ce qu'il faut synthétiser, et ce que contient le corpus. On fait l'hypothèse que le corpus est suffisamment vaste pour que, à la limite, tout ait déjà été prononcé, par petits morceaux. Cet espoir est évidemment déçu en pratique, mais cependant la qualité de synthèse demeure excellente, car il s'agit d'échantillons naturels. Augmenter la taille du corpus n'augmente pas vraiment la qualité. Le problème de ce type de synthèse est qu'elle est mal adaptée à la nuance et à l'introduction de variabilité. D'une part pour obtenir une voix de qualité différente il faut réenregistrer un nouveau gros corpus, ce qui en pratique limite à quelques unités les types de qualité vocale possibles. D'autre part, passer de façon continue d'un type de voix à

l'autre, par exemple réaliser une voix qui devient plus faible de façon réaliste, est difficile. La synthèse manque en somme de flexibilité.

2.2. Le paradigme génératif

L'approche générative en synthèse, ou « synthèse par règles », est apparue à peu près en même temps que la synthèse par concaténation. Mieux fondée dans les théories linguistiques, cette approche met en œuvre les propositions de la phonétique acoustique, de la phonologie et de la grammaire générative. Du point de vue acoustique, le son est calculé par un synthétiseur à formants, dispositif qui simule le modèle source/filtre de l'appareil vocal. Les paramètres de ce modèle sont calculés par des règles (une grammaire « générative ») qui associent à chaque phonème des paramètres acoustiques. Par exemple les voyelles seront définies par les valeurs cibles de leurs « formants » spectraux. Les consonnes par les lieux de focalisation de leurs formants, les vitesses de transitions et autres paramètres acoustiques spécifiques. La prosodie est elle-même calculée par des règles de réécriture, qui associent aux consignes d'entrée une courbe mélodique, en fonction d'un certain modèle prosodique. Les règles sont élaborées par une analyse attentive d'un petit corpus, dont les échantillons sonores ne sont pas utilisés comme tels pour la synthèse, mais seulement comme source de connaissances. Le travail d'analyse permet de concentrer dans un jeu de règles la connaissance explicitement acquise par l'expertise accumulée. La synthèse par règles a obtenu dans les années 1970-1990 un bon niveau de qualité. Elle est économique et élégante. Cependant, sa qualité sonore a été dépassée par celle de la synthèse par corpus.

Le paradigme génératif est dérivé de celui de l'intelligence artificielle, avec le développement de règles déterministes et explicites. Il est issu de modèles pour les processus de décision ou d'analyse, mais il montre clairement des limites dans le cas de la synthèse de la parole. Les règles ne sont pas vraiment adaptées à la simulation du contrôle moteur complexe mis en œuvre lors de la phonation. Dans le principe, ce paradigme est flexible, mais en pratique le développement de règles demande un effort trop important pour être réalisable sur une vaste échelle. Il n'est pas possible de traiter un vaste corpus par une analyse explicite. Les avantages de cette méthode sont une grande économie en terme d'empreinte mémoire (taille du logiciel), la lisibilité et la réutilisabilité des règles, l'apport de connaissances explicites.

2.3. Le paradigme probabiliste

L'approche paramétrique statistique combine certains éléments des deux approches précédentes, mais dans une perspective nouvelle. Comme la synthèse par concaténation, l'approche probabiliste utilise des procédés d'apprentissage automatique et de gros corpus, et non une expertise explicite obtenue par observation d'un petit corpus. Mais comme en synthèse par règles, le synthétiseur utilise un modèle paramétrique source/filtre de l'appareil vocal, autorisant une grande flexibilité à la synthèse. Ce paradigme, utilisé depuis une trentaine d'années en reconnaissance de la parole, s'inspire des travaux anciens du mathématicien russe Andreï Markov. La parole est considérée comme une chaîne d'éléments, dont les probabilités de succession dépendent fortement de leur voisinage, et peuvent être apprises par la statistique. Markov travaillait sur la probabilité de succession des lettres dans les romans. Ces « chaînes de Markov » ont depuis été généralisées à tous les niveaux linguistiques, probabilités d'enchaînement des paramètres spectraux du modèle source / filtre, d'enchaînement des phonèmes, des mots, des syntagmes, des accents mélodiques.

Après avoir enregistré et étiqueté un corpus de parole de grande taille, on considère que les phonèmes et les courbes prosodiques observés ont été produits par des modèles de Markov sous-jacents (ou « cachés »). On apprend alors les statistiques de ces modèles cachés en effectuant des comptages sur le corpus. Au moment de la synthèse, étant donné une suite de phonèmes et des consignes prosodiques, les paramètres acoustiques les plus probables sont calculés par des modèles de chaînes de Markov appris lors de « l'entraînement » du système sur le corpus. Le son est calculé par un modèle acoustique source/filtre de synthèse, comme en synthèse par règles.

Ce paradigme est celui de l'apprentissage probabiliste. Étant donné un corpus, aussi gros que possible, on cherche à la synthèse la façon la plus probable dont l'énoncé aurait été prononcé, en fonction de ses spécifications (par exemple prosodiques ou expressives), s'il s'était trouvé dans le corpus. Cette approche est très flexible, puisqu'il suffit de changer les spécifications pour changer le résultat. L'apprentissage d'une nouvelle voix ou de nouveaux contextes est rapide et aisé. Ce paradigme a montré sa puissance dans le cadre de la reconnaissance de parole, et devrait occuper la scène de la synthèse pour les quelques années à venir. La qualité est pour le moment inférieure à celle de la synthèse par concaténation directe, mais la flexibilité, l'élégance mathématique et l'économie de ce paradigme de synthèse lui donnent un fort potentiel de développement.

3. LES ANALYSES LINGUISTIQUES

Les différents paradigmes de synthèse décrits précédemment n'opèrent pas directement sur le texte d'entrée, mais sur une représentation phonétique du texte, c'est-à-dire une chaîne de phonèmes enrichie d'étiquettes ou de consignes prosodiques plus ou moins élaborées. Cette représentation doit être dérivée du texte à l'aide d'une suite d'analyses linguistiques.

3.1. Normalisation du texte

La première étape, une fois le texte saisi, est d'identifier les unités lexicales, les mots, ainsi que la ponctuation. Lorsque les systèmes de traitement automatique de la parole se sont trouvé confrontés à des sources de textes réelles et variées, il a fallu s'occuper des nombreuses « anomalies » graphémiques des textes, et de leur « normalisation ». Il faut transcrire en toutes lettres les sigles, les abréviations, les

acronymes, les diverses sortes de mots composés, les symboles, les divers types de nombres et de chiffres. Les signes de ponctuation et les caractères typographiques non alphanumériques doivent être analysés.

Tout cela nécessite des prétraitements sophistiqués pour former une suite de « lexèmes », ou unités lexicales à partir du texte d'entrée. De nombreuses conventions de notations, implicites, doivent être explicitées et mises en algorithmes. L'apparition de textes électroniques peu ou mal structurés, utilisant des conventions peu formalisées, introduit des « fautes » plus ou moins systématiques, par rapport aux règles grammaticales ou typographiques. Ainsi, courriers électroniques, traductions automatiques, textes sans accents, SMS, etc., posent des problèmes spécifiques et à chaque fois nouveaux, que de nouveaux algorithmes doivent être capables de traiter automatiquement.

3.2. Analyse lexicale

Le prétraitement a transformé un texte d'entrée en une suite d'unités lexicales et de marques de ponctuation. L'analyse lexicale consiste à analyser les informations associées à chaque lexème.

Le lexique contient les classes lexicales de chaque lexème, avec des conventions spécifiques à chaque système. La classe lexicale peut se réduire à une information fonctionnelle, par exemple « mot outil/mot plein », ou à des catégories grammaticales plus fines, comme les parties du discours.

La taille des lexiques varie significativement suivant les systèmes de synthèse. Certains systèmes utilisent des lexiques restreints, par exemple les mots outils et les verbes (quelques centaines à quelques milliers de mots). Les lexiques complets, contenant toutes les formes fléchies pour les mots et l'ensemble de leurs dérivations morphologiques se composent de 400 000 à 1 000 000 de mots et de locutions, à comparer à un dictionnaire de langue qui comporterait environ 50 000 articles. Notons que la catégorie grammaticale d'un lexème hors contexte est très souvent ambiguë. La recherche lexicale peut être accompagnée d'une d'analyse morphologique, afin de décomposer le lexème en morphèmes, correspondant aux préfixes, suffixes, désinences et racines. Le lexique comporte aussi les exceptions, les mots qui ne se décomposent pas suivant les règles morphologiques du français standard.

3.3. Analyse syntaxique

Les lexèmes étant pourvus d'une ou plusieurs catégories grammaticales, l'analyse syntaxique doit d'une part lever les ambiguïtés, et d'autre part construire les constituants syntaxiques. Pour affecter la catégorie de chaque mot, on utilise des règles contextuelles ou des modèles statistiques, en s'appuyant sur les catégories possibles des mots adjacents. Les règles contextuelles décrites par la grammaire de la langue demandent une expertise linguistique. À l'inverse, les modèles probabilistes ne demandent qu'une connaissance sommaire de la langue, puisqu'il s'agit d'observer et d'apprendre sur des corpus les successions de catégories les plus probables. La suite de catégories grammaticales permet de grouper les lexèmes en constituants syntaxiques. Ce découpage, ou « parenthésage » de la phrase, permettra de proposer une structuration prosodique de la phrase. Le calcul des relations de dépendance entre les constituants syntaxiques est difficile à automatiser : il reste en général sommaire dans les systèmes de synthèse, bien qu'il puisse être utile pour le calcul de l'accentuation.

3.4. Phonétisation

La forme orthographique des mots, particulièrement en français, ne rend pas directement compte de leur prononciation. Une étape de transcription graphème-phonème, ou phonétisation, est nécessaire afin de transformer le texte en une suite de phonèmes. La phonétisation est généralement réalisée grâce à un système de règles combiné à un lexique. Parmi les problèmes de phonétisation difficiles, on rencontre les noms propres, les mots nouveaux et inconnus, les mots de provenance étrangère, les variantes de prononciation. La phonétisation pose des questions de phonologie, comme celles du « e » dit muet ou schwa, de la coupe syllabique, des liaisons, de l'harmonie vocalique, de l'emprunt de phonèmes d'autres langues, les différents dialectes, idiolectes ou sociolectes.

La phonétisation n'est pas dépourvue d'ambiguïtés : la même chaîne orthographique peut être associée à différentes transcriptions phonétiques, à cause des « homographes hétérophones », mots qui s'orthographient de la même façon mais se prononcent différemment. On en compte en français standard environ 150 homographes hétérophones, la plupart des cas pouvant être désambiguïsés par la catégorie grammaticale. À la suite de la phonétisation, il est utile d'enrichir la chaîne de phonèmes de marques syllabiques. En effet, les noyaux syllabiques forment les unités rythmiques de base pour la prédiction prosodique.

3.5. Prédiction de la prosodie

En plus de la représentation sous forme de phonèmes, le système de synthèse doit calculer la prosodie, c'est-à-dire le rythme de prononciation, les pauses dans l'énoncé, la mélodie, le rythme syllabique, les variations de force des syllabes.

Sans prosodie en effet, la parole devient difficile à comprendre et perd son pouvoir expressif et beaucoup de son pouvoir informatif. Pour cette raison, la prosodie, domaine relativement négligé de la linguistique des textes, s'est de suite imposée aux oreilles des acousticiens, phonéticiens, linguistes et informaticiens s'occupant de synthèse de parole. Dans la linguistique traditionnelle, avant les traitements automatiques, la prosodie était surtout abordée sous l'angle du mètre et de « l'accent », et relativement peu formalisée. C'est aujourd'hui un domaine majeur dans le champ des études sur la parole : il est certain que la synthèse de la parole a motivé et nourri tout un pan de recherche en linguistique, en phonétique, en traitement automatique des langues autour de ces questions.

La prosodie est par excellence le domaine de la nuance dans le discours et le royaume de l'orateur. Pour prédire la prosodie dans un contexte donné, il faudrait comprendre parfaitement la situation de communication et d'énonciation, c'est-à-dire la structure et le sens de l'énoncé, les intentions véhiculées, les positions respectives du locuteur et de l'auditeur, le rapport du locuteur à son discours, avec ce que cela implique d'attitudes ou d'émotions. Ces informations ne sont pas données par le texte, et restent difficile à traiter automatiquement. Les premiers systèmes de prédiction prosodique ont utilisé des règles liant la syntaxe et des considérations sur la phonotactique (rythme syllabique) à la prosodie. Ainsi la modalité, la distribution des groupes prosodiques, leur position dans la phrase, le nombre de syllabes dans les groupes, etc., permettent de proposer une organisation prosodique pour la phrase. Cette prosodie est souvent très systématique et répétitive.

De façon surprenante, la prosodie obtenue sans règles explicites en synthèse par sélection et concaténation est beaucoup plus variée et naturelle. Dans cette approche, on recherche dans le corpus sonore des segments aussi longs que possible (plusieurs phonèmes, voire syllabes), qui satisfont un certain nombre de critères phonétiques et prosodiques. Par exemple, on recherche les segments en début de phrase, avec un nombre de syllabe donnée proche d'un signe de ponctuation particulier. Ce type de critère suffit pour qu'en recollant directement les segments obtenus, la prosodie semble naturelle. En revanche, il n'y a pas possibilité de la contrôler et, par exemple, de spécifier un type d'expression particulier. En synthèse paramétrique statistique, tout comme en synthèse par règles, la prosodie est un paramètre explicite de la synthèse, et donc parfaitement contrôlé. Mais ici, les courbes prosodiques sont apprises automatiquement sur le corpus, par des modèles markoviens. Les règles ne sont pas élaborées par une analyse explicite d'exemples, mais déduites automatiquement du corpus. Comme les variations prosodiques sont paramétrées, il est possible de spécifier des variations particulières si on souhaite réaliser un effet prosodique particulier.

3.6. Parole audiovisuelle

En plus de la génération du son, les systèmes de synthèse peuvent être associés à des visages parlants : c'est la synthèse de parole audio-visuelle. Ces représentations plus ou moins stylisées de sujets parlants posent des questions nouvelles sur la coordination entre geste de parole visible et parole audible. La synthèse audio visuelle consiste le plus souvent à appliquer sur un visage parlant les consignes issues de la synthèse du son. Aux phonèmes qui représentent les plus petites unités sonores sont associés des « visèmes », ou plus petites unités visuelles, qui rendent compte des mouvements visibles associés à la parole. La correspondance entre phonèmes et visèmes est relativement directe, par un jeu de règles. Les conséquences visuelles de la prosodie ont fait l'objet de moins d'études à ce jour, mais ce domaine est en plein essor puisque l'intérêt actuel pour la parole expressive concerne tout autant les aspects sonores que les aspects visuels. En effet, un froncement de sourcils est susceptible de modifier la perception de la prosodie et de l'expression globale.

4. DISCUSSION

4.1. Machine parlante ?

La possibilité d'une machine parlante n'est pas une question nouvelle. Des dispositifs pour reproduire certains aspects de la voix humaine ont été proposés dès la première moitié du XVIIe siècle (dans l'Harmonie Universelle de Marin Mersenne), anticipant de quelques décennies les travaux linguistiques de Port-Royal. Ces première machines, en fait des instruments imitant l'appareil vocal humain à l'aide de technologies dérivées de la facture d'orgue, ont été accompagnées d'une réflexion philosophique sur la parole, le langage, et les automates. Descartes envisageait tout-à-fait qu'une machine puisse imiter de la parole, sans la comprendre plus que ne le font les oiseaux parleurs. Les moyens technologiques et les connaissances scientifiques contemporains nous placent-ils dans une situation épistémologique radicalement différente ? Ce n'est pas certain, puisque les systèmes de synthèse à partir du texte sont toujours des machines à imiter la parole, capables de la modéliser ou de la prédire, mais sans la « comprendre » et bien entendu sans la liberté ou la volonté de la générer.

La ligne de démarcation entre véritable parole et reproduction mécanique ne peut être franchie sans toucher à la question même du sujet parlant. Cependant le développement de machines parlantes plus sophistiquées est possible, en améliorant l'illusion d'une réponse intelligente, c'est-à-dire adaptée, expressive, voire émotive, modulée en fonction de la situation de communication. De nombreuses recherches se développent autour des questions de l'adaptation au contexte, de la compréhension des textes et de l'expression.

4.2. Synthèse de l'expression

Une culture dominée par l'écrit incline à considérer la parole principalement comme un moyen de transmettre un sens par la fonction linguistique, mais il faut se garder de la confiner à ce rôle. La parole est premièrement, dans l'ontogénèse comme dans la phylogénèse, un moyen de communication et d'expression. Dans l'usage quotidien de la parole, beaucoup d'actes communicatifs ont un contenu linguistique très réduit, mais un contenu expressif très important. La synthèse de parole s'est au début confinée au domaine de la « signification », de nature logico-grammaticale, en s'appuyant sur le signe, le langage étant considéré comme moyen de représentation, ancré dans le cognitivisme, pour lequel les aspects lexicaux et syntaxiques prédominent. Plus récemment, la synthèse des attitudes et des expressions a abordé le domaine du « sens », qui relève de la rhétorique et de l'herméneutique, et prend pour objet les textes écrits et/ou oraux. Le langage est alors considéré comme moyen de communication, donc par nature situé, incarné et expressif, avec une prédominance des aspects prosodiques. Bien entendu certaines tâche dévolues à la synthèse à partir du texte portent surtout sur la signification : s'il s'agit de lire un

journal, on attend du système un haut niveau d'intelligibilité, et peu de nuances expressives. Au contraire, s'il s'agit d'utiliser la synthèse pour un jeu vidéo, ou pour l'accueil du public, le sens de l'interaction importe autant que la signification, la parole doit avant tout être expressive, le contenu sémantique étant plutôt limité et fortement prévisible. A la limite, la synthèse à partir du texte peut être utilisée pour synthétiser de la voix chantée, pas forcément très intelligible, mais très expressive.

4.3. Synthèse performative

Depuis quelques années se développe une approche de contrôle gestuel de la synthèse (« performative speech synthesis »). Le synthétiseur est considéré comme un instrument (évolué, puisqu'il inclut des procédures linguistiques sophistiquées, comme la transcription graphème-phonème) dans laquelle les aspects expressifs sont pris en charge par un opérateur, qui joue de/avec la parole en temps réel. Cette approche ouvre de nouvelles pistes de recherche sur l'expression et la communication parlée, mais évidemment nécessite une intervention humaine et ne convient pas à toutes les applications de la synthèse à partir du texte.

Christophe d'ALESSANDRO
LIMSI-CNRS (Orsay)

BIBLIOGRAPHIE SUCCINCTE

D'ALESSANDRO C., RICHARD G. (2013), « Synthèse de la parole à partir du texte », *Techniques de l'ingénieur*, Réf. H7288. [Un article de présentation de la synthèse en direction de l'entreprise et de l'ingénieur.]

ALLEN J., HUNNICUTT M.S., KLATT D. (1987), *From text to speech : the MITalk system*, Cambridge Univ. Press. [Une description du paradigme de synthèse par règles et des analyses linguistiques, pour l'anglais.]

DUTOIT T. (1997), *An Introduction to Text-To-Speech Synthesis*. Kluwer Academic Publishers. [Une description du paradigme de synthèse par concaténation et des analyses linguistiques, pour le français.]

SÉRIS J.-P. (1995), *Langages et machines à l'âge classique*. Hachette.

MARTINET A. (1960), *Éléments de linguistique générale*, Paris, Armand Colin. [Contient, entre autres, des discussions sur les machines parlantes.]

D'ALESSANDRO C. ET TZOUKERMANN E. (2001), *Synthèse de la parole à partir du texte*, Hermès Science Publications. [Ce livre présente une collection d'articles, dont une revue de la synthèse en français depuis les origines jusqu'en 2001 et d'un disque d'exemples sonores, contenant tout les systèmes en français de l'origine à 2001, accessible à : http://groupeaa.limsi.fr/corpus:synthese:start]

RASTIER F. (1989), *Sens et textualité*, Hachette. [Une discussion, entre autres, sur le sens et la signification.]

TAYLOR P. (2009), *Text-to-Speech Synthesis*, Cambridge University Press. [Ce livre présente les techniques de synthèse modernes, dont le paradigme paramétrique statistique.]

Les développements de la synthèse à partir du texte font régulièrement l'objet d'articles dans les revues comme :

IEEE Transactions on Speech and Audio Processing
Speech Communication, Computer Speech and Langage
EURASIP Journal on Audio, Speech, and Music Processing
Journal of the Acoustical Society of America,
Acta Acustica United with Acustica,
Language Ressources and Evaluation.

Et dans les actes de conférences comme :

IEEE International Conference on Audio, Speech and Signal Processing (ICASSP),
Annual Conference of the International Speech Communication Association (INTERSPEECH),
International Conference on Phonetic Sciences (ICPhS)
Speech Synthesis Workshop de l'International Speech Communication Association

Quelque sites de compagnies proposant des produits de synthèse à partir du texte

Acapela
http://www.acapela-group.com/
Voxygen
http://voxygen.fr/
Ivona
http://www.ivona.com/en/
Nuance
http://www.nuance.fr
creawave
http://crea-wave.com/
ATT
http://www.research.att.com/projects/Natural_Voices/index.html
Dec Talk
http://www.speechfxinc.com /
IBM
http://www-01.ibm.com/software/pervasive/tech/demos/tts.shtml
Yamaha — Vocaloid
www.vocaloid.com/en/