LES THÉORIES DE LA PAROLE DANS L'ÉPROUVETTE DE LA SYNTHÈSE

La synthèse de la parole ne se résume pas à l'inversion des résultats de l'analyse phonologique et phonétique de la parole. Les préoccupations traditionnelles de la phonologie sont orientées vers les fonctions communicative et distinctive, ainsi que vers l'établissement de structures et universels. Parallèlement, les préoccupations traditionnelles des sciences phonétiques concernent les mécanismes de production de la parole, ainsi que les concepts physiques et psychologiques permettant de comprendre la création, la transmission, la perception et la compréhension de la parole. Ces informations ont une importance indéniable pour la synthèse de la parole. Cependant, des informations supplémentaires sont requises pour recréer une parole naturelle, caractérisée par la pleine panoplie de marques typiques d'un individu particulier, fonctionnant dans une communauté linguistique spécifique. Un survol rapide des tentatives de créer une synthèse de la parole durant le dernier demi-siècle nous révèle que les réussites, tout comme les défaillances, des synthèses existantes sont directement liées aux préoccupations traditionnelles en analyse linguistique et phonétique de la parole. Certaines perspectives prometteuses pour l'exploration des connaissances manquantessont discutées, ainsi que les conséquences plus générales de ces considérations pour l'épistémologie de notre domaine.

1. Introduction

Avec le lancement du Laboratoire d'analyse informatique de la parole (LAIP) à l'Université de Lausanne en 1991, nous nous sommes lancés le défi de développer un système complet de synthèse de la parole pour la langue française. Par ce projet, nous souhaitions vérifier si certaines des théories psycholinguistiques et phonétiques[1] portant sur la

[1] Terminologie: La *psycholinguistique* s'occupe de la modélisation des processus «mentaux», se produisant en temps réel durant tout traitement linguistique. Les aspects de la *prosodie* qui nous intéressent dans le contexte de ce volume concernent, avant tout, les processus psycholinguistiques associés à la production des aspects temporaux et intonatifs d'énoncés oraux. La *phonétique* s'occupe de la modélisation des aspects articulatoires, acoustiques et perceptives associés spécifiquement au langage oral, donc de la parole. L'élément fédérateur entre ces trois domaines est leur partage d'un *substrat neurolinguistique*: les processus psycholinguistiques, tout comme les processus phonétiques articulatoires et perceptifs, se déroulent nécessairement dans le substrat physiologique des systèmes neuro-musculaires humains. Une théorie englobante de la *production de la parole* tiendra donc compte de toutes évidences (expérimentales ou naturelles) portant sur le traitement en temps réel du matériel linguistique, aboutissant à la production d'une parole naturelle. Dans cette perspective, une *théorie psycholinguistique* alimentant une *synthèse de la parole* tentera, dans la mesure du possible, de faire correspondre les algorithmes alimentant un logiciel informatique aux processus «mentaux» se déroulant en temps réel dans le locuteur humain.

production de parole pouvaient être vérifiées dans le cadre d'une modélisation informatique. Précédemment[2] nous avions soupçonné qu'un ensemble de règles à la fois plus restreint que plus directement inspiré par des concepts psycholinguistiques, serait en mesure de produire une parole bien plus naturelle que la synthèse généralement disponible durant les années 1970-90.

L'implantation de ce programme de recherches a connu des réussites tout autant que certains échecs. Par exemple, un système de regroupements de mots inspirés par des recherches psycholinguistiques nous a permis de construire une charpente temporelle pour la prédiction des durées sonores plus proche des habitudes humaines en lecture à haute voix que les modélisations implantées dans la majorité des systèmes de synthèse précédents du français (voir l'article par B. Zellner dans ce volume). Par contre, la modélisation des mouvements articulatoires, ainsi que des ondes acoustiques qui en résultent, s'est avérée trop difficile, du moins pour l'instant, dans le cadre d'un système de synthèse fonctionnant en temps réel.

Cela étant, je ne souhaite pas évoquer les résultats de nos travaux dans ces quelques pages. Plutôt, j'aimerais explorer les bases théoriques de la modélisation algorithmique de la production de la parole. En effet, l'exercice auquel je souhaite me livrer ici semblera peut-être surprenant au premier abord: j'aimerais montrer que notre compréhension des phénomènes phonétiques et psycholinguistiques sous-jacents à la parole reste sérieusement déficiente sans modélisation par une synthèse. Selon nous, la *reconcrétisation d'un phénomène à partir des principes structuraux et algorithmiques* — donc ici, la (re-)synthèse de la parole — est indispensable à la bonne compréhension de la production de la parole humaine. Plus important encore, je soutiendrais qu'une compréhension exclusivement basée sur l'analyse — sans tentative de synthèse parallèle — induirait des visions de limitatives, c.-à-d., des visions structurales figées qui peuvent finir par retarder l'approfondissement de notre compréhension de la production de la parole humaine.

J'aimerais donc présenter des arguments en faveur d'un outil de recherche supplémentaire qui s'impose, selon nous, obligatoirement dans les sciences humaines. La méthode traditionnelle et dominante pour amorcer l'appréciation des phénomènes de l'expression humaine consiste à effectuer une identification expérimentale des éléments centraux, puis à élaborer une structure reliant ces éléments de manière logique et raisonnée. Mais selon nous, cette méthode n'est que le premier pas dans une démarche à plus longue haleine, car elle devrait être complétée par une modélisation dynamique. Dans cette vision de recherche, la

[2] Dans le cadre de ma thèse puis en tant que professeur-chercheur à l'Université du Québec à Montréal.

modélisation est considéré comme un outil complémentaire à *l'analyse*, en ce sens qu'elle vérifie la justification ainsi que la qualité de l'analyse originale.

Appliquons ces concepts à la parole. L'identification de la structure centrale de la production de la parole est ancrée dans une longue tradition d'analyses acoustiques, articulatoires et perceptives. Ainsi, les grandes lignes des mécanismes de l'articulation de la parole sont bien établies, tout comme les événements neurologiques sous-jacents, ou la génération des sons de la parole, ainsi que leur perception distinctive. Ces analyses ont été améliorées et raffinées ces dernières décennies par l'introduction de méthodes statistiques, en particulier les analyses discriminante et factorielle. Nous pouvons aujourd'hui cerner la structure centrale d'un phénomène phonétique donné, et jusqu'à un certain point, du comportement humain sous-jacent à la production de la parole.

Cependant, ce n'est que par la synthèse que nous pouvons vérifier si notre compréhension de cette structure est juste et adéquate. A quel moment devient-elle «juste et adéquate»? La réponse est évidente: quand elle suffit à une reconstruction complète et acceptable, voire indiscernable, du phénomène d'origine. Pour utiliser une figure architecturale, si l'analyse traditionnelle nous permet — la plupart du temps — de cerner la charpente d'une maison, ce n'est qu'en reconstruisant la maison que nous pouvons comprendre l'importance des détails laissés de côté dans la considération initiale des éléments centraux de la structure. Ce n'est que durant la reconstruction d'une maison ou d'une église du moyen-âge, par exemple, que nous comprenons pleinement les limites des matériaux et des outils utilisés à l'époque. Et ce n'est que par un processus de reconstruction que nous serons amenés à comprendre en détail les astuces et techniques élaborées à l'époque pour palier à de telles déficiences.

Cette ligne d'argumentation nous permet de reformuler les déficiences des divers systèmes de synthèse de parole contemporains en termes particulièrement utiles, car elle pose la question de l'outil épistémologique lui-même. Si les systèmes de synthèse de la parole avaient l'air de robots parlants dans le passé (et parfois aujourd'hui encore), est-ce peut-être parce que nous nous sommes permis de croire que «la charpente faisait la maison»? En particulier, est-ce que le «moindre dénominateur commun» obtenu par des méthodes analytiques nous a dérouté, en nous induisant à penser pendant deux décénnies ou plus, que pour recréer la parole, il suffisait d'inverser les principes identifiés au moyen de nos analyses? A la lumière des recherches effectuées au cours de la dernière décennie, nous sommes disposés à admettre que cela soit le cas. En effet, une parole basée exclusivement

sur les aspects partagés entre locuteurs d'une langue donnée, semble plutôt pauvre en texture. La question de la validité de l'inversion des résultats analytiques est donc tout à fait motivée.

Si les recherches sur la synthèse sont aujourd'hui largement sorties de cette conception limitative de la science (voir p.ex., l'article de J. Local dans ce volume), ce n'est pas encore le cas de la généralité des sciences humaines. Voici donc raison suffisante pour illustrer la relation entre l'analyse et la synthèse par rapport à la synthèse de la parole. Nous illustrerons nos propos par un exemple tiré d'un aspect particulièrement bien connu de la parole, c.-à-d., la reconstruction de l'onde acoustique au moyen d'une «synthèse par formants». Nous verrons que cette approche envers la synthèse ne circonscrit pas seulement le problème de l'inversion, mais contient en elle les germes de solutions au problème.

2. Analyse et resynthèse acoustique de la parole

a. Les fondements. Il est bien établi que la structure acoustique de la parole est fortement déterminée par l'information vocalique et que les voyelles, à leur tour, sont largement caractérisées par des «formants». Ce phénomène est facilement mis en évidence par l'analyse spectro-temporelle traditionnelle des signaux de la parole. Il s'agit d'une décomposition de l'onde acoustique en ondes sinusoïdales de différentes fréquences au moyen d'une analyse de Fourier, ainsi que par une estimation de l'amplitude relative de chaque type d'onde. A partir d'une telle analyse, nous obtenons les «spectrogrammes» bien connus de la phonétique acoustique (Figure 1).

Dans une telle représentation tri-dimensionnelle, où l'axe *x* correspond au temps, l'axe *y* à la fréquence et le degré de noirceur à l'amplitude relative, on perçoit les structures dominantes de la parole et de la voix humaine, c.-à-d. les *formants* (les «tâches épaisses» du milieu de la Figure 1) et les *fréquences harmoniques* (les «tâches minces» du bas de la Figure 1). Si un spectre est appliqué de manière à regrouper les résultats d'un grand nombre d'ondes (p.ex., sur une largeur de bande de 125 Hz), on perçoit des bandes appelées des «formants», avant tout durant les périodes de temps correspondant aux voyelles. Il s'agit de renforcements et de diminutions relatives dans les résonances de l'onde acoustique, dont les fréquences précises sont déterminées par les degrés d'ouverture du conduit vocal le long de son étendue, ainsi que par diverses réflexions intra- et extra-orales.

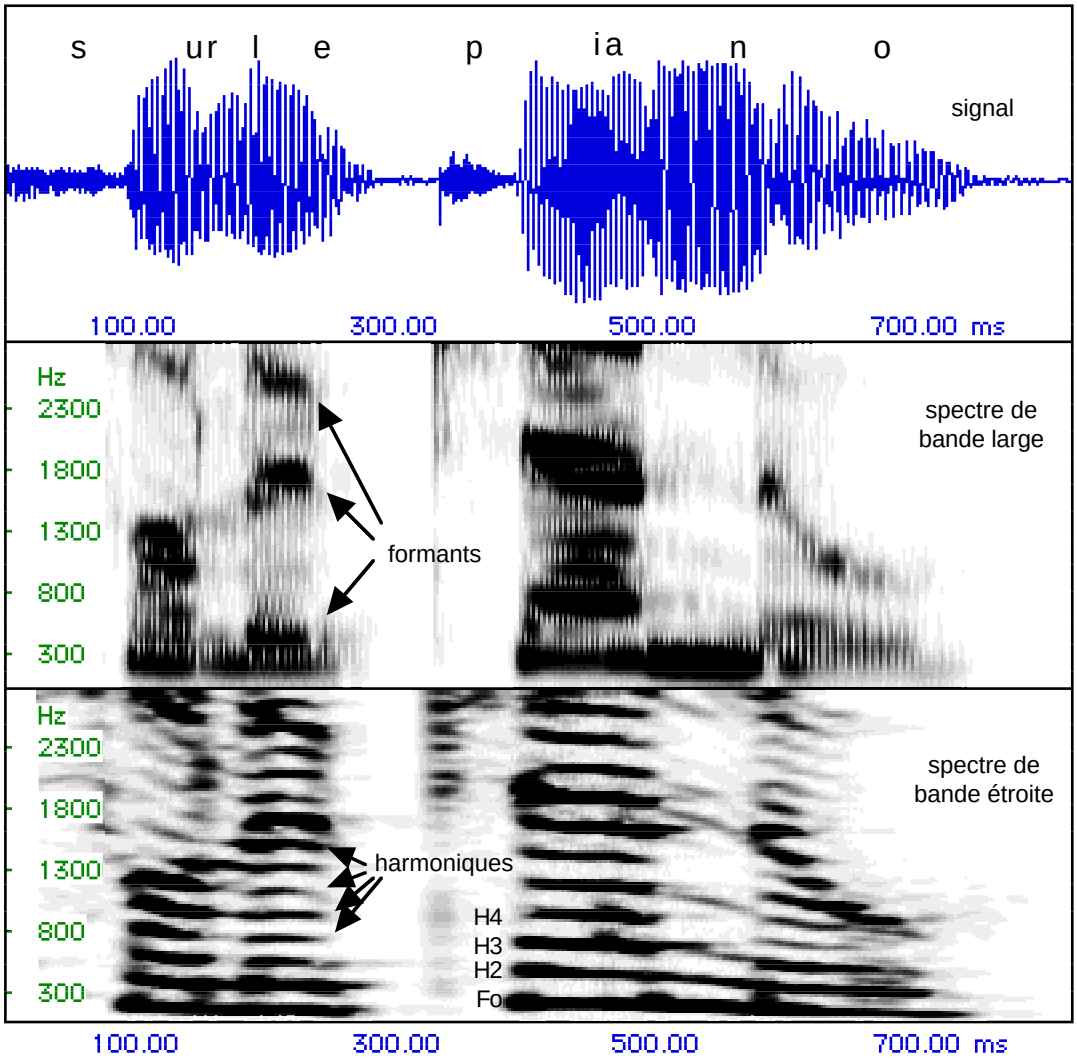

Figure 1. Différentes analyses de l'onde sonore correspondant à l'énoncé «sur le piano». Haut: le signal acoustique montrant l'évolution générale de l'amplitude sur le temps. Milieu: spectre de bande large (125 Hz) qui fait ressortir les «formants». Bas: spectre de bande étroite (25 Hz) montrant les fréquences harmoniques de la voix. Cet énoncé intervient en fin de la phrase «Pierre avait posé son porte-feuille sur le piano» et est donc caractérisé par une mélodie (Fo) tombante.

L'onde acoustique de la parole elle-même prend son origine dans l'interaction entre l'activité des cordes vocales et celle de la cage thoracique. En effet, le bas de la Figure 1 illustre que les véhicules acoustiques des «formants» ne sont en fait rien d'autre que les fréquences harmoniques de la voix. Il s'agit d'une part de la fréquence fondamentale (Fo), qui résulte en grande mesure de l'activité régulière de la glotte, et d'autre part, des harmoniques supérieures de cette même fréquence (H2, H3, etc.). Les fréquences harmoniques sont mises en évidence par une analyse de Fourier plus détaillée du même matériel sonore, c.-à-d., par un regroupement d'un plus petit groupe d'ondes. (Dans la Figure 1, on a regroupé les ondes sur une largeur de bande de

25 Hz). Retenons que la hauteur des fréquences harmoniques est largement responsable de la hauteur mélodique perçue.

Ceci illustre le la nature multidimensionnelle de la structure acoustique des voyelles. Les ondes harmoniques, porteuses de la mélodie, sont en même temps codées dans leurs amplitudes formantiques. Au moyen du «microscope» plus détaillé de l'analyse à bande étroite, nous comprenons que les «formants» ne sont rien d'autre qu'un renforcement de certaines fréquences harmoniques sur certaines périodes de temps.

b. Premières resynthèses formantiques. Dès lors, une stratégie de resynthèse de la parole se suggère. Serait-il possible de resynthétiser la parole à partir des fréquences harmoniques, ainsi qu'à partir des renforcements et des diminutions formantiques? Il suffirait de recréer un ensemble de fréquences harmoniques évoluant sur le temps, reproduisant les aspects spectraux de l'onde originale tant par rapport à leur hauteur de fréquence que par rapport à leur amplitude relative.

Une telle chose est effectivement possible et elle est pratiquée depuis un demi-siècle environ. La structure harmonique de la voix ainsi que l'existance des formants étaient déjà connus dans les années 1930 (Potter, 1930; Steinberg, 1934; cités dans Koenig et al., 1946). Mais ce n'est que suite à l'invention du spectrographe durant la deuxième guerre mondiale (Koenig et al., 1946) que la relation essentielle entre les fréquence formantiques et les voyelles sous-jacentes a été mise en évidence. Cette relation a été fortement popularisée sous la forme du triangle (ou rectangle, ou trapèze) vocalique (p.ex., Joos, 1948; Peterson & Barney, 1952; Stevens & House, 1955, Fant, 1956)[3].

Durant les années d'après-guerre, ces connaissances ont été appliquées à la synthèse de la parole, particulièrement dans le cadre de la reconstruction de syllabes alimentant une série d'études perceptives. Dans ces expériences, des «formants» stylisés ont été peints à la main sur des transparents et ont été reconvertis en onde acoustique en passant les transparents sous un senseur optique relié à un générateur de sons[4] (p.ex. Figure 2). Les signaux sonores résultant de cette

[3] De manière générale, les voyelles dites «articulatoirement hautes» montrent un formant 1 relativement plus bas que les voyelles dits «articulatoirement basses». Parallèlement les voyelles dites «articulatoirement antérieures» montrent un formant 2 relativement plus haut que les voyelles dites «articulatoirement postérieures». Cependant, ces tendances sont assujeties à des modifications contextuelles importantes. De plus, l'identification de formants n'est pas toujours immédiate. Par exemple, la représentation acoustique de l'enregistrement montré dans la Figure 1 a été obtenue selon les meilleures méthodes acoustiques et informatiques en utilisation courante. Cependant, l'identification des formants durant la production de la diphthongue /ja/ de «piano» n'est pas tout à fait évidente.

[4] La méthode était connue sous le nom de "pattern playback". La méthode de reproduction sonore est résumée comme suit dans Liberman et al., 1956, p. 129 (traduction libre): «Le "playback" utilise une

procédure n'étaient pas de qualité extraordinaire. Toutefois, ils ressemblaient suffisamment aux voyelles humaines pour permettre des jugements systématiques concernant leur appartenance à un groupe ou un autre de voyelles. Et plus important dans le contexte actuel, ces expériences ont servi à renforcer la notion que la synthèse n'était essentiellement qu'une analyse inversée.

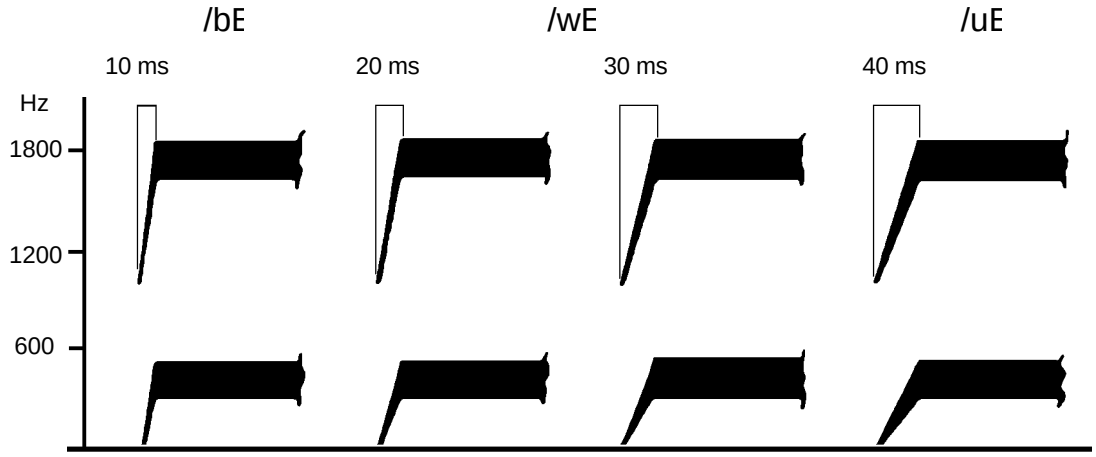

Figure 2. Patterns formantiques peints à la main, resynthétisés en syllables audibles au moyen de la méthode de "pattern playback", et utilisés dans le cadre d'expériences psychoacoustiques dans les années 1950 et 1960. Dans l'expérience illustrée ici, une manipulation de la durée de montée initiale des formants 1 et 2 produit des syllabes proche de /bE/ quand la montée est brève, et des syllabes plus proche de /uE/ quand la montée est prolongée (illustration d'après Liberman et al., 1956, p. 130). Les manipulations de ce type ont permis déceler les structures formantiques sous-jacentes aux interactions entre voyelles et consonnes contigües, tandis que d'autres manipulations ont permis d'identifier les structures formantiques sous-jacentes aux différentes voyelles.

c. La synthèse de Klatt. Avec l'arrivée des méthodes informatiques dans les années '70 et '80, une synthèse informatisée basée sur la même idée fondamentale fut développée, notamment en Suède par le groupe de Fant et aux Etats-Unis, par le groupe de Klatt et collègues au MIT (Klatt, 1980; 1997; Klatt & Klatt, 1990; pour une revue de cette approche, voir Styger & Keller, 1994). Il devenait dorénavant possible de tester l'inversion des résultats de l'analyse acoustique par rapport à des énoncés entiers, voire par rapport à la production de la parole continue. De plus, un raffinement important était implanté en ce sens que l'outil informatique permettait de définir une évolution dynamique

---

roue produisant un ton d'intensité variable au moyen d'une modulation de la lumière d'un arc de mercure. Ceci produit une fréquence fondamentale de 120 Hz ainsi que toutes les fréquences harmoniques, jusqu'à la 50e harmonique, à 6000 Hz. Les raies de lumière sont projetées sur le spectrogramme à l'échelle appropriée. Quand le spectrogramme, peint à la main, est passé devant la lumière, la peinture blanche reflète les raies. Ces raies sont captées et acheminées à un amplificateur qui traduit les fréquences reflétées en son audible. »

nettement plus importante des formants ainsi que des fréquences fondamentale et harmoniques. Enfin, la qualité sonore des consonnes était approximée par l'introduction de différents bruits de plosion et de friction dans le signal.

En employant ce type de recomposition du signal à partir d'éléments primaires, un système complet de traduction texte-à-parole (angl. "Text-to-speech", ou "TTS"), fut développé par Dennis Klatt et ses collègues. Afin de connaître les valeurs précises des formants dans l'ensemble des voyelles, adapés aux différents contextes consonantiques, Klatt a utilisé une base d'informations étendue sur sa propre prononciation de l'anglais. Outre les valeurs formantiques, cette base de données incluait des règles de durée pour chaque type de phonème (toujours adaptées au contexte) ainsi que quelques règles élémentaires concernant l'évolution de la courbe de fréquence fondamentale. La synthèse basée sur cette approche a été fortement répandue, d'abord par la publication de son code source (Klatt, 1980), puis par l'implantation du système "DecTalk", produit en collaboration avec la compagnie DIGITAL, et largement distribué depuis les premières années 1980.

La synthèse par formants représentait une avance décisive dans le domaine de la synthèse, mais elle accusait néanmoins quelques défauts importants — des défauts qui ont inspiré une panoplie de recherches dans la quinzaine d'années suivante[5]. D'une part, la création de nouvelles voix et le développement d'un jeu de règles pour d'autres langues que l'anglais constitue une lourde entreprise, du fait du nombre impressionnant de paramètres à contrôler et à faire varier en continu (p.ex., un nouvel ensemble de 40-80 valeurs tous les 5-10 millisecondes, où chaque valeur requiert son propre modèle[6]).

D'autre part, la qualité sonore de cette synthèse était toujours déficiente — elle était plutôt «robotique» et son timbre de voix était peu naturel. Ces deux limites découlent naturellement de la méthode de

---

[5] Note personnelle: Quand j'ai rendu visite au laboratoire de la parole du MIT en 1982, Dennis Klatt m'a fait écouter des exemples de phrases qu'il avait synthétisées. Tout en le félicitant de la qualité sonore qu'il avait atteint au moyen de son système, j'attirais son attention sur les défauts du système (notamment la qualité sonore des /r/ et des /l/). Klatt était visiblement déçu de ma réaction, à l'époque ainsi que six ans plus tard quand je l'ai rencontré une dernière fois quelques mois avant sa mort. Ce n'est que récemment, quand nous avons tenté nous-mêmes de créer de la parole continue en utilisant ces mêmes outils plutôt élémentaires, que je me suis rendu compte des véritables merveilles accomplis à l'époque par Dennis Klatt — ce qui m'a amené à regretter le ton critique que j'avais assumé envers D. Klatt. En discutant les défauts de cette approche, il ne faudrait donc pas oublier les accomplissements considérables de ces chercheurs.

[6] En effet, les collègues et les successeurs de Klatt ont pu considérablement réduire le nombre de paramètres à contrôler (Stevens, 1995). Au lieu des 80 paramètres définis par le modèle évolué de la fin des années '80, le modèle de Stevens n'en requiert qu'une petite dizaine. Cependant, les autres problèmes discutés ici affectent toute la classe des synthèses paramétriques, incluant les dernières versions de la synthèse par formants selon Klatt.

création de cette voix synthéthique. Afin de mieux comprendre ces déficiences, il faudra brièvement approfondir la conceptualisation de cette méthode de synthèse.

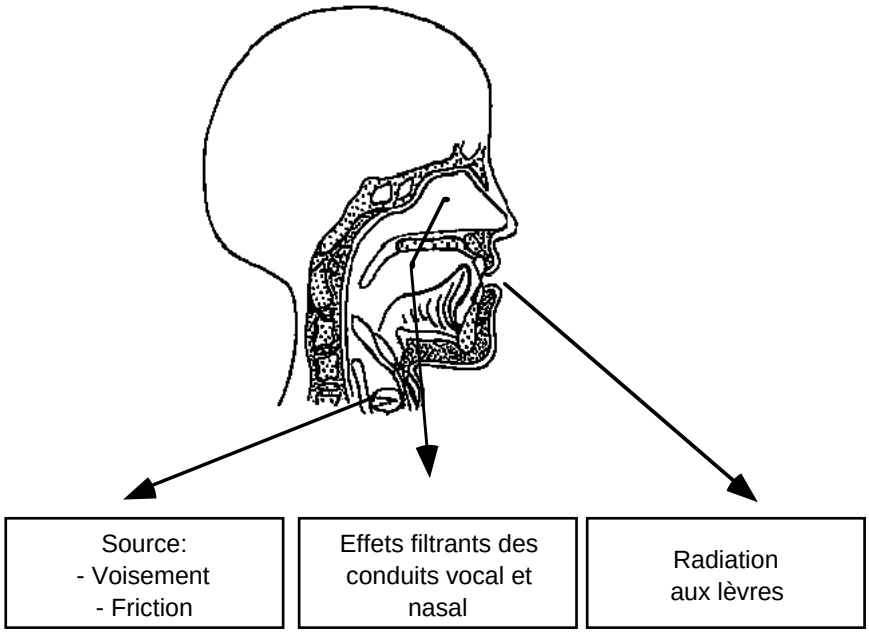

Figure 3. Conceptualisation fondamentale du modèle source-filtre. Le modèle est divisé en trois parties, la source (le voisement, la friction), le filtre (simulation des effets filtrants des conduits oral et nasal), et la radiation aux lèvres. Illustration adaptée de Styger & Keller, 1994.

d. Le fonctionnement d'une «synthèse par formants». La synthèse de Klatt est basée sur trois composantes principales (Figure 3). D'une part on modélise le signal de source de la voix. Dans le cas le plus simple, celui-ci peut être approximé par un signal sinusoïdal, mais des formulations plus récentes définissent différents modèles de source selon différents modes de production vocale (p.ex., Klatt & Klatt, 1990: glotte plutôt fermée — «voix laryngée», glotte normale — «voix normale», glotte plutôt ouverte — «voix aspirée»). Le choix de la forme du signal de base détermine les caractéristiques spectrales de la Fo et des fréquences harmoniques fournies au départ.

La prochaine composante est la phase la plus cruciale dans le processus de génération du signal. Dans le cas de voyelles et de consonnes voisées, une série de filtres approximant les 4-6 premiers formants est imposée à l'onde de source (Figure 4). Pour les fricatives, un générateur de signaux aléatoires fournit des sources de bruit convenables, à modulér par des filtrages subséquents. Dans le cas de plosives, une période silencieuse est suivie par un bruit de plosion, ainsi que par une transition légèrement affriquée vers le son suivant.

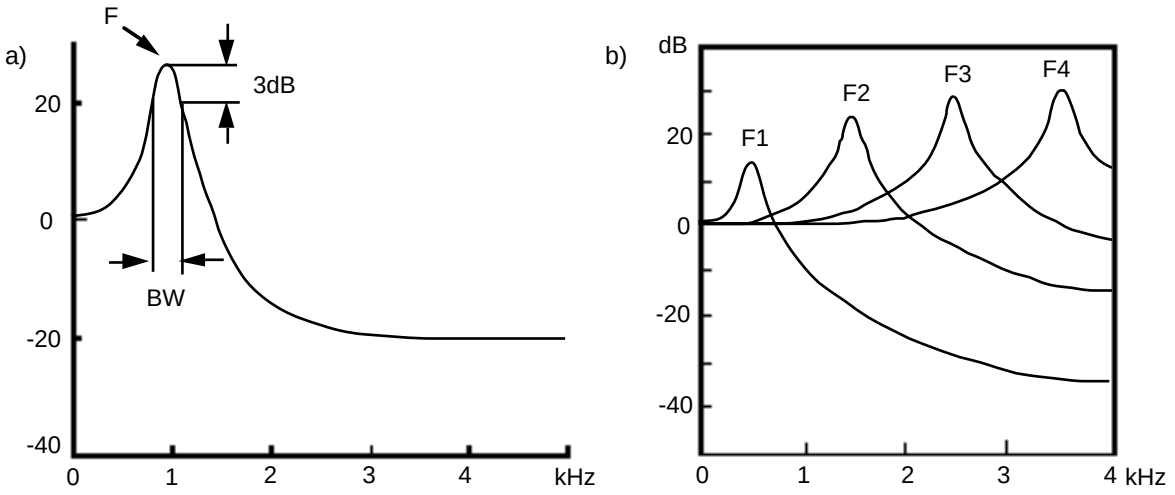

Figure 4. Simulation des formants au moyen du filtrage successif de l'onde de source par une série de filtres. (a) La forme du filtre est déterminé par F, sa fréquence centrale et par BW ("Bandwidth") sa largeur de bande entre les deux points de diminution par 3 db du maximum. (b) Les effets de superposition de plusieurs filtres de ce type s'approchent du spectre naturel de la voix en production de parole.

Retenons que ce type de simulation reste approximative dans deux sens. Premièrement, le spectre d'un tel signal ne ressemble que partiellement au spectre d'une voix naturelle. La formulation du paragraphe précédent laisse déjà sous-entendre que la simulation des consonnes, par exemple, était fort approximative. Par rapport aux voyelles, nos mesures indiquent que les formants synthétiques de cette approche sont généralement plus «plates» (c.-à-d., montrent moins d'amplitude) que les formants naturels (Figure 5). Malgré l'excellence de cette solution initiale, ni les concepteurs de cette approche, ni les auditeurs, n'éprouvaient aucun doute sur la qualité "robotique" de cette voix synthétique.

Deuxièmement, une telle représentation idéalisante laisse nécessairement de côté beaucoup de détails pertinents. Par exemple, l'examen du spectrogramme de la Figure 1 montre non seulement les formants attendus, mais également des résonances supplémentaires, difficiles à classer. Nous savons que certaines de ces résonances contribuent au timbre individuel de la personne. La synthèse par formants traditionnelle ne génère pas ce type de résonances supplémentaires.

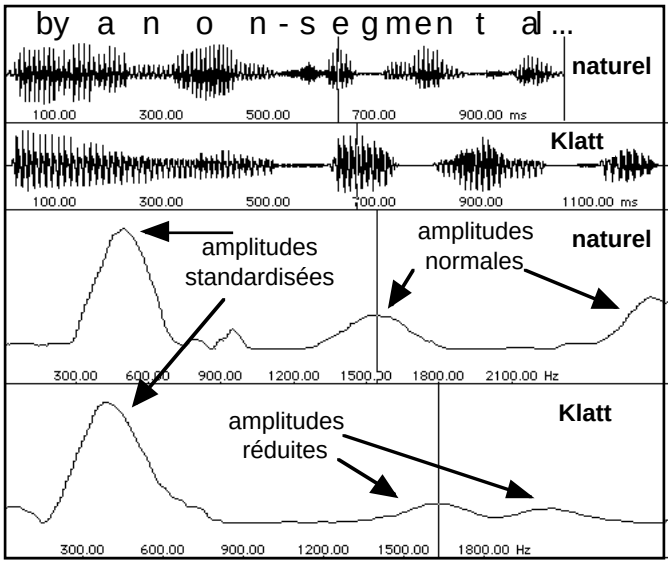

Figure 5. Comparaison de spectres pris sur la voyelle /E/ de "non-segmental synthesis system". Haut: locuteur humain. Bas: synthèse par formants, méthode de Klatt adaptée à l'Université de York, UK. Les pics des formants 2 et 3 de la synthèse sont fortement réduits, en comparaison avec les pics du spectre naturel. La représentation du Formant 1 a été standardisée à la même hauteur par le logiciel d'affichage graphique.

De plus, une voix humaine produit toute une série de «petits bruits» parasites et supplémentaires (p.ex., bruits de respiration, claquements de langue etc.). Ces bruits contribuent encore davantage à la qualité naturelle de la parole, mais la presque totalité des synthèses actuelles les excluent[7]. Le signal de ce type de synthèse était donc, jusqu'à un certain point, trop «idéal», puisqu'il négligait les composantes individuelles de la parole[8]. L'inversion d'un modèle analytique avait donc montré ses limites: pour reconstruire la parole, il ne suffisait pas d'inverser les abstractions spectro-temporelles décelées au moyen de méthodes analytiques.

3. A la recherche du naturel de la parole

Comment pallier à ce problème? Comment peut-on compléter l'information manquante du signal, y-compris sa qualité individuelle, afin de le rendre plus naturel, plus agréable, plus acceptable?

a. Une solution transitoire. Ce problème est plus difficile à résoudre qu'on ne se l'imagine au premier abord. Pour l'instant, la communauté de la synthèse de la parole a généralement choisi de contourner le

[7] Exception notable: la synthèse concaténative de ATR, Nara, Japon, développé sous la direction de N. Campbell, qui inclut les pauses de respiration.

[8] D'autres limites de la synthèse de Klatt concernaient avant tout l'aspect prosodique. Les modèles de l'époque étaient nécessairement moins sophistiqués que les modèles actuels.

problème en adoptant la voie concaténative[9]. Selon cette méthode, le signal de parole est construit à partir de différents segments de signal soigneusement extraits d'enregistrements et concaténés ensemble. Dans la plupart de ces systèmes, les segments sélectionnés englobent une transition entre deux sons, il s'agit donc de systèmes dites «diphoniques». Certains autres systèmes emploient des groupes de deux ou plus de sons, ce sont donc des systèmes «polyphoniques». Au moyen de tels systèmes, une qualité de parole plutôt satisfaisante peut être obtenue.

Le problème de l'absence de qualité naturelle est donc résolu, pour l'instant, au moyen du préstockage de segments de parole soigneusement sélectionnés et enregistrés par des locuteurs particulièrement doués pour la tâche astreignante de lecture de corpus contenant toutes les transitions de la langue en question. Cependant les limites de cette approche sont rapidement identifiés: même en appliquant la panoplie des outils de segmentation automatique moderne, la constitution de chaque nouvelle voix requiert plusieurs mois de travail, ainsi que de longues heures d'enregistrement sonore. Afin d'obtenir une qualité satisfaisante pour chaque transition, les signaux sonores ainsi que les analyses extraites de ces signaux doivent être vérifiés manuellement. De plus, la taille de la base de données diphoniques reste non négligeable. Avant toute compression, une base diphonique pour une langue comme le français mesure aux alentours de 5 Mo. Une base d'environ 8 Mo est requise pour une langue distinguant clairement les voyelles accentuées et non-accentuées, comme l'anglais ou l'allemand.

De plus, cette solution laisse en suspens le problème original de l'inversion des résultats de l'analyse — ce qui est un problème fort intéressant du point de vue de la science phonétique, tout autant que du point de vue de l'histoire des sciences humaines en général. Comment aller au-delà des limites posées par l'inversion du modèle analytique? Le problème n'est pas seulement du domaine académique. Pour celui qui apporte des solutions à cette énigme, la promesse principale est la création d'une parole naturelle et personnalisée, à partir uniquement de modèles mathématiques de la parole, donc sans aucun besoin d'enregistrements sonores supplémentaires. En outre, une base sonore complète pourrait probablement être extraite à partir d'un bref enregistrement d'une voix particulière, et de nouvelles voix pourraient être créées à volonté. Au-delà de la recherche fondamentale, les possibilités pour l'industrie (p.ex., les dessins animés, pour la publicité,

[9] Tout comme nous, d'ailleurs. Après avoir exploré la synthèse par formants pendant une période de trois ans, nous nous sommes tournés — comme presque tout le monde — vers une synthèse concaténative.

ainsi que pour la vente automatisée de produits et d'informations via le téléphone) sont considérables.

b. Aux limites des sciences phonétiques. Il est possible d'argumenter qu'il serait peut-être possible d'améliorer la qualité de la prédiction du signal sonore au moyen d'un *modèle articulatoire*. Dans cette perspective, on utiliserait les connaissances acquises sur la géométrie tridimensionnelle du conduit vocal, celles portant sur le fonctionnement des organes d'articulation, ainsi que les connaissances sur la génération et la modulation d'un son acoustique évoluant dans un tel espace, pour générer les sons de la parole. Ces sons seraient ensuite assujetis à un système d'apprentissage par réseau neuromimétique, afin d'en améliorer la qualité sonore tout en apprenant par induction à mieux gérer la multitude de paramètres qui contrôlent le modèle.

En théorie, cette approche présenterait un grand nombre d'avantages. Mis à part la clarté théorique qu'elle apporte à notre modélisation de l'acte articulatoire, elle promettrait la génération d'une multitude de voix au moyen d'une simple modification d'un ensemble de paramètres articulatoires. De plus, un tel modèle ne consommerait que peu de mémoire, car un modèle articulatoire pourrait en théorie se résumer à un code binaire aussi petit qu'un demi mégaoctet.

Les recherches effectuées dans le cadre d'un projet ESPRIT récemment achevé ("SpeechMaps") nous ont montré que cet objectif était probablement trop ambitieux au point actuel des connaissances scientifiques. La qualité sonore d'une resynthèse réalisée dans cette optique était plutôt décevante. Les raisons de cette insuffisance étaient multiples: manques de connaissance de base (en particulier, précision insuffisante dans la spécification de la forme du conduit vocal pour différents sons de la parole dans différents contextes phonétiques, puisque les bases cinéradiographiques de cette connaissance sont statistiquement faibles et relativement imprécises), limites de la modélisation (modèle trop peu détaillé du conduit vocal), excédent de paramètres à contrôler, et dynamique articulatoire encore relativement peu comprise.

Une autre raison à ces limites, nos propres recherches portant sur la structure temporelle nous ont montré que le regroupement lexical et syllabique avait des effets non négligeables sur la production articulatoire (nous estimons que ces effets sont responsables de 25% à 35% de la variation totale des durées syllabiques), des effets qui ne font pas traditionnellement partie d'un modèle articulatoire. Finalement, le modèle final — en particulier le calcul de l'onde à partir des résonances du conduit vocal — était fort complexe et donc d'une lourdeur computationnelle considérable. Même si les algorithmes eux-mêmes ne

consommaient que peu de code binaire, ils étaient lents du fait des nombreux calculs qu'ils exigeaient.

L'inversion des résultats de l'analyse a donc dû accuser un autre échec embarrassant. Même si les connaissances acquises en explorant cette lignée de recherche sont d'une grande importance ultime pour les sciences et les technologies de la parole, ce n'est pas du côté d'un modèle articulatoire que nous pouvons aujourd'hui espérer trouver les solutions les plus directes aux problème du naturel dans la synthèse de la parole.

c. Vers des solutions. Il me semble que les solutions les plus intéressantes à ce problème se trouvent dans deux autres directions. Premièrement, nos systèmes de synthèse devraient être enrichis par des *informations phonétiques systématiques allant au-delà des distinctions phonologiques*. Rappelons que la recherche phonétique et phonologique de ce siècle a été largement caractérisée par la recherche de distinctivité et de structure. Le souci principal était (et est toujours) d'identifier les éléments humains et universels, permettant la communication efficace au moyen d'un système de signes distinctifs. Actuellement, la majorité des chercheurs ne se préoccupe donc pas trop du problème de l'inversion de l'analyse en appui de la synthèse de la parole.

Or il semble que les expériences en synthèse de la parole décrites ici imposent une telle orientation sans ambiguïté. Dans ce volume, John Local suggère quelques pistes que de telles recherches pourraient prendre. Par exemple, il indique que l'aspiration d'une plosive produite à la fin d'un tour de parole est un marqueur phonétique systématique en anglais britannique. Ou il rapporte des distinctions phonétiques fines mais systèmatiques pour deux éléments linguistiques qui sont traditionnellement transcrits de la même manière (p.ex. "lime" et "I'm" ont la même transcription phonologique /-ajm/). Nous pouvons soupçonner que notre utilisation de la parole est en fait pleine de tels détails phonétiques, tout à fait pertinents pour une synthèse véritablement naturelle.

Deuxièmement, nous devons impérativement avancer le travail sur *les caractéristiques individuelles* en production de parole. Ce n'est qu'en modélisant les caractéristiques de voix spécifiques, appartenant à des locuteurs véritables, que nous pouvons espérer combler les lacunes laissées par une concentration exclusive sur «la charpente» théorique de la parole. Aujourd'hui la recherche sur cette question ne fait pas encore partie intégrante des sciences phonétiques ou psycholinguistiques. Cette lacune devra être comblée rapidement, si la synthèse paramétrique veut espérer passer au-delà du stade d'un modèle appauvri.

Les recherches courantes sur l'identification du locuteur peuvent représenter un point de départ vers cette nouvelle orientation. Il faut insister sur la formulation «point de départ», car traditionnellement, les recherches sur l'identification du locuteur (p.ex., dans une optique de la sécurisation de l'accès à une boîte vocale ou de l'entrée dans un édifice), se sont rabattu sur des méthodes statistiques agissant sur des transformées relativement abstraites des aspects spectraux de la voix (p.ex. Li & O'Shaughnessy, 1997; Lee, 1997). Ceci est avant tout dû au fait qu'en identification du locuteur, il faut éviter d'influencer les résultats par le canal de transmission, c.-à-d. une personne devrait être reconnue ou rejetée malgré les différences spectrales introduites par l'utilisation de différents appareils de téléphone. Ce type de problématique ne touche évidemment pas la synthèse. Il est donc possible de songer à des analyses plus proches du matériel de base — et par conséquent plus transparentes — qui devraient nous permettre l'identification des composantes individuelles d'une voix.

4. Conclusion

Une vision intéressante pour les recherches à accomplir dans notre domaine se dégage de ces considérations. Nous avons vu que la question des aspects de la parole généralement partagés entre locuteurs a permis d'établir «la charpente» de nos connaissances sur la parole. Si ces informations ont été fort utiles au cours des travaux traditionnels en synthèse, cette structure s'est ultimement avérée trop circonscrite pour la recréation fidèle d'une parole naturelle.

En effet, nous avons argumenté que les limites actuelles de la synthèse par formants sont dues à une orientation presqu'exclusive vers des tentatives d'inversion des connaissances linguistiques et phonétiques traditionnelles. Ces connaissances sont généralement axées sur les informations et structures qui sont universelles ou du moins, spécifiques à une communauté linguistique donnée. Nous avons considéré que ces connaissances ne suffisent pas pour une tentative de recréation de synthèse. Dans notre esprit, des études approfondies sur les aspects individuels de la production de la parole sont requises.

Si les résultats ultérieurs finissent par étayer cette vision de la recherche en synthèse de la parole, il pourrait s'avérer utile d'en considérer les conséquences pour les méthodes expérimentales en parole, et peut-être au-delà, en sciences humaines en général. Actuellement un résultat scientifique dans notre domaine n'est considéré comme véritablement intéressant que quand il caractérise une tendance générale ou même universelle. Pour emprunter une figure prise dans le monde de la prédiction statistique, ce n'est que la ligne droite (ou

courbée) caractérisant la relation directe entre valeur prédite et valeur mesurée qui intéresse. Les résultats et notions présentées ici suggèrent que le prochain pas scientifique consistera à établir pourquoi un point spécifique diverge de la ligne centrale — ou dans un espace multidimensionnel — quelles sont les relations qui relient les points divergents des diverses lignes centrales. Ce n'est que par ce type de savoir que nous pouvons commencer à combler nos lacunes en synthèse de la parole et ce n'est que par cet exercice que nous comprendrons véritablement les mécanismes de la production humaine de la parole tels que projetés sur des êtres humains véritables.

Références

Fant, G. (1956). On the predictability of formant levels and spectrum envelopes from formant frequencies. In M. Halle, H. Lunt, & H. Maclean (eds.), *For Roman Jakobson* (pp. 109-120). Mouton.

Joos, M. (1948). Acoustic Phonetics.*Language, 24,* Suppl., 1-136.

Klatt, D.H. (1980). Software for a cascade / parallel formant synthesizer. *Journal of the Acoustical Society of America, 67,* 971-995.

Klatt, D.H. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America, 82,* 737-793.

Klatt, D.H., & Klatt, L.C. (1990). Analysis, synthesis and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America, 87,* 820-857.

Koenig, W., Dunn, H.K., & Lacy, L.Y. (1946). The sound spectrograph. *The Journal of the Acoustical Society of America, 17,* 19-49.

Lee, C.-H. (1997). A unified statistical hypothesis testing approach to speaker verification and verbal information verification. *Speech Technology in the Public Telephone Network: Where are we Today?* Proceedings of COST Workshop (pp. 63-72). Rhodes, Greece.

Li, W.-Y., & O'Shaughnessy, D. (1997). Hybrid network based on RBFN and GMM for speaker recognition. *Eurospeech '97* (Rhodes, Greece), *Session T4C,* CD-ROM Edition.

Peterson, G.K., & Barney, H.L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America, 25,* 175-184.

Potter, R.K. (1930). *Proceedings I.R.E., 18,* 581-648. (cité dans Koenig et al., 1946).

Steinberg, J.C. (1934). *The Journal of the Acoustical Society of America, 6,* 16-24. (cité dans Koenig et al., 1946).

Stevens, K.N. (1995).

Stevens, K.N., & House, A.S. (1955). Development of a quantitative description of vowel articulation. *The Journal of the Acoustical Society of America, 27,* 484-493.

Styger, T., & Keller, E. (1994). Formant synthesis. In E. Keller (ed.), *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges* (pp. 109-128). Chichester: John Wiley.

Eric Keller
Laboratoire d'analyse
informatique de la parole (LAIP)
Section d'informatique

et de méthodes mathématiques
Université de Lausanne
Eric.Keller@imm.unil.ch