UNIVERSITE PARIS 7 – DENIS DIDEROT

DESS COMMUNICATION ET INFORMATION SCIENTIFIQUES, TECHNIQUES ET MEDICALES

ANNEE UNIVERSITAIRE 2001 - 2002

MAITRISER LA VOIX, DE LA PEDAGOGIE A LA SYNTHESE ARTIFICIELLE : UN PHENOMENE SONORE AUX FRONTIERES DES SCIENCES

Par François LASSAGNE

Directeur de mémoire : Thierry LEFEBVRE

Sommaire

Remerciements

A Thierry Lefebvre, directeur de ce mémoire, qui a cru au sujet que je lui proposais,

A Christophe d'Alessandro (LIMSI, CNRS) et Xavier Rodet (IRCAM), qui m'ont accordé temps et enthousiasme,

A Enrique Pardo, qui m'a reçu au Théâtre National de Chaillot et fait partager sa passion,

A mes proches, qui ne sont pas lassés de ma préoccupation monomaniaque et m'ont écouté leur parler…de vive voix.

Introduction

La voix humaine, outil spontané de communication, ne permet pas seulement l'échange d'informations objectives. Elle engendre aussi des modifications de l'état psychique du locuteur et de son auditoire. Livrée à elle-même, elle laisse transparaître des émotions ou des états mentaux ; maîtrisée, elle peut être perçue comme un instrument au pouvoir déconcertant.

Economiser sa voix, être persuasif, ne pas laisser transparaître ses émotions lors de la prise de parole, retenir l'attention, apaiser, étendre son registre, séduire, chanter juste… Autant de capacités que cherchent à acquérir, suivant le cas, les personnalités politiques, les chanteuses et chanteurs, les professionnels de la communication (conseillers, agents publicitaires, créatifs…), les comédiens …

Pour aboutir à la maîtrise vocale nécessaire à l'atteinte de tels objectifs, il existe différentes approches, des démarches centrées sur l'audition à l'entraînement presque sportif des organes vocaux auquel s'astreignent les interprètes lyriques.

Par ailleurs, de nombreux systèmes de synthèse et de reconnaissance vocale commencent à conquérir des applications grand public et quittent les laboratoires associant linguistique, phonétique et modélisation mathématique où ils avaient été développés dès les années 70. Les progrès de la microélectronique, les performances croissantes des processeurs et l'intégration progressive des études menées jusqu'à nos jours étendent champs d'application et performances, laissant entrevoir dans un avenir proche l'avènement de nouveaux systèmes de communication interpersonnelle et homme-machine, ou encore l'enrichissement des " réalités virtuelles ", sur un plan ludique ou artistique.

Quels liens les pratiques vocales à l'œuvre dans les écoles de chant ou les cours d'art dramatique, par exemple, entretiennent-elles avec les axes de recherches visant à élaborer les voix de synthèse de demain ?

Répondre à cette question ne peut s'envisager sans construire une réflexion sur la nature même de la connaissance contemporaine des phénomènes présidant à la production de la voix. De la simple évocation d'un don naturel à la définition des paramètres acoustiques permettant à la machine informatique de la faire naître, la voix est un matériau qui se prête à de multiples stratégies d'investigation.

Après avoir esquissé un aperçu de sa nature physiologique et de ses potentiels, nous tenterons de saisir dans un premier temps la manière dont les pédagogues des arts lyriques et dramatiques tentent de s'approprier les mécanismes de la phonation. Nous confronterons alors ce premier niveau de préhension de la voix aux stratégies de recherche des laboratoires travaillant sur la synthèse vocale.

Cette mise en relation nous conduira à des frontières au-delà desquelles la science se lie à l'art dans la recherche d'un inouï vocal, se soumettant à l'irréductibilité de la voix, universel anthropologique qui se refuse à toute description définitive.

A. Entre disposition innée et complexité physiologique : la voix, un territoire à conquérir

A.1. Les voix de « FIP » : un don et une sensibilité à la source d'une identité radiophonique

Le média radiophonique est, par essence, un média bruyant. Le flux continu déversé sur les ondes par l'industrie musicale alimente un océan sonore où l'oreille de l'auditeur risque à chaque instant la noyade, jusqu'à trouver, au fil des fréquences, l'instant et la station où respirer. Car les tubes, flashes, jingles et autres présentateurs à la logorrhée invasive laissent en général peu d'espace à la sérénité, se satisfaisant d'alimenter autant qu'ils renvoient en miroir la frénésie de la culture occidentale contemporaine.

Certaines stations de radio ont su cependant prendre le contre-pied de cette tendance à la cacophonie, misant sur la création d'un environnement sonore apaisant, apte à offrir à l'auditeur une écoute agréable et reposante. France Inter Paris, créée en 1971 à l'initiative de Pierre Codou et Jean Garetto, en est sans doute un des meilleurs exemples.

Imaginée à l'origine pour accompagner des automobilistes parisiens s'improvisant stratèges de l'itinéraire, à une époque où les quatre-roues s'imposaient comme le fléau indispensable des grandes agglomérations, FIP devait guider les conducteurs –les conductrices étaient alors minoritaires- à travers les embouteillages, les détendre et leur proposer des informations pratiques et culturelles relatives à Paris. Mais conduire requiert une part d'attention que l'écoute d'une émission ne saurait occulter. Aussi, les programmateurs de la station furent-ils invités à élaborer des enchaînements musicaux fluides, où les œuvres diffusées n'étaient annoncées qu'à l'issue d'une série de quatre ou cinq titres. Il s'agissait de ne pas solliciter l'attention de l'automobiliste trop curieux de connaître l'interprète d'une chanson inédite. Jazz, musique classique, standards de la variété étaient donc assemblés en une alchimie confiée à l'inspiration d'une petite équipe de programmateurs peu préoccupés par les tendances commerciales de l'industrie musicale.

Si la musique électronique, la world music, l'acid jazz, le trip-hop ou le rap s'intègrent, aujourd'hui, à la programmation de FIP, c' est en perpétuant la volonté d'établir un « fonds sonore » discret mais pas insipide, à écouter aussi bien sur les grands axes asphyxiés que dans le calme d'un salon. La station conserve donc sa stratégie musicale. Mais, surtout, elle continue de s'appuyer sur le deuxième pilier fondateur de son identité : ses voix féminines.

En 1972, un chauffeur-livreur auditeur de FIP confiait : « Je trouve que leur voix est excitante. Quand je les entends, il me semble que j'ai affaire à des minettes perverses » [1].

Tous les auditeurs, aujourd'hui comme hier, n'affichent pas une réaction aussi marquée à l'écoute des « fipettes ». Néanmoins, il est probable que peu prétendent y être indifférents. Les animatrices de la station ne tiennent pourtant pas de propos particulièrement suggestifs : quartiers à éviter pour cause d'embouteillages, horaires et lieux de concerts, anecdotes sur un artiste programmé, météo, renseignements pratiques… De plus, leurs interventions n'excèdent pas deux minutes. Leur « pouvoir » vient donc d'ailleurs. Jean Garetto offre une définition de cet « ailleurs », interrogé sur la façon dont les « fipettes » étaient sélectionnées pour la station à sa création : « […] notre seul critère était l'audition. Et le charme. Voilà le mot que je cherchais, c'était le charme. Quand on parle du charme vocal d'une femme, on ne s'attend pas à ce qu'elle vous claironne les choses, mais plutôt qu'elle soit douce et enjouée » . [2] Dans une interview antérieure, il avait indiqué : « Nous avons établi le prototype de voix avec Kriss, de l' « Oreille en coin » ; une voix qui a du timbre, légèrement sophistiquée –comme celle dite de « l'hôtesse de l'air »- une bonne dose d'humour et de fantaisie en plus »[3].

En 1972, le succès de la station auprès des auditeurs et, en conséquence, l'enracinement de son identité, reposait donc en grande partie sur le « charme vocal » de ses huit animatrices. Celles-ci, recrutées sur audition, avaient alors entre 25 et 28 ans.

Aujourd'hui, si l'identité de FIP a peu évolué, c'est qu'elle s'appuie encore sur le même type de voix : féminines, jeunes, charmantes. Dominique Pensec, actuelle directrice de la station, précise cependant que les recrutements récents, toujours réalisés par audition des candidates, visent à retenir celles dont la voix évoque le plus de naturel : la spontanéité fait office de leitmotiv, là où la sophistication constituait un élément central de la marque de fabrique de la station au début des années 70.

Repérer les candidates potentielles tient sans doute plus de l'opportunisme que de la quête éclairée, dans la mesure où il semble difficile de revendiquer sa compétence vocale comme on attesterait de ses qualifications en produisant un diplôme. L'empirisme et la chance constituent donc l'essentiel de la démarche. On peut néanmoins s'interroger sur l'intégration des heureuses élues à leur environnement professionnel. Le « don » vocal, qui

---

[1] Propos rapportés dans « La radio pour les automobilistes parisiens part à la conquête de la province », *Renault-Promotion* , n°79, mai 1972

[2] Propos rapportés par Guy Robert, « Une chaîne à contre-courant : FIP 514 », *Cahiers d'Histoire de la Radiodiffusion*, , n°70, 2001, p.70

[3] in *Signature*, n°136, septembre 1981

leur offre une place derrière un micro, suffit-il à en faire des animatrices efficaces, capables, notamment, de régler leur débit de parole au temps limité de leurs interventions ? Le ton perçu par l'auditeur est-il rigoureusement le même que celui qu'elles auraient dans une conversation courante ? Autrement dit, une part de travail, d'entraînement, intervient-elle dans la création d'une « fipette » ?

D'après Jean Garetto, la formation des animatrices était minime. «On leur disait de ne pas parler brutalement, de ne pas scander trop leurs syllabes, et de parler doucement. Mais on ne les entraînait pas. On avait du mal à les choisir et à les trouver, et après elles avaient leur ton, elles avaient ce ton-là dans la vie »[4]. Minime mais prépondérante puisque aujourd'hui encore un ton doux, parfois proche du murmure, caractérise les interventions des « fipettes ». Plus importante encore est sans doute l'influence de Kriss, la pionnière. Elle anime actuellement l'émission « Portraits sensibles », sur France Inter, où sa « voix de gamine »[5] contribue sans doute à susciter les confidences de ses interlocuteurs anonymes. Il est fort probable que sa position initiale de modèle vocal pour FIP se perpétue à présent tacitement tant lors du recrutement des animatrices que dans la façon qu'ont celles-ci de tenir leur rôle.

Ainsi, les voix de FIP exploitent-elles à la fois leurs capacités physiologiques naturelles et leur manière d'être pour créer une ambiance, un ton caractéristique, sans recourir véritablement à un exercice méthodique de leur talent. Macha Béranger, autre voix de radio caractéristique qui officie sur France Inter et s'identifie sans ambiguïté possible à travers son timbre usé, rauque et chaud, ne déroge pas à ce constat.

Il apparaît donc clairement que certains domaines d'activités où l'usage de la voix est le cœur du métier n'exigent guère plus de leurs représentants que… d'être eux-mêmes. Si nous venons de l'illustrer dans le cas de FIP, il en serait tout autrement à considérer les chanteurs professionnels. Bien entendu, quiconque ne s'improvise pas chanteur : une certaine conformation physiologique des organes vocaux constitue le bagage minimum d'une carrière artistique vocale. Néanmoins, le travail, l'entraînement et, surtout, la connaissance de ses propres capacités, fondent la plupart du temps le succès et la longévité des interprètes féminins et masculins. De même, les personnes dont le métier exige une haute qualité de présentation face à un public –comédiens, hommes politiques, présentateurs de télévision, dirigeants de grandes entreprises…- tendent à recourir à des

[4] Guy Robert, *loc. cit.*
[5] Télérama , n°2738, 3 juillet 2002

activités relevant de l'entraînement, tout au moins de la découverte, de leurs capacités vocales.

Entre la spontanéité recherchée des « fipettes » et la maîtrise technique d'une cantatrice d'opéra, la voix humaine se prête donc à l'investigation et est susceptible d'être modifiée, voire améliorée, relativement à la finalité qu'on se fixe.

Une première façon d'appréhender la voix, pour quiconque entend agir sur elle, consiste à mieux comprendre la physiologie qui la sous-tend.

A.2. Les organes de la voix, prémisses de la maîtrise vocale

Bien qu'intimement lié à ce qui fait la nature de l'Homme, le phénomène sonore qu'est la voix humaine met en jeu des organes qui ne lui sont pas spécifiquement dévolus. On parle ainsi d' « appareil de la phonation », ou d' « appareil vocal », même si cette terminologie ne traduit que le résultat fonctionnel des organes en jeu dans la production de la voix.

A l'instar des instruments à vent dont l'embouchure est munie d'une anche, l'appareil de la phonation se décompose en trois parties fonctionnelles. La première est la source énergétique de la voix : c'est la «soufflerie », autrement dit l'appareil respiratoire dans son ensemble. La deuxième correspond (en première approche) à la anche des instruments à vent, il s'agit du larynx, à l'origine de la vibration sonore, et la troisième, enfin, est formée par les diverses cavités de résonance dans lesquelles le son émis au niveau du larynx va se voir amplifié et modifié.

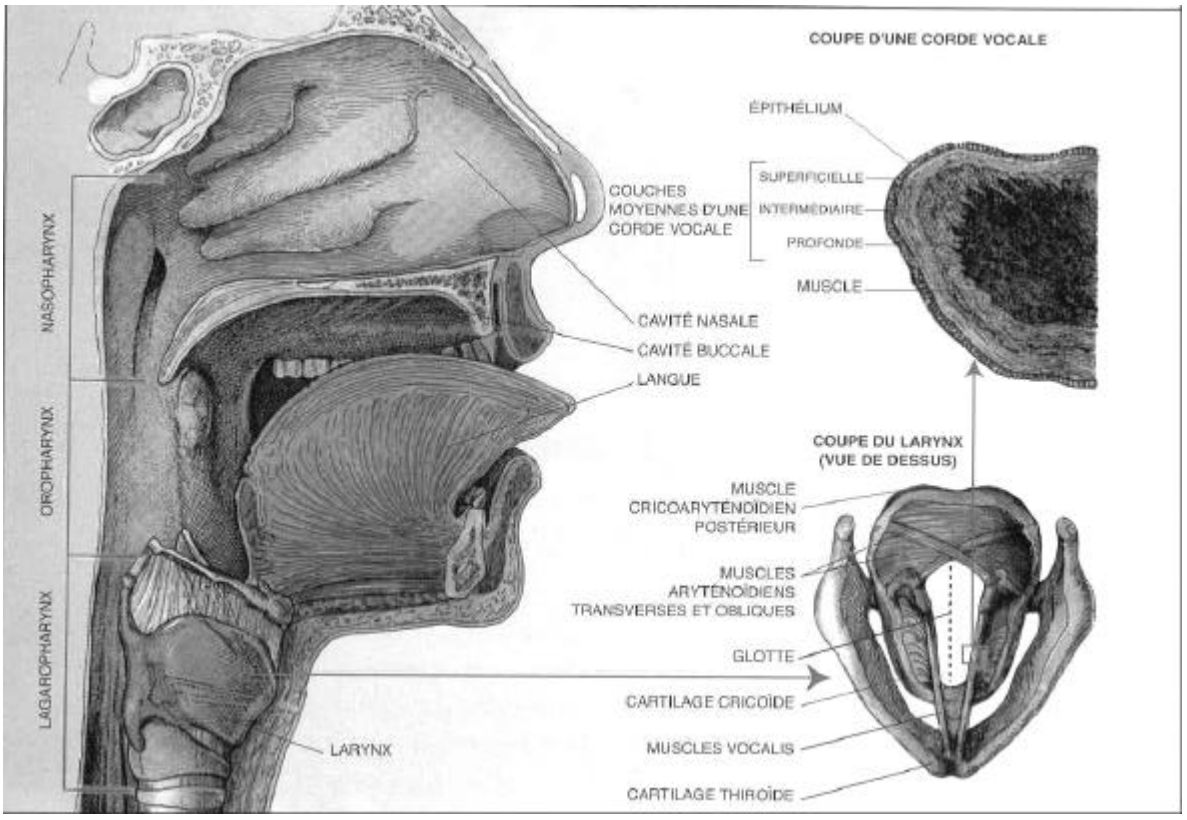
Lorsque l'on parle ou que l'on chante, le mouvement respiratoire s'adapte d'une manière très différente à la conformation de la respiration normale. Alors que la phase expiratoire, où l'air est expulsé des poumons, se rallonge considérablement, la durée de la phase inspiratoire est, elle, raccourcie. En corollaire, les volumes d'air mobilisés dans les poumons sont beaucoup plus importants que ceux observés dans le cas de la respiration. De même, la pression de l'air expulsé des poumons est plus grande, car le flux expiré rencontre avant sa sortie l'obstacle des cordes vocales, dont il va provoquer la vibration.

Cette vibration des cordes vocales s'effectue, comme nous l'avons indiqué, au niveau du larynx. Celui-ci s'étend du haut de la trachée à l'épiglotte, et cette dernière correspond au «couvercle » qui ferme les voix respiratoires aux aliments ou à la salive en provenance du pharynx, lors de la déglutition.

Le larynx lui-même est un organe particulièrement complexe, mettant en jeu plusieurs faisceaux musculaires ainsi que divers cartilages mobiles ou fixes. Rattaché dans sa partie supérieure au crâne et à la mâchoire inférieure, il est reliée par le bas aux premiers anneaux de cartilage de la trachée. Trachée et mâchoire inférieure étant susceptibles de se déplacer, le larynx est donc un organe extrêmement mobile, verticalement aussi bien qu'horizontalement.

Les cordes vocales à proprement parler correspondent en fait à une paire de muscles et de ligaments superposés. Elles se situent dans un plan sensiblement horizontal, s'insérant à l'avant sur le cartilage thyroïde (la «pomme d'Adam ») et à l'arrière chacune de part et d'autre de la glotte, à l'extrémité de cartilages mobiles. Les cordes vocales sont donc

mobiles elles aussi et, puisqu'il s'agit de muscles, capables de se contracter. Les fibres nerveuses et la muqueuse qui les constituent s'organisent cependant de façon très spécifique, en cinq couches précisément, dont les propriétés mécaniques diffèrent. Cette particularité leur confère la capacité d'être le siège de mouvements ondulatoires, à la source des vibrations sonores qu'elles émettent lors de la phonation. Toutefois, les cordes vocales ne font pas vibrer l'air comme celles d'un violon. Elles libèrent l'air sous pression en provenance des poumons, en ouvrant et fermant alternativement la glotte. C'est ce « hachage » de l'air qui produit la vibration acoustique, par un mécanisme qui ressemble au claquement de mains. Ainsi, si l'on ralentit fortement un enregistrement d'une personne qui parle, on finit par percevoir non plus un phénomène sonore continu, mais une série de « tops » parfaitement séparés les uns des autres.



Organes de la phonation *- Illustration tirée de R. Sataloff, « La voix humaine », Pour la Science Dossier « Le monde des sons », n°32, juillet-octobre 2001, p.10*

En résumé, l'air expulsé par les poumons frappe les cordes vocales, celles-ci le font vibrer à des fréquences[6] audibles par l'oreille humaine… et la voix s'élève ? Pas tout à fait. En effet, le son émis au niveau des cordes vocales est encore assez faible, et, si cela était possible, on peinerait à l'entendre, sans équipement d'amplification adéquat, en le

[6] La fréquence du son laryngé, mesurée en hertz (Hz), correspond au nombre de cycles d'ouverture et de fermeture de la glotte en une seconde.

cherchant au niveau du larynx. Comme une contrebasse dont on aurait supprimé la caisse, il reste encore au son émis à traverser les différentes cavités où il va être amplifié et modulé.

Le pharynx, région de la gorge située entre la bouche et l'œsophage, ainsi que les cavités buccale et nasale agissent justement comme une série de résonateurs où l'intensité de l'émission sonore laryngée va augmenter fortement. Toutefois, on ne saurait comparer ce mécanisme à celui qu'on observe dans la plupart des instruments à vent. Dans une trompette, par exemple, le son émis au niveau de l'embouchure par les lèvres du musicien est lui aussi amplifié, dans son trajet vers le pavillon, par le tube qu'il traverse et qui fait office de résonateur. Ce tube est dimensionné une fois pour toute, et il est fixe. Dans la phonation, le système est beaucoup plus complexe, puisque les parois des résonateurs sont flexibles et leur forme variable. Par ailleurs, nous verrons plus loin que l'amplification de la source laryngée n'est absolument pas linéaire. Autrement dit, les résonateurs n'agissent pas simplement comme un potentiomètre de volume mais sélectionnent certaines parties du son, qu'ils vont amplifier, alors que d'autres sont au contraire atténuées. De plus, la langue, les lèvres et les dents vont encore modifier considérablement la nature du son en provenance des cordes vocales.

Sans entrer dans le détail des mécanismes physiologiques qui donnent corps à la voix, il est clair d'ores et déjà que les leviers susceptibles d'être actionnés pour la modifier sont nombreux. Le contrôle de la respiration –les volumes d'air échangés, leur pression, leur rétention dans la trachée, leur débit-, celui du larynx –sa propre position, la position et la tension des cordes vocales en son sein- ainsi que les différentes ouvertures possibles de la bouche et les nombreux placements de la langue offrent autant de degrés de liberté à qui souhaite jouer de l'instrument extrêmement souple qu'est la voix.

Un exemple assez frappant illustre parfaitement cette étendue des capacités de l'appareil de la phonation. Il s'agit du chant diphonique. Phénomène étonnant mis en évidence essentiellement grâce aux travaux de différents ethnologues, il met en jeu, pour l'essentiel, un usage particulièrement bien maîtrisé des résonateurs pharyngé, buccal et nasal. Il va nous donner l'occasion d'aborder la manière dont on représente le son vocal. Instrumentation électronique et représentation graphique idoine tiennent en effet une part importante dans la compréhension de la phonation et constituent un référentiel commun à de nombreuses approches, dès que l'on souhaite focaliser le travail de la voix sur un de ses différents « paramètres » de contrôle.

A.3. Le chant diphonique : une performance auscultée par la science

Avant d'illustrer l'importance du rôle des résonateurs pharyngé, buccal et nasal à travers le phénomène du chant diphonique, il est indispensable de préciser plus avant la nature du son vocal, et la manière dont les cavités naturelles de résonance le transforment.

Le son laryngé lui-même fait l'objet, depuis les années 1950, d'une investigation quantitative. Certains équipements permettent ainsi d'étudier les mouvements d'ouverture et de fermeture de la glotte. Une solution consiste notamment à éclairer les cordes vocales avec un stroboscope : c'est la strobovidéolaryngoscopie. L'électroglottographie mesure, quant à elle, les variations de pression de l'air au niveau de la glotte, alors que l'électromyographie permet d'étudier les influx nerveux contrôlant les mouvements des muscles du larynx. Si ces instruments sont à l'origine de nombreuses découvertes sur les phénomènes physiologiques qui président à l'émission vocale laryngée, ils ne renseignent en revanche pas sur le rôle des résonateurs.

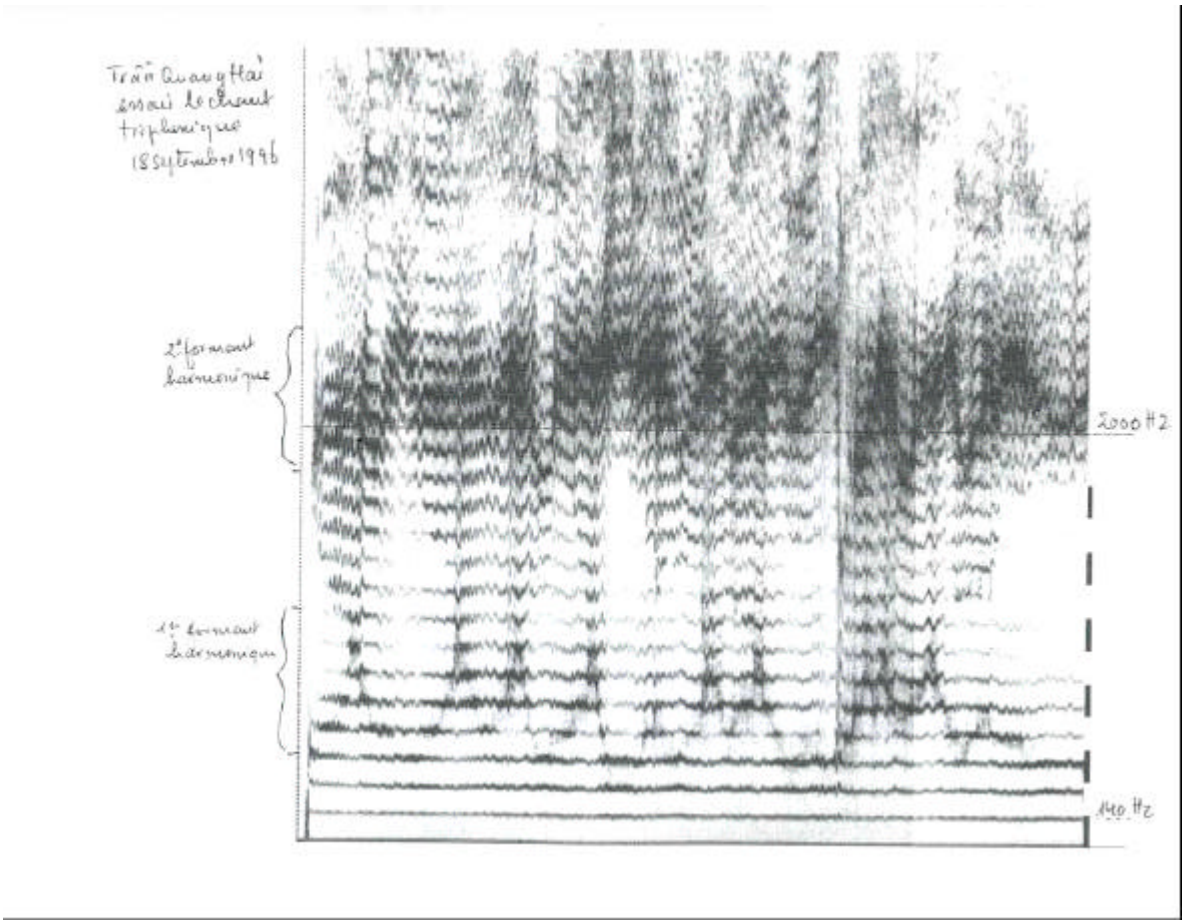
Le son émis au niveau du larynx par les cordes vocales est un son complexe, composé d'une fréquence fondamentale et de plusieurs harmoniques, ou partiels, de fréquences supérieures. Ce qu'on appelle le « timbre » d'une voix correspond en première approche au rapport moyen des intensités des différentes harmoniques par rapport à la fréquence fondamentale. On a pu constater qu'au niveau laryngé, ce rapport, appelé aussi spectre du son, est quasi identique pour les chanteurs débutants ou professionnels[7]. On perçoit pourtant bel et bien un timbre différent, suivant le locuteur ou le chanteur que l'on écoute, de la même façon que l'on différencie le son d'une flûte ou d'un violon qui joueraient la même note fondamentale. La maîtrise du contrôle des résonateurs s'annonce donc comme un des paramètres fondamentaux dans l'obtention d'une voix maîtrisée.

Cela suppose de réussir à ajuster les dimensions du « tractus vocal » –pharynx, bouche, lèvres- de façon assez fine pour sélectionner les fréquences à amplifier et celles que l'on souhaite atténuer. Les cavités naturelles agissent en effet comme des filtres qui, suivant leur dimension, vont amplifier ou atténuer l'énergie de chaque harmonique en provenance du larynx et transformer le bourdonnement inaudible des cordes vocales en sons expressifs, à la manière d'un souffle devenant son lorsqu'il est dirigé sur le goulot d'une bouteille. Chez les chanteurs, en général quatre ou cinq bandes particulières de fréquences sont

[7] R. Sataloff, « La voix humaine », *Pour la Science Dossier « Le monde des sons »*, n°32, juillet-octobre 2001, p.10-15

amplifiées. On les appelle « formants ». L'intensité de la source vocale laryngée diminue uniformément sur tout le spectre, sauf pour ces plages de fréquences particulières, où elle est augmentée.

L'observation de la répartition des harmoniques se fait en général à partir de l'enregistrement analogique ou numérique de la voix que l'on cherche à étudier. Cet enregistrement est alors appliqué à l'entrée d'un appareil appelé Sona-Graph, qui permet d'obtenir l'image du son sur un seul graphique : un sonogramme. En abscisse figure l'information *temps*, en ordonnée l'information *fréquence*, et l'épaisseur du tracé traduit l'*intensité* observée pour chaque fréquence. Cette représentation n'est d'ailleurs pas sans rappeler la décomposition par diffraction d'une source lumineuse en ses différentes composantes fréquentielles.



Sonogramme - *Trân Quang Haï, « Recherches introspectives sur le chant diphonique et leurs applications »,* Penser la voix*, La Licorne, UFR Langues Littératures, Poitiers, 1997, p.213*

Il y a peu de chances pour que les Touvas, bergers semi-nomades de la république autonome éponyme, située entre Russie et Mongolie, se soucient de la transcription graphique sur sonogramme du chant qui a fait leur renommée en occident depuis une trentaine d'années. Il s'agit pourtant d'une performance vocale qui, analysée à la lumière des connaissances évoquées jusqu'ici sur les mécanismes de la phonation, ne cesse d'étonner.

Les bergers Touvas émettent, dans les étendues silencieuses de la taïga, un chant à deux mélodies, produites *simultanément* par le même chanteur. «La première mélodie est un son fondamental grave, soutenu, semblable au bourdon d'une cornemuse. La seconde est une série d'harmoniques à un ou plusieurs octaves d'intervalle, qui rappelle le sifflement d'un oiseau, les rythmes syncopés d'un torrent de montagne ou le petit galop d'un cheval.»[8].

Cette technique du chant diphonique est répandue « non seulement dans toute la partie du monde se trouvant autour du Mont Altaï en Haute Asie et peuplée de Mongols, Touvas, Khakash, Bachkirs, Altaiens, mais également à un certain degré, parmi les Rajasthanais de l'Inde, les Xhosas d'Afrique du Sud, et les Moines tibétains des monastères Gyütö et Gyüme »[9]. Elle a fait son chemin jusque dans la musique contemporaine, dans les courants de la world music, du new age, dans le jazz…

Comme nous l'avons précisé plus haut, lorsque l'on parle ou que l'on chante, la majeure partie de l'intensité acoustique est concentrée dans le son fondamental, alors que les harmoniques déterminent le timbre de la voix. Dans le chant diphonique, une harmonique particulière acquiert une intensité telle qu'elle est perçue comme un son distinct, proche d'un « sifflement désincarné »[10].

Ce sifflement, pour devenir chant, doit réussir à suivre une ligne mélodique. Autrement dit, l'auditeur doit être capable de percevoir des changements de tonalité, discerner les variations de hauteur du son. Or la capacité à distinguer des hauteurs de son différentes est fortement liée aux modalités individuelles de perception acoustique, et impose aux chanteurs Touvas une manière très particulière de sélectionner les harmoniques qu'ils choisissent de renforcer.

« Les manuels d'acoustique classiques disent que la hauteur des sons harmoniques, donc les sons comportant un fondamental de fréquence F et une suite d'harmoniques F1, F2, F3… multiples de F, est donnée par la fréquence du premier son fondamental. Cela n'est pas tout à fait juste car il est possible de supprimer électroniquement ce fondamental sans pour cela changer la hauteur subjective du son perçu. Si cette théorie était vraie, une chaîne électro-acoustique ne reproduisant pas l'extrême grave changerait la hauteur des sons. Il n'en est rien car le timbre change mais pas la hauteur. […] La définition du spectre à formant est l'accentuation en intensité d'un groupe d'harmoniques constituant un

---

[8] T. Levin et M. Edgerton, « Le chant des Touvas », op. cit, p. 16-19
[9] Trân Quang Haï, « Recherches introspectives sur le chant diphonique et leurs applications », *Penser la voix*, La Licorne, UFR Langues Littératures, Poitiers, 1997, p.197
[10] T. Levin et M. Edgerton, *loc.cit.*

formant, donc une zone de fréquences où l'énergie est grande. Considérant ce formant, une deuxième notion de la perception de la hauteur se fait jour. On s'est en effet aperçu que la position du formant dans le spectre sonore donnait la perception d'une nouvelle hauteur. […] Cette théorie doit être nuancée, car certaines conditions s'imposent. Comme expérience, on chante trois DO (DO1, DO2, DO3) à une octave entre deux DO en portant la voix comme pour s'adresser à un grand auditoire. On constate avec un sonogramme que le maximum d'énergie se trouve dans la zone sensible de l'oreille humaine (2 à 3 KHz). Il s'agit bien d'un formant situé entre 2 et 4 KHz. On enregistre ensuite trois DO dans la même tonalité, mais cette fois en posant la voix pour s'adresser à un auditoire restreint, et on observe la disparition de ce formant. Dans ce cas, la disparition du formant ne change pas la hauteur des sons. La perception de la hauteur par la position du formant n'est possible que si celui-ci est très aigu, à savoir que l'énergie du formant n'est répartie que sur deux ou trois harmoniques. Donc, si la densité d'énergie du formant est grande, et que le formant est étroit, celui-ci donnera une information de hauteur en plus de la tonalité globale du morceau chanté. Par ce biais, on arrive à la technique du chant diphonique. »[11]

Des études expérimentales basées sur des observations en radiocinématographie et naso-endoscopie, réalisées auprès de certains chanteurs Touvas, ont montré que ce mécanisme spécifique de renforcement des harmoniques s'appuie sur trois composantes corrélées. Tout d'abord, chaque harmonique constitutive de la mélodie se superposant au « bourdon » grave est ajustée au centre d'un formant. Ensuite, la durée de la phase de fermeture des cordes vocales est allongée. Enfin, l'intervalle des harmoniques affectées par le formant est réduit. Ces trois composantes ont pour rôle essentiel d'affiner la sélectivité des formants et d'augmenter l'intensité des harmoniques aiguës. Physiologiquement, pour parvenir à ce résultat, les chanteurs qui pratiquent le chant diphonique ont tendance à contracter les muscles de la gorge. Cela a pour conséquence de provoquer une fermeture brutale et prolongée des cordes vocales. Par ailleurs, ils amincissent les lèvres, arrondissent la bouche qu'ils maintiennent presque fermée : les pertes d'énergie acoustique sont amoindries, et les harmoniques, avant d'être renforcées par un formant donné, sont déjà plus perceptibles que dans le chant « normal ».

La méthode de la sélection des harmoniques et du renforcement acoustique suffisant à garantir leur perception, reste à créer la mélodie, c'est-à-dire faire entendre plusieurs harmoniques différentes successivement…en conservant toujours la fréquence grave

[11] Trân Quang Haï, *loc. cit*. p. 201

fondamentale du «bourdon ». Là encore, ce sont des déplacements précis des cavités de résonance qui le permettent, selon des techniques différentes en fonction du chanteur pratiquant le chant diphonique. Il en existerait quatre chez les Touvas. A titre d'exemple, l'une d'elles consiste à bomber la langue progressivement, pour sélectionner des harmoniques de plus en plus aiguës. Et, pour accorder avec précision le formant à chaque nouvelle harmonique, il « suffit » d'ouvrir périodiquement les lèvres. Certains auteurs préconisent même une « méthode » du chant diphonique :

« On applique la « recette » décrite ci-dessous :

1. Chanter avec une voix de gorge
2. Prononcer la lettre L. Dès que la pointe de la langue touche le centre de la voûte palatine, maintenir ainsi cette position.
3. Prononcer ensuite la voyelle Ü avec, toujours, la pointe de la langue collée fermement contre le point de fixation entre le palais dur et le palais mou.
4. Contracter les muscles du cou et ceux de l'abdomen pendant le chant comme si on essayait de soulever un objet très lourd.
5. Donner un timbre très nasalisé en amplifiant les fosses nasales.
6. Prononcer ensuite les deux voyelles I et Ü (ou bien O et A) liées mais alternées l'une après l'autre en plusieurs fois.
7. Ainsi sont obtenus le bourdon et les harmoniques en pente ascendante et pente descendante selon le désir du chanteur. On varie la position des lèvres ou celle de la langue pour moduler la mélodie des harmoniques. La forte concentration musculaire augmente la clarté harmonique. »[12]

Il n'est pas certain qu'une telle «recette » suffise à faire de chacun un virtuose du chant diphonique. Néanmoins, ceux qui le pratiquent aux quatre coins du monde ne présentent aucune disposition physiologique particulière, aucun don qui les privilégierait dans cet art. L'exemple, certes très spécifique, du chant diphonique montre donc bien que la maîtrise de la voix est avant tout une question d'expérimentation et de connaissance. Etre capable de saisir le rôle du souffle, ressentir les mouvements du larynx et parvenir à les contrôler dans la voix chantée comme dans la voix parlée, ajuster volume et tonalité, jouer sur la forme de la bouche… Comédiens et chanteurs sont parmi les usagers professionnels

[12] Trân Quang Haï, *loc. cit.* p. 205

de la voix les plus familiers de ces notions, qu'ils mettent au service de leur art à travers un certain nombre d'exercices.

Bien entendu, théâtre et chant ne se préoccupent guère de sonogrammes ou de radiocinématographie. Néanmoins ils en exploitent les résultats, parfois inconsciemment, parfois en toute connaissance de cause, lorsqu'ils choisissent leurs écoles et méthodes d'éducation vocale.

B. Construire la maîtrise vocale

B.1. Enseigner la voix au théâtre

Nous empruntons les lignes suivantes à Sarah Bernhardt, qui prodigua dans un petit ouvrage quelques conseils à une génération de jeunes comédiens qu'elle considérait avec bienveillance.

« La voix est l'instrument le plus nécessaire à l'artiste dramatique. C'est elle qui fixe l'attention du public, c'est elle qui lie l'artiste et l'auditoire. Il faut qu'elle ait toutes les harmonies, graves, plaintives, vibrantes et métalliques. Jamais une voix ayant des trous ne permettra à un artiste le développement complet de son art, quelle que soit son intelligence. Il se butera à un obstacle intangible et cependant matériel »[13].

La grande comédienne que fut Sarah Bernhardt avait une vision peut-être romantique, attachée à des styles vocaux propres à un théâtre classique dont nombre de comédiens contemporains se défieraient. Elle situait néanmoins la voix comme l'enjeu central de la représentation théâtrale, et nous verrons plus loin que ce choix non seulement se confirme aujourd'hui dans divers courants de création, mais qu'il s'enrichit d'ouvertures qui dépassent largement le seul souci de complétude mis en avant dans la citation précédente. En se limitant à ce stade à la nécessité pour les comédiens de disposer d'une voix capable de se prêter à « toutes les harmonies », l'espace ouvert au travail, à l'apprentissage, est considérable. Car le talent vocal des artistes dramatiques ne peut se limiter à leurs caractéristiques individuelles, ou aux effets qu'ils maîtrisent naturellement. « Bien souvent des comédiens et comédiennes n'ont retenu l'attention du public que par le timbre de leur voix. Cela ne suffit pas. La voix n'est qu'un instrument dont l'artiste doit apprendre à user avec souplesse et sûreté comme de ses membres »[14].

Ce nécessaire entraînement, qui poursuit des finalités bien différentes suivant les écoles où il est pratiqué, repose au moins sur des bases communes incontournables. La gestion du souffle, l'articulation, le débit et l'intensité forment ainsi les paramètres physiques sur lesquels il est indispensable de concentrer efforts et attention. Naturellement, ils ne sauraient être indépendants. Le débit aura tendance à augmenter si la réserve d'air contenu dans les poumons est trop faible, alors que l'intensité diminuera. De même,

---

[13] Sarah Bernhardt, *L'art du théâtre. La voix, le geste, la prononciation*, L'Harmattan, coll. Les introuvables, Paris, 1993, p.41
[14] Sarah Bernhardt, *op. cit.*, p.52

l'articulation pâtira peut-être d'un débit trop rapide ou d'une voix trop peu appuyée. Suivant les capacités propres de l'élève, en fonction du texte à jouer, de l'acoustique du lieu de représentation, en prenant en compte la sensibilité des autres comédiens, du metteur en scène et de celle, attendue, du public, il appartiendra à chacun, lors des répétitions, de définir sa propre projection vocale, celle qui saura faire vivre le texte et susciter l'émotion.

Certaines troupes s'entraîneront essentiellement à travers des exercices respiratoires, d'autres s'essaieront au cri, d'autres encore pourront tenir des joutes de diction… Dans tous les cas, tout comme le chant diphonique ne disposait d'aucune règle définitive d'apprentissage –hormis les «recettes » proposées, dont le succès dépend avant tout de la patience et de l'expérimentation personnelle de l' apprenti chanteur- la performance de la voix au théâtre ne saurait se garantir par les recommandations fixées dans un quelconque manuel.

Il en va différemment avec la voix chantée. La connaissance des mécanismes physiologiques engagés dans le chant est, depuis fort longtemps, mise à profit dans des écoles. Ces lieux spécifiques d'apprentissage, où un maître alterne cours théoriques et séances d'apprentissage, forment un cadre où l'élève évolue le plus souvent selon des directions bien balisées. La voix y est, là aussi, l'instrument-roi, mais la part de découverte empirique y est beaucoup moins marquée qu'au théâtre, au profit de méthodes et de stratégies pédagogiques en relation étroite avec les représentations scientifiques relatives à la phonation.

B.2. Enseigner la voix aux chanteurs

Là où la description physiologique de l'appareil phonatoire distinguait trois fonctions essentielles dans la production de la voix et mettait l'accent, entre autres, sur le rôle de la respiration, un vieux dicton italien affirme que « l'art de chanter est l'école de la respiration ». De nombreux auteurs reconnaissent d'ailleurs à l'Italie une place de premier plan dans une vision du chant basée sur le souffle, en particulier dans le style musicale dit du « Bel Canto ». Les castrats du XVIII^{ème} siècle se livraient à des joutes consistant à mesurer le temps de leur prestation vocale sans prises de respirations successives. Les chanteurs lyriques professionnels, s'ils ne se prêtent plus à de tels concours, continuent à accorder une importance prépondérante au souffle. «Pourquoi le souffle compte-t-il tant ? Parce que bien dispensé, il est le grand élément physiologique par lequel le chanteur a

quelque prise pour assurer aux cordes vocales leur plein essor »[15], écrivent Jacques et Catherine OTT, qui se sont livrés à une minutieuse et sans complaisance étude des différentes écoles de chant européennes contemporaines.

Hauteur et intensité, tenue des notes, qualité d'un vibrato, dépendent effectivement de façon étroite du dosage du souffle et de la capacité respiratoire du chanteur. De même, la direction du volume d'air expiré et mis en vibration par les cordes vocales influe directement sur la nature du son perçu, rencontrant suivant le cas majoritairement les cavités nasale ou buccale. Dans la même perspective, la différenciation opérée entre respiration ventrale et respiration thoracique, qui figure en bonne place dans la pratique du yoga, contribue grandement à la qualité de la voix chantée. L'air au sortir de la bouche n'attend pas de quitter le corps qui l'émet pour vibrer, et la « colonne d'air » qui a sollicité les cordes vocales peut, selon les cas, amplifier ou atténuer le son émis.

La respiration joue donc un rôle majeur et permanent dans le chant, mais elle n'est pas le seul élément de la physiologie de l'appareil vocal à offrir des stratégies de travail aux pédagogues du chant. La notion de registre tient elle aussi une place de choix.

Un registre correspond à l'étendue vocale sur lequel le timbre d'un individu reste à peu près identique, alors que les « passages » s'identifient aux notes au-delà desquelles s'effectue un changement dans le mode de production de la voix. Un même sujet peut avoir plusieurs registres, séparés par plusieurs passages. L'observation instrumentée de l'émission de la voix chantée a permis de découvrir que les registres sous-tendent, pour l'essentiel, un fonctionnement différent des cordes vocales. Néanmoins, cette découverte est très récente.

« Il est remarquable que les pays latins –surtout l'Italie- aient pris conscience du phénomène des registres bien avant leur confirmation physiologique datant des découvertes neuro-phonatoires de 1950, car seule la théorie musculo-élastique de la glotte régnait dans l'Europe romantique, incitant quantité d'enseignants à nier la registration italienne puisque le souffle, uniquement, devait créer le son en passant au travers des rubans vocaux considérés comme une anche d'instrument à vent. Les registres façonnent dans l'organisme des sensations vibratoires nettes et localisées mais secondaires au travail des cordes vocales. Ces sensations ressenties soit dans le thorax soit dans la tête ont été prises, autrefois, pour la cause du registre lui-même : il ne s'agit là, en fait, que d'une conséquence indirecte, quoique ces diverses localisations vibratoires influencent bénéfiquement la

[15] Jacques et Catherine OTT, *La pédagogie et les techniques européennes du chant*, éditions EAP, Issy les Moulineaux, 1994, p. 81

conduite de la voix par la perception qu'elles fournissent au chanteur du bon fonctionnement nerveux et sonore de ses cordes vocales »[16].

Les registres sont nommés en relation avec la tessiture de la voix qu'ils recouvrent : registre grave, registre medium, registre aigu. D'anciennes dénominations, qui se sont intégrées au vocabulaire courant, font correspondre le registre «de poitrine » au registre grave, le registre « de fausset » au registre medium et le registre « de tête » au registre aigu, en raison de la localisation des sensations vibratoires produites par chaque registre dans le corps.

Sur le plan physiologique, le registre grave se caractérise par le fait que les cordes vocales sont le siège de vibrations de grande amplitude, en position de faible élongation. Le registre aigu se distingue a contrario par des vibrations de faible amplitude et à forte élongation des cordes vocales. Selon le sexe et la physionomie, les différents registres s'étalent sur des plages de notes variables, le registre aigu étant, par exemple, plus étendu chez les femmes. En conséquence, un sujet aura plus ou moins de facilité à effectuer les passages d'un registre à l'autre, et sera plus ou moins à l'aise dans un registre donné.

En terme de pédagogie, si certaines écoles continuent à nier l'existence des registres, malgré leur fondement physiologique, d'autres ne les exploitent pas nécessairement sans risque pour l'élève. Il semble que l'essentiel de l'apprentissage devrait consister à faire sentir les notes de passage, afin que l'élève soit à l'aise dans les transitions d'un registre à l'autre et, ainsi, parvienne à découvrir et maîtriser ses propres ressources. Malheureusement, certains pédagogues tentent au contraire de forcer un registre plutôt qu'un autre, influençant sélectivement la tessiture de la voix et la fatiguant dangereusement.

Appréhension du mécanisme de la registration et contrôle du souffle offrent, quand ils sont bien menés, deux axes de travail centraux dans la recherche performative de la voix. Néanmoins, comme le suggérait la description succincte des fondements physiologiques de la phonation que nous avons tenté d'esquisser, d'autres facteurs significatifs, quoique moins déterminants, entrent en ligne de compte.

L'exemple du chant diphonique soulignait ainsi la potentialité dégagée par une maîtrise assurée des cavités de résonance constituées, grossièrement, du pharynx, de la bouche et de la langue et, dans une mesure largement moindre, du nez. La mise à profit judicieuse des formants, puisqu'elle est susceptible de générer, pour un seul chanteur, deux lignes mélodiques distinctes, a de fortes chances d'influencer considérablement la qualité

[16] Jacques et Catherine OTT, *op.cit.*, p.119

vocale dans le chant « classique ». Avant d'étayer cette hypothèse au travers des techniques qui la mettent en œuvre dans la pratique vocale des chanteurs, nous devons préciser ici ses propres limites. Ceux-là même qui l'exploitent la circonscrivent en effet avec fermeté, condamnant par ailleurs les promoteurs d'une approche du travail vocal majoritairement centré sur les phénomènes de la résonance.

« D'emblée, disons-le, ceux qui s'attachent trop à la théorie systématique de la résonance vocale et aux résonateurs de tête font partie de ceux qui ignorent un mécanisme élaboré du souffle et des directivités vocales qui en découlent. Chez les tenants de la résonance, l'énergie sonore n'est pas reconnue en tant que phénomène nerveux lié aux modalités de la pression de l'air ; on y parle plutôt d'un son originel glottique atrophié, mais timbré et agrandi par des cavités corporelles. […] La résonance, en quelque sorte, accorde une couleur acceptable, une couleur d'emprunt à une voix qui n'est pas «placée » par ailleurs »[17] dénoncent ainsi Jacques et Catherine OTT. Ils renforcent d'ailleurs leur position en relatant les propos du professeur Guearti, pédagogue italien de Milan qui s'est livré à une expérience étonnante. « Cette expérience très désagréable faite sur bien des chanteurs éminents et sur moi-même, qui avons consenti à nous prêter à cette démonstration, a permis d'établir qu'aucune différence ni aucune diminution de vibration ne se remarque dans l'émission des sons et des phrases produits avant ou après l'injection d'eau distillée qui remplit les sinus. Moi-même n'ai ressenti aucune difficulté pour chanter, mais seulement une gêne produite simplement par cette ponction et par le liquide remplissant mes sinus »[18].

Les détracteurs de la théorie résonantielle exposent pourtant d'autres ressors de la résonance utiles aux chanteurs. A l'inverse des pédagogues qu'ils incriminent, ils situent cependant le phénomène et son intérêt non pas dans les sinus, le nez, les os de la tête… mais, comme c'est le cas pour le chant diphonique, dans la forme du tractus vocal : position du larynx, forme du pharynx et de la bouche, position des lèvres et de la langue. Et, à l'instar du chant diphonique, il ne s'agit pas de créer ni d'apporter une intensité uniformément renforcée aux sons émis au niveau de la glotte par les cordes vocales, mais bien de sélectionner et de mettre en valeurs certains composants sonores qui font la « couleur » d'une voix .

Dans cette optique, les lèvres doivent suivre un précepte simple : se faire oublier. Pour cela, la plupart des pédagogues conseillent de chanter « en souriant ». La bouche ainsi

---

[17] Jacques et Catherine OTT, *op.cit.*, p.201
[18] *Ibid*.

élargie, les lèvres sont fines et ramenées sur les dents. Cela doit éviter de grossir la voix, de la rendre sombre, car des lèvres en avant, en « entonnoir », influencent fortement le timbre en atténuant, notamment, les harmoniques les plus aiguës.

Le souci d'obtenir un son le plus régulier possible commande quant à lui de ne pas imposer au maxillaire inférieur des mouvements de grande amplitude. La bouche est ainsi en général à demi-ouverte, afin de ne pas provoquer de mouvements intempestifs du larynx. Celui-ci dépend en effet indirectement des déplacements de la mâchoire, et il est préférable de ne pas superposer ces effets aux positionnement choisis de manière autonome par le chanteur. « L'ouverture très modérée du maxillaire inférieur avec un étirement en largeur des commissures des lèvres s'accompagne toujours d'une élévation de l'os hyoïde et du larynx, seule attitude à donner aux cordes vocales leur fonctionnement complet de portée sonore »[19].

La langue pour sa part constitue pour le chant, comme l'écrivit Esope, la « meilleure et la pire des choses ». Si elle permet l'articulation et la coloration des voyelles, elle risque aussi, mal placée, de gêner considérablement la progression du son en provenance de la glotte : elle s'impose comme son obstacle principal.

En conséquence, la plupart des professeurs de chant recommandent de la maintenir haute et bombée, pour dégager au maximum le passage du larynx aux lèvres. « La langue, pointe en avant et légèrement arrondie dans sa forme générale, restera en relation plutôt étroite avec la voûte palatine. La langue épouse ainsi (sans se coller pour cela au palais osseux) l'arrondi du plafond buccal en formant un « biseau vocal ».[…] Ce « biseau vocal » ajoute, dans son exacte dimension, une finition sonore au timbre premier de la voix »[20].

La langue, organe complexe aux faisceaux musculaires nombreux et, par conséquent, à la mobilité quasi illimitée, se prête difficilement à la description exhaustive des liens existant entre ses configurations et les modulations du son qu'elle permet, notamment dans la formation des voyelles. Jacques et Catherine OTT, dans l'ouvrage cité ici, ne formulent pas par eux-mêmes d'indications plus précises qu'une position haute à observer préférentiellement dans la phonation propre au chant. Ils ne reprennent pas non plus les éventuelles recommandations préconisées par les diverses écoles de chant européennes et préfèrent en dernier lieu s'en remettre à la sensibilité individuelle des interprètes : « On nous permettra de na pas pousser plus avant l'analyse du rôle lingual dans

[19] Jacques et Catherine OTT, *op.cit.*, p.183
[20] *Ibid.*

ses modifications de formes utiles à la prononciation des voyelles : le vocaliste fera confiance à son oreille […] »[21]

Ce renvoi à l'audition comme guide finalement le plus sûr de l'éducation vocale des interprètes n'eût-il pas du être mentionné plus avant, lorsque, par exemple, nous abordions la notion de registre ? Car si le pédagogue s'appuie sur son expérience et les théories existantes pour initier un élève aux techniques vocales, celui-ci, au-delà des appréciations de son professeur, n'a d'autre recours que sa propre écoute pour juger de ses échecs comme de ses progrès.

Sans doute cette idée n'a-t-elle pas lieu d'être formalisée dans une école de chant. Dans une perspective d'éducation vocale, la musique s'impose comme le débouché naturel d'une formation réussie à la maîtrise de sa voix, autant qu'elle imprègne implicitement toute sa progression. De fait, lors de la formation des interprètes, l'écoute n'est qu'un médium implicite du travail : elle n'en est ni l'objet, ni la finalité. Cela reste vrai dans le contexte éminemment empirique des bergers Touvas, qui se transmettent la technique du chant diphonique de façon traditionnelle. Les plus jeunes observent les plus vieux, les écoutent, s'exercent, jusqu'à maîtriser eux aussi le contrôle de l'émission des harmoniques, sans s'appuyer sur une connaissance explicite du phénomène. Là aussi, l'écoute est fondamentale mais implicite. Elle est plus une sensibilité qu'une stratégie d'appropriation d'une technique vocale originale.

Ne pourrait-on pas imaginer, pourtant, appréhender le travail de la voix au travers d'une écoute méthodique, où la perception auditive serait auscultée dans un but performatif ? Tout comme la compréhension du fonctionnement de l'appareil phonatoire a permis d'étayer et, parfois, de susciter, les modalités de l'enseignement des techniques vocales, l'écoute servirait de support aux progrès espérés par, selon le cas, les interprètes lyriques, les comédiens, ou les gens souffrant de pathologies de l'appareil vocal.

Nous allons voir par la suite qu'un tel courant de recherche existe, et qu'il a donné lieu à nombre de résultats aussi étonnants qu'ils paraissent, a posteriori, basés sur des évidences empiriques.

[21] *Ibid.*

B.3. L'apport de la technologie : contrôler la voix par l'audition, l'audio-psycho-phonologie d'Alfred Tomatis

Le docteur Alfred Tomatis, oto-rhino-laryngologiste, a ouvert le champ d'une discipline nouvelle, dès 1947, baptisée audio-psycho-phonologie. Ce terme composite renvoie autant à une approche singulière du phénomène de la phonation qu'à une pratique thérapeutique éprouvée, mise à profit dans des établissements appelés « centres Tomatis ».

L'audio-psycho-phonologie se base sur la conviction de son inventeur que « lâcher un son, c'est d'abord l'auto-contrôler, puis élaborer un son ou un cri, c'est l'imaginer tel qu'on le voudrait, puis le jeter dans l'espace et l'écouter pour juger s'il répond bien à ce que nous pensions créer »[22]. Pour A. Tomatis, la voix ne dépend donc pas tant des capacités intrinsèques de l'appareil phonatoire d'un individu que de la qualité d'écoute qu'il est capable de mettre au service de son émission vocale. «Entendre et s'entendre. Ecouter et s'écouter. Telle est l'étape à laquelle nous [sommes] parvenus dans la structuration de notre conditionnement auditif du langage. Elle résume on ne peut mieux ce que signifie audio-phonologie. S'entendre parler en est la définition la plus rapide et la plus raccourcie »[23].

Cette approche du phénomène vocal se base autant sur l'observation empirique que sur la vérification expérimentale instrumentée. «Dans un premier temps, il nous est apparu qu'en général il n'y avait aucune correspondance réelle entre la description anatomique classique et les qualités vocales que l'on pouvait espérer chez un chanteur –tel gros larynx n'engendrait qu'une voix fluette tandis qu'un autre, parfois minuscule, était capable de fournir les premières impulsions d'une voix wagnérienne. Au surplus, combien de larynx, apparemment endommagés, émettaient des sons exceptionnels, tandis que d'autres, strictement normaux, voire même anatomiquement exceptionnels, étaient incapables, dans leur jeu acoustique, de se montrer susceptibles de créer un son de qualité. […] C'est vers l'oreille que nous nous sommes alors dirigés »[24]. Il est permis de mettre en doute les appréciations de qualité anatomique d'un larynx, comme celles évoquant un « son de qualité », émises par le docteur Tomatis pour justifier sa démarche. Une «correspondance réelle entre la description anatomique classique et les qualités vocales » laisse en effet une part importante au jugement subjectif de l'observateur qui, d'ailleurs, appuiera un tel jugement avant tout sur ses propres capacités auditives…

---

[22] A. Tomatis, *L'oreille et le langage*, Editions du Seuil, collection Le rayon de la science, Paris, 1963, p. 75
[23] A. Tomatis, *op. cit.* p. 87
[24] A. Tomatis, *op. cit.* p. 92

Mais ce qui nous préoccupe ici n'est pas de renvoyer l'audio-psycho-phonologie à sa propre mise en abîme. Son inventeur, d'ailleurs, en reconnaît certaines limites, et ne nie pas complètement l'intérêt des travaux menés par les pédagogues de la voix que nous évoquions auparavant dans le cas, notamment, de la formation des chanteurs. Il précise ainsi que « la voix ne reproduit que ce que l'oreille entend […] Si le sujet ne peut émettre que ce qu'il entend, il n'émet pas pour autant tout ce qu'il entend. Cette limitation tient compte des impossibilités de notre appareil phonatoire »[25].

Néanmoins, l'approche proposée par le docteur Tomatis s'ancre dans une description de la fonction auditive qui ne laisse aucun doute sur ses implications dans la phonation. Ce ne sont pas les étonnantes considérations sur l'ontogenèse de l'oreille avancées pour la soutenir qui retiennent notre attention, bien qu'elles soient assez séduisantes : « Somme toute, l'oreille moyenne dans sa totalité, si elle est un tout –et nous savons qu'elle est un tout fonctionnel- entraîne ipso facto une unité fonctionnelle, bouche-face, et mieux encore bouche-face-oreille »[26], écrit ainsi l'oto-rhino-laryngologiste. En revanche, aborder comme il le fait les interactions ente l'audition et la voix sous l'angle de la cybernétique, et, par là, ouvrir la voie à un traitement instrumenté de la phonation, nous semblent un terrain d'investigation particulièrement fertile.

La cybernétique, telle que définie par le Petit Robert, est la « science constituée par l'ensemble des théories relatives au contrôle, à la régulation et à la communication dans l'être vivant et la machine »[27]. Cette seule définition suppose déjà l'existence d'une analyse commune à l'homme et à l'instrument technique, dès lors qu'il s'agit de penser le contrôle. A. Tomatis pousse, lui, le parallèle jusqu'à proposer que l'étude des relations entre l'oreille et la voix fournit une introduction tout à fait pertinente à la science cybernétique. Mais, surtout, il met ainsi en image sa théorie audio-phonologique, tout en rendant transparente la démarche pragmatique de traitement qui en découle.

L'oreille est donc considérée avant tout comme un capteur sonique intervenant à la fois en amont et en aval de la phonation. Les sons émis au sortir de la bouche sont analysés par l'oreille, qui, en retour, permet d'ajuster volume, tonalité, intensité et articulation de leur production. « L'oreille devient donc l'organe majeur du contrôle de notre information dirigée vers l'extérieur, de notre geste vocal informationnel, de notre langage. L'oreille est comparable dès lors au régulateur d'un système qui apparaît asservi et sous sa dépendance.

---

[25] A. Tomatis, *op. cit.* p. 104
[26] A. Tomatis, *op. cit.* p. 59
[27] *Le Petit Robert*, dictionnaire de la langue française, édition 2001

Il y a plus exactement une interdépendance qui relève plus d'une association, d'une symbiose de deux fonctions, tributaires l'une de l'autre, indispensables l'une à l'autre »[28]. Sans entrer dans le détail des paramètres acoustiques du son vocal, A. Tomatis précise les modalités du contrôle exercé par l'oreille : « Le capteur auditif […] sait écouter […] dans les limites de certaines fréquences qui s'étalent de 16 à 20 000 hertz. Il peut, au surplus, dans ce large champ fréquentiel, se satisfaire de bandes passantes qu'il préfèrera électivement. Il y localisera sa sélectivité, son affinité. En cela, le capteur déterminera la qualité du débit et lui imposera la bande passante de son contrôle. A côté de cette régulation fréquentielle s'inscrit celle qui répond aux intensités. Parler plus ou moins fort n'est que la traduction d'un contrôle auditif, plus ou moins aiguisé, des intensités acoustiques. Enfin, il existe un dernier élément qu'introduit notre capteur sur son retour et qui s'ajoute aux précédents, dont l'ensemble réalise le « gain », cet autre élément est le temps. […] Cette mémoire de contrôle a pour charge d'agir sur l'élaboration de l'acte futur, sans que son intervention l'entrave en quoi que ce soit, bien au contraire. Elle doit agir de telle manière que son rôle oscille à tous moments entre celui de l'accélération et du frein aux contre-réactions judicieusement combinées »[29].

Il pourrait sembler, à ce stade, que l'inventeur de l'audio-psycho-phonologie ne fasse qu'inscrire des évidences (l'oreille, en plus de sa fonction sensorielle de saisie d'informations sonores en provenance du milieu extérieur, sert d'étalon pour jauger et ajuster notre propre émission vocale) dans le cadre théorique opportun de la cybernétique. Il n'en est rien, à partir du moment où l'expérimentation vient dépouiller des concepts de leur apparence triviale pour en extraire une stratégie effective de traitement de la phonation.

Le docteur Tomatis, au départ pour confirmer expérimentalement sa théorie puis, par la suite, pour développer les outils de son exploitation commerciale, mit rapidement au point un appareillage relativement simple. Il s'agissait – et, mis à part les perfectionnements rendus possibles par les progrès de l'électronique, le principe reste inchangé aujourd'hui – d'introduire, dans la chaîne du signal sonore quittant la bouche pour parvenir aux oreilles, une série de filtres électroniques susceptibles d'amplifier, d'atténuer, voire de supprimer, tout ou partie des composantes fréquentielles du son vocal. Pour cela, la voix est captée par un microphone, traitée par les filtres dont on ajuste les paramètres de traitement à volonté, puis restituée aux oreilles par l'intermédiaire d'un casque.

---

[28] A. Tomatis, *L'oreille et le langage*, Editions du Seuil, collection Le rayon de la science, Paris, 1963, p. 88
[29] *Ibid*.

Cet équipement, dont les composants doivent répondre à des critères de qualité élevés pour garantir que le son ne sera ni bruité ni distordu de façon intempestive, cristallise ainsi le rôle de contrôleur de l'oreille en autorisant toutes les modifications voulues sur les caractéristiques de traitement « naturelles » d'un individu donné.

De nombreux chanteurs ont accepté, au début des travaux d'A. Tomatis, d'exercer leurs talents devant le microphone, en se couvrant au préalable les oreilles du casque d'où sortait leur voix modifiée. Un auditeur placé dans le laboratoire où se tenait l'expérience, ou les expérimentés eux-mêmes à qui l'ont proposait après-coup d'écouter un enregistrement de leur voix naturelle pendant l'expérience, pouvait observer les résultats rapportés par l'oto-rhino-laryngologiste :

« Muni de [l'] ensemble expérimental, on constate sans exception que :

1° Si la bande au-dessus de 2 000 hertz est tronquée, la voix devient terne, dénuée de la richesse harmonique de départ, plus frêle, plus postérieure et blanche, surtout en montant ; la justesse est conservée. Seule apparemment la qualité change.

2° Si la bande comprise entre 1 000 et 2 000 hertz est seule éliminée, tout le reste étant respecté, la voix conserve sa qualité antérieure, sa richesse de départ, mais le contrôle de la hauteur tonale a disparu. La reproduction juste est impossible.

3° Si la zone limitée entre 500 et 1 000 hertz se trouve à son tour modifiée, c'est alors l'expression de la justesse globale qui est altérée. L'expérimenté est soudain incapable de juger de la justesse de toute musique qui est exécutée à ses côtés. En même temps son affinité musicale s'émousse.

4° Si toute la courbe est altérée dans la zone comprise entre 500 et 2 000 hz, on aboutit à une amusicalité »[30].

On imagine sans peine l'étonnement, puis le malaise, d'un chanteur lyrique professionnel renommé confronté à un telle expérience. Des années de travail, de répétitions, de vocalises, à parfaire un art où un timbre riche et une grande justesse de ton peuvent construire une carrière, s'effondrent en quelques minutes si l'oreille n'entend plus comme elle le fait d'ordinaire.

L'appareillage mis au point par le docteur Tomatis n'a heureusement pas pour intérêt premier de faire prendre conscience aux grands interprètes de la fragilité de leur métier. Il contribue, a contrario, à influencer l'audition d'individus dont l'oreille ne remplit pas correctement son rôle, entraînant des dysfonctionnements de la phonation. En revanche,

[30] A. Tomatis, *op. cit.* p.99

si l'expérience décrite ci-avant semble aisément reproductible, puisqu'il s'agit d'imposer un handicap virtuel à une oreille opérant normalement, il en va bien différemment lorsqu'on se donne le but inverse. En effet, il s'agit alors de réunir un ensemble de filtres qui vont définir un « profil d'audition » correct, apte à entraîner une phonation de qualité.

Confronté à ce nouveau problème, alors qu'il venait de mettre en évidence expérimentalement le contrôle actif de l'oreille sur la phonation, le docteur Tomatis se basa à nouveau sur une approche empirique, menée en collaboration avec des chanteurs professionnels, pour cerner les modalités de la construction de ce que nous serions tentés d'appeler « une machine à écouter correctement ».

« […] Nous construisions des hypothèses sur les manières d'entendre de Caruso, de Tita Rufo, Benjamino Gigli et bien d'autres, et nous tentions de réaliser des montages nous permettant d'obtenir des écoutes, lors de l'émission, comparables à celles que devaient avoir nos chanteurs si exceptionnels. Ce travail de longue haleine nous permit d'obtenir des ensembles électroniques capables de recréer à volonté des modes d'auto-contrôle identiques à ceux de nos sujets choisis comme tests expérimentaux. […] Dès l'instant où un sujet bénéficie d'une telle modification de son auto-contrôle, autrement dit du gain de son capteur, son émission change ; elle s'enrichit électivement dans les mêmes plages qui lui sont livrées auditivement. Son timbre s'allume, il devient identique sur le [sonogramme] au modèle choisi. […] Ainsi, par exemple, l'imposition du contrôle du type Gigli entraîne une phonation très antérieure, au niveau des lèvres, automatiquement en *mezza voce* ; les lèvres s'allongent comme une moue qui s'avance ; le nez se pince à la base, la tête se défléchit légèrement, la respiration devient profonde, abdominale, l'écoulement du souffle se régule, se ralentit très sensiblement. La manière de chanter peut, par un mécanisme profond, devenir celle de Gigli. Certes, ce n'est pas tout, la qualité reste propre à chaque individu, et il en est des chanteurs comme des instruments, les uns sont des stradivarius et d'autres des violons de moindre qualité. Mais le mode d'excitation de l'ensemble pneumo-phonologique –et c'est en-cela que le phénomène est intéressant- s'oriente vers une identité d'action chez tous les sujets soumis à l'expérimentation, qui rappelle en tout point les procédés qu'utiliseraient les grands techniciens »[31].

Le procédé d'A. Tomatis consiste donc à élaborer de façon empirique des modèles d'écoute destinés à remplacer ceux que les sujets souffrant de pathologies de la phonation ou souhaitant améliorer leurs capacités vocales utilisent d'ordinaire. Nous ignorons la

[31] A. Tomatis, *op. cit.* p. 107

manière dont ces modèles sont construits précisément, et sans doute est-ce un secret bien gardé puisque l' « oreille électronique » du docteur Tomatis fait l'objet d'un brevet. Peut-être une solution consiste-t-elle à soumettre un « sujet-modèle » à sa propre voix modifiée par une série de filtres paramétrés aléatoirement. A priori, le rendu vocal d'un tel traitement conduit à une voix différente de la voix naturelle du sujet. Reste ensuite à ajuster les paramètres –c'est peut-être ici que se situe toute la difficulté d'une telle démarche empirique menée « à tâtons »- jusqu'à ce que la voix modifiée coïncide avec la voix naturelle, révélant alors les réglages des filtres spécifiques à l'écoute naturelle du sujet.

Nous ne saurions spéculer plus avant sur le détail des procédés techniques en jeu dans la méthode d'A. Tomatis. En revanche, il est important de considérer le fait que l'éducation vocale basée sur l'audition pratiquée dans les « centres Tomatis » ne se pratique que sur des temps relativement longs. Le docteur Tomatis parle ainsi de « conditionnement audio-vocal ». En effet, si un individu voit ses capacités phonatoires profondément augmentées à partir du moment où son écoute est modifiée, le phénomène cesse dès l'instant où il retire le casque substituant à ses propres oreilles « l'oreille électronique » et ses filtres fréquentiels.

Pour asseoir les bénéfices du contrôle auditif électroniquement assisté de la phonation, il est donc nécessaire de procéder à une démarche relevant de l'entraînement, où le sujet apprend peu à peu à aligner, sans renforts artificiels, son écoute naturelle sur celle, programmée, qui lui est proposée.

Ainsi, dans les « centres Tomatis », des programmes sont-ils définis en fonction de l'âge des individus et des motifs qui les poussent à vouloir agir sur leur écoute, après qu'un bilan auditif a permis de mieux connaître leurs capacités. A titre indicatif, le schéma type d'un traitement se compose d'une première série de séances de deux heures par jour pendant quinze jours consécutifs. Suit une période de « pause » de trois à six semaines, complétée pour finir par une reprise des séances pendant huit jours environ. Lors des séances, les sujets sont soumis à diverses stimulations auditives passives, ou bien sont invités à reformuler vocalement ce qu'ils ont entendu dans un casque relié à une « oreille électronique » réglée en fonction de leur propre profil.

Les modalités des séances dépendent fortement des capacités initiales du sujet, de son vécu et de ses attentes. En dehors du traitement de la qualité vocale, sur lequel nous avons focalisé notre attention jusqu'ici, la théorie audio-psycho-phonologique d'A. Tomatis prétend aussi donner des moyens d'action sur un certain nombre d'autres capacités individuelles. Elle distingue en effet, dans le pouvoir de contrôle qu'exerce l'oreille sur la phonation, l'oreille gauche et l'oreille droite. Cette dernière est reconnue comme oreille

directrice, et ses relations avec les circuits nerveux du cerveau droit ont amené le docteur Tomatis à lui conférer une capacité d'incidence sur la qualité d'élocution –il a ainsi réalisé de nombreuses expériences où l' « oreille électronique » améliorait très sensiblement, voire soignait totalement, le bégaiement ou la dyslexie, y compris la dyslexie dans la lecture et l'écriture- voire même sur le comportement général des individus.

C. Synthèse artificielle de la voix

C.1. De la machine à parler de Von Kempelen aux synthétiseurs informatiques

Avec l' « oreille électronique » d'A. Tomatis, nous sommes passés d'une approche purement organique de la voix à une démarche instrumentée, où le recours à un système artificiel appuie, conditionne même, les transformations apportées au phénomène de la phonation. Là où la seule relation de maître à élève, agrémentée des connaissances relatives aux mécanismes physiologiques présidant à l'émission vocale, permettait à un individu, au sein des écoles de chant en particulier, de perfectionner ses capacités, l'électronique propose une véritable alternative pédagogique où le sujet progresse de façon quasi-autonome. Les corrections du professeur se sont muées en auto-ajustements, permis par une machine compensant les défauts propres à l'élève.

Un tel système, en dépit des preuves expérimentales de ses performances, n'aurait sans doute pas pu voir le jour indépendamment de la construction de la théorie audio-psycho-phonologique, puisqu'il qu'il n'est rien d'autre que son prolongement matériel. En cela, il s'inscrit dans un cadre de pensée que son auteur, pour l'avoir affiné, ne cherche pas à généraliser : l'audition en est le centre et le moteur, et l'acte phonatoire ne saurait s'en émanciper.

Vouloir pénétrer les mystères de la production sonore vocale par le biais des machines et, ainsi, se donner éventuellement les moyens de contrôler l'acte phonatoire, relève cependant d'une quête largement antérieure aux seules prémisses de l'audio-psycho-phonologie.

C'est au milieu du XVIII^{ème} siècle que le Hongrois Wolfgang Von Kempelen, présent à la cour de l'impératrice Marie-Thérèse à Vienne, mit au point l'une des premières « machines à parler ». Il s'agissait moins de divertir un public courtois amateur de curiosités savantes que d'essayer, grâce à cet objet, de comprendre comment l'être humain produisait les sons si complexes dont il alimentait une de ses singularités devant l'animal : le langage.

Von Kempelen, jaugé à l'aune de nos acquis théoriques contemporains, est décrit comme le pionnier de la phonétique expérimentale. Cela souligne le caractère profondément pragmatique de ses recherches : il voulait décortiquer le son, et pensait qu'en le reproduisant par l'artifice il s'octroierait la capacité d'en saisir les processus de

production de façon suffisamment sûre pour en déduire d'éventuelles applications thérapeutiques.

Aujourd'hui, il semble bien illusoire que la machine de Von Kempelen ait pu suffire à imaginer des traitements aptes à soigner bègues et dyslexiques, comme l'autorise dans certains cas l' « oreille électronique » de l'oto-rhino-laryngologiste Tomatis.

Ingénieuse, la « machine à parler » demeure en effet assez rudimentaire. Dans son ensemble, elle reproduit les diverses composantes des organes de l'appareil phonatoire. Un soufflet en cuir mime les poumons, une anche en ivoire reproduit une corde vocale rudimentaire, alors qu'un cornet rigide figure le volume buccal.

La production des sons est largement liée au savoir-faire de l'expérimentateur car, outre les trois éléments principaux qui la composent, la « machine à parler » est équipée de divers leviers et sifflets actionnés manuellement. Ils permettent de modifier la façon dont l'air expulsé du soufflet vient frapper la anche, en jouant sur les volumes qu'il traverse ou en introduisant des obstacles dans son parcours, de la même façon que la langue ou la morphologie du pharynx altèrent la production naturelle du son formé au niveau du larynx. Le cornet qui tient lieu de bouche n'étant en lui-même pas déformable, la mobilité des lèvres est remplacée par la main de l'expérimentateur, qui, suivant la façon dont elle vient boucher plus ou moins l'orifice final de la machine, permet de produire des sons différents. D'après Von Kempelen, trois semaines d'entraînement suffisaient pour réussir à faire émettre des sons suffisamment distincts les uns des autres pour les assimiler à certaines voyelles du langage naturel. Les consonnes exigent quant à elles autant la performance de la personne qui manie la machine que l'indulgence de l'auditeur, et ne sont d'ailleurs pas toutes reproductibles, même à tolérer leur large imperfection. Ainsi les d, t, g et k ne se plient-elles pas à leur imitation artificielle, sauf à espérer compensation par une écoute à l'imagination trop grande pour servir de caution à une expérience qui se veut rationnelle.

Il serait tentant de penser que l'originalité de la « machine à parler » ne fut jamais vouée qu'à stimuler une curiosité vive mais finalement peu déterminante pour la compréhension du phénomène phonatoire, même si son inventeur la développa dans le cadre de recherches aux ambitions plus larges que la représentation spectaculaire.

Imparfaite, peu fidèle à ce qu'elle était supposée imiter, elle a cependant ancré la reproduction synthétique de la voix dans une réalité technologique qui a, depuis, largement prospéré. La réalisation purement mécanique du geste vocal s'est ainsi améliorée, les inventeurs successifs complexifiant chaque nouvel appareil et profitant des progrès enregistrés dans le domaine des matériaux ou de l'automatique.

Aujourd'hui, certains laboratoires perpétuent cette conception mécaniste de la synthèse vocale artificielle. Au Japon, Hideyuki Sawada, chercheur à l'université Kagawa, a ainsi fabriqué ce qui n'est rien d'autre que la version moderne de la machine de Von Kempelen. Un compresseur à air alimente un équivalent de tractus vocal en matériaux polymères souples, dont la géométrie est contrôlée par une série de pistons pneumatiques, eux-mêmes coordonnés par un « réseau de neurones » électronique. La anche en ivoire a été remplacée par une «corde vocale » en caoutchouc. Si la dextérité et l'entraînement d'un expérimentateur ne sont plus nécessaires, le système dans son ensemble étant piloté par un programme informatique, les performances ont, elles, peu progressé en regard des solutions adoptées au XVIII^{ème} siècle. Le « robot vocal » reste incapable d'articuler des locutions plus complexes, sur le plan sonore, qu'un simple « hello », déjà assez peu convaincant.

Ces travaux n'en sont qu'à leur balbutiement. Il n'est pas dit que, d'ici quelques années, ils soient à même de produire des résultats honorables. Cela dit, la motivation qui les sous-tend paraît tenir plus de la quête de l'exploit technologique –l'attrait nippon pour la robotisation y prenant sans doute une part non négligeable- que de la mise au point de modèles artificiels susceptibles, dans leur imitation de la production naturelle de la voix humaine, de mieux la faire comprendre.

A l'inverse, c'est sans doute plus dans une réelle optique de maîtrise globale de la production physique du langage –fût-il chanté- que se développe une approche désormais dominante : celle de la synthèse artificielle de la voix par l'outil informatique.

La voix, si on la considère du point de vue purement acoustique, est avant tout une onde et c'est sa nature vibratoire qui a permis de la capter sous la forme d'un signal électrique. Alors que les sonogrammes évoqués à propos de l'étude du chant diphonique traduisaient presque directement la nature composite de ce signal –décomposition en fréquences et intensités dans le temps-, et là où l' « oreille électronique » du docteur Tomatis altérait ce même signal au moyen de filtres fréquentiels, l'informatique permet un traitement autrement plus complet du son vocal. Il devient ainsi possible de recréer une « voix » à partir d'un grand nombre d'échantillons numériques sonores, au lieu de mimer physiquement sa production.

C.2. Synthèse vocale à partir du texte

C.2.1. Des synthétiseurs à diphones aux synthétiseurs à formants

Avant de détailler les stratégies de recherche engagées dans la synthèse vocale, il est important de souligner que la dynamique qui l'imprègne fait se rejoindre en un tout à la fois l'importance cruciale de l'audition chère à A. Tomatis et celle non moins déterminante des modalités pratiques de la production de la voix. Clairement, le potentiel de création offert par l'informatique ne peut être véritablement mis à profit que si ceux qui l'emploient parviennent à extraire des paramètres pertinents des voix de synthèse produites, afin d'estimer en retour les façons les plus efficientes de modifier leurs modèles initiaux. Cela n'a rien de trivial…

Précisons ici que nous nous concentrerons essentiellement sur la synthèse vocale à partir du texte. Celle-ci propose qu'un logiciel auquel on soumet tout texte dactylographié dans un langage naturel, et dans la limite des langues auxquelles le logiciel en question est adapté, doit être capable de le faire entendre «à haute voix », par l'intermédiaire de hauts parleurs connectés à l'ordinateur utilisé pour l'opération. « Text-to-speech synthesis is simply a simulation of the human process of reading a text out aloud »[32].

Cette définition, bien que partielle, nous semble la meilleure pour évoquer la synthèse de la voix « à proprement parler ». Elle met en effet de côté les innombrables procédés de traitement numériques appliqués «après coup » à des extraits enregistrés de voix humaines naturelles. Pour préciser cette distinction à travers une comparaison dans le domaine de l'image, de tels procédés relèveraient de la retouche numérique de photographies existantes, alors que la synthèse à partir du texte pourrait s'apparenter à la création *in abstracto* d'images de synthèses, qui ne sont définies à l'écran que par les équations leur imposant formes et couleurs.

Il ne s'agit cependant pas d'une synthèse vocale *ex nihilo.* Pour construire une voix artificielle, il est évident que certaines briques de base sont indispensables. En l'occurrence, il s'agit de briques sonores. Soit les sons qui vont servir à fabriquer une voix artificielle seront extraits d'une voix humaine réelle, pour être réorganisés en un flux de parole qui n'obéit dans son agencement qu'aux lois de la machine qui le produit, soit ils seront élaborés à partir de générateurs de signaux et de filtres, donc, effectivement, totalement

[32] Christopher Baber and Janet M. Noyes, *Interactive speech technology. Human Factors issues in the application of speech Input/Output to computers*, Taylors & Francis Ltd, 1993, p. 26

artificiels, mais structurés selon les observations menées sur des enregistrements de voix naturelle, pour leur garantir un rendu auditif qui se rapproche le plus possible d'une voix humaine réelle.

Dans ce qui retient notre attention, le matériau déterminant de la synthèse vocale s'avère quoiqu'il en soit n'être autre que le texte. Par conséquent, prétendre faire jaillir le son des machines à la façon dont les humains lisent un texte… présuppose de connaître la manière dont ceux-ci eux-mêmes y parviennent. Dès l'antiquité, le langage et la parole furent l'objet de réflexions poussées, d'études nombreuses. De nos jours, phonéticiens, linguistes, sémiologues contribuent continuellement à l'élaboration de théories complexes, confinant au débat philosophique dès lors qu'il s'agit de démêler ce qui préside à l'émergence de la pensée de ce qui conditionne l'acquisition du langage, quand il ne s'agit pas de s'interroger sur le sens même d'une telle quête…

A priori, les ressources théoriques sur lesquelles baser une méthode de production artificielle de la voix, s'appuyant sur le traitement informatique d'un texte et sa concrétisation sonore, devraient se compter en nombre suffisant pour garantir le succès d'une telle opération. Oui et non. Oui, si l'on s'en tient aux faits. Car aujourd'hui les systèmes de synthèse vocale fonctionnent d'une manière tout à fait honorable même si, nous le verrons, ils sont confrontés à des limites qui tendraient à ne pas se laisser dépasser. Non, car les ressources théoriques nécessaires sont à la fois nombreuses et contradictoires, et se prêtent difficilement à leur mise en œuvre pratique au sein de systèmes informatiques : « The early researchers in the 1960s relied heavily on what linguists and psychologists had to say about human speech production, since it is not possible to engineer anything successfully without having a good deal of reliable theory on which to build. It seemed self-evident that the specialists in language and speech would be in a position to supply some reliable information with which to underpin the simulation attempt –after all, linguistics had been around for some four thousand years. The linguists were happy to supply the information, though some of them were bemused that anyone would want to get a machine to read a text. Remember, this was early on in the development of computers. Ironically it was the fact that the information about the nature of language and how people speak was so readily available which ultimately held up the development of speech technology for such a long period of time. The information offered to those developing the simulation was wrong. Nobody at the time realized that the metatheorical position adopted by linguists and psychologists was quite unsuited to the task in hand –the simulation of what a human being is doing. […] An obvious question to be answered when setting out to design any speech synthesis system is: what are the basic building blocks of speech ? We need to know the

size and nature of the units which are to be assembled to create the speech output. The answer seems self-evident. They are the sounds, which we can all make, which are strung together to form the words we speak: there are three of them in a world like *cat* and four of them in a world like *text* –or should that be three, or perhaps five ?"[33].

Le problème principal de la synthèse vocale informatique à partir du texte était donc, à son origine, de choisir la façon dont les mots devaient être découpés en unités sonores suffisantes pour générer, une fois répertoriés et après avoir attribué une concrétisation sonore artificielle à chacun, un flot sonore s'apparentant de façon satisfaisante à la lecture humaine à voix haute.

Un premier choix se porta sur les phonèmes, comme l'évoque l'auteur cité ci-avant avec les mots «cat » et «text ». Les phonèmes forment «la plus petite unité du langage parlé, dont la fonction est de constituer les signifiants et de les distinguer entre eux »[34]. On distingue les phonèmes vocaliques, consonantiques, oraux, nasaux, sourds, sonores… Le français en comprend 36, répartis en 16 voyelles et 20 consonnes, qui ne correspondent pas toujours, de fait, aux voyelles et consonnes de l'alphabet.

Les premiers synthétiseurs vocaux, dispositifs purement électroniques développés dès les années 1960, donc avant l'avènement de l'informatique, utilisaient les phonèmes selon un procédé finalement assez proche de la machine de Von Kempelen. En effet, ils faisaient appel en premier lieu à une source vocale constituée d'enregistrements distincts de tous les phonèmes de la langue dans laquelle on souhaitait travailler, équivalente de l'ensemble soufflet-anche en ivoire. Ces phonèmes étaient alors soumis à un ensemble de filtres électroniques fréquentiels régulés par un système de règles, à l'instar du son émis par la anche d'ivoire altéré par les positions des sifflets et de la main de l'expérimentateur devant le cornet faisant office de bouche.

A la différence de la « machine à parler », ces synthétiseurs produisaient une imitation de parole continue, limitée par la seule longueur du texte soumis. Les règles déterminant les paramètres des filtres utilisés reproduisaient en fait les configurations des formants vocaux (cf *infra*.) observés dans la voix naturelle, de façon à lier les phonèmes entre eux de la façon la plus fidèle possible à la lecture humaine à voix haute. Pour le son « a », par exemple, un premier formant est observé à 800 hertz, un second à 1 200 hertz, un troisième à 3 000 hertz, etc. Le passage du phonème «a » au phonème «beu », comme dans le mot « abri » par exemple, sera assuré par une règle indiquant jusqu'à quelle

---

[33] *Ibid*.
[34] *Le Petit Robert*, dictionnaire de la langue française, édition 2001

fréquence va augmenter le premier formant, pendant que le second devra diminuer jusqu'à une autre fréquence donnée, etc.

Ces synthétiseurs, en usage jusque dans les années 80, offraient certains avantages : d'utilisation souple, ils requéraient des traitements informatiques peu complexes, et pouvaient s'appliquer facilement d'une langue à l'autre. Leur rendu s'avérait en revanche assez médiocre, et les voix produites ne trompaient personne : c'était bel et bien la machine qui parlait.

Actuellement, une autre stratégie régit la majorité des synthétiseurs vocaux en usage. Ils reposent non pas sur la succession ordonnée par des règles de phonèmes, mais sur le recours aux «diphones ». A la différence des phonèmes, les unités sonores de base appelées « diphones » sont extraites de la langue pour, prises toutes ensemble, refléter toutes les possibilités d'enchaînement d'un son à l'autre dans la langue considérée. Leur utilisation part du constat qu'un même son n'aura pas le même profil acoustique si on le fait suivre par des sons consécutifs différents. Un «a » suivi d'un « gueu », observé sur un sonogramme par exemple, n'aura pas la même « signature acoustique» que s'il est suivi d'un « keu ». Ou bien, typiquement, « ch » plus « a » n'est pas égal à « chat ».

La langue française peut ainsi être caractérisée de façon exhaustive par, environ, un millier de diphones. Contrairement aux synthétiseurs « à formants », les synthétiseurs «à diphones » ne requièrent donc pas de règles pour assembler les unités phonétiques de base sur lesquels ils reposent, celles-ci impliquant leurs propres enchaînements.

La mise au point de ce type de synthétiseurs repose sur l'enregistrement de plusieurs dizaines de minutes de lecture à voix haute par des volontaires bien vivants : il est indispensable de couvrir toutes les combinaisons sonores possibles, de façon à pouvoir traiter n'importe quel cas de figure dans les textes soumis à un synthétiseur. L'interrogation de tels corpus de diphones, très volumineux, ne saurait se faire à une vitesse trop lente, préjudiciable à une vitesse de lecture artificielle acceptable par son auditeur.

Si leur rendu est largement meilleur que leurs prédécesseurs historiques, les synthétiseurs à diphones nécessitent des moyens informatiques relativement exigeants. L'accès impérativement rapide aux bases de données répertoriant l'ensemble des sons propres à une langue donnée n'a été permis que par l'abaissement exponentiel des coûts des composants de stockage numérique de l'information et leurs progrès en terme de rapidité d'exécution des opérations.

Performants, bien maîtrisés, les modèles les plus récents de synthétiseurs, développés dans les laboratoires de recherche publics aussi bien que dans les départements de recherche et développement de nombreuses entreprises de télécommunication, ont pris le

chemin de leur démocratisation. Systèmes d'information embarqués dans les véhicules, consultation vocale de bases de données, services en ligne automatiques pour les réseaux de téléphonie mobiles ou fixes… Les applications se multiplient, et devraient profiter du développement de l'informatique personnelle auprès du grand public.

Dans l'autoproclamée « ère de la communication », la synthèse vocale semble assurée d'un bel avenir. Néanmoins, si elle tend à se départir peu à peu de la confidentialité propre à toute technologie nouvellement adoptée par les grands systèmes commerciaux, elle pêche encore par une faiblesse qui suscite une grande part des efforts de recherche actuelle dans le domaine.

C.2.2. Une faiblesse congénitale : l'inhumaine voix de la machine

Nous avons vu que la synthèse par diphones autorisait le traitement de n'importe quel texte de façon intelligible. Les comportements des utilisateurs confrontés à cette technologie montrent cependant que le critère d'intelligibilité, pour être éminemment nécessaire, n'entre pas seul en compte dans le succès ou l'échec de la banalisation des voix artificielles. A cet égard, le cas des ordinateurs de bord embarqués dans les voitures particulières est significatif. Ces systèmes sont équipés d'un synthétiseur vocal qui avertit les conducteurs de l'état du véhicule : consommation ,nécessité de remplir le réservoir, ceinture non bouclée, signalisation lumineuse en marche, porte mal fermée, coffre non verrouillé… Autant d'avertissements censés garantir la sécurité ou éviter d'éventuels désagréments prévisibles. S'il est possible que la fréquence à laquelle se manifeste le système de contrôle soit mal adaptée aux nécessités jugées par le conducteur lui-même, ou si des dysfonctionnements conduisent à des informations erronées, il est clair que le caractère « métallique », « irréel » de la voix artificielle véhiculant tous les messages s'avère bien souvent être une des motivations premières de sa mise hors circuit.

La lenteur mécanique du débit, la pauvreté du timbre, le caractère égrillard, l'absence de respiration… et surtout, l'impression généralement inhumaine qui se dégagent de nombre de voix de synthèse « courantes » tendent à les disqualifier rapidement dans l'esprit de la personne qui les entend. Les meilleurs synthétiseurs ne « sonnent » pas comme une voix humaine réelle. En particulier, le fait que chaque diphone soit reproduit par la machine à l'identique dès lors qu'il est entouré des mêmes prédécesseurs et successeurs, au fil d'un texte lu artificiellement, conduit immanquablement un auditeur à reconnaître le caractère inhumain de ce qu'il entend.

Les instruments de musique artificiels actuels ont eux, en revanche, fini par faire leur preuve. Nombreux sont ceux qui, pour se laisser différencier de leurs modèles

originaux, devront être soumis à des oreilles mélomanes expertes et particulièrement exercées. La voix, de son côté, résiste. Les axes de recherche ouverts actuellement pour pallier cette déficience balbutient pour la plupart, ou se heurtent encore à l'hétérogénéité des approches empiriques. Il n'est pas sûr cependant que le temps seul suffise à faire évoluer cet état de fait, en dépit des découvertes théoriques et des progrès technologiques à venir. Avant d'étayer cette hypothèse, il importe de se pencher plus avant sur les pistes ouvertes qui doivent mener à rapprocher la voix de synthèse de son modèle naturel.

Dans ce véritable défi, l'analyse la plus exhaustive de la voix humaine est une étape incontournable. L'artifice, pour être fidèle, se doit de comprendre au mieux ce qu'il tente de reproduire. Mais la voix ne se plie que très difficilement aux tentatives de description qui l'appréhendent. Elle est mouvante, tout d'abord. D'un individu à l'autre et, ce qui est plus délicat encore à cerner, pour un même individu, elle va changer de coloration, d'intensité, de timbre, au gré de l'état psychologique du locuteur, de sa fatigue, des personnes à qui il l'adresse. Elle est complexe, ensuite. Elle n'a rien d'un son pur unique, tel le LA que l'on entend lorsqu'on décroche le téléphone, mais naît d'un agrégat continuellement fluctuant d'harmoniques de sons purs, de bruits d'écoulement, de souffles, de claquements. Elle dit beaucoup plus que les mots qu'on lui fait prononcer, enfin. Voilà qui complique singulièrement la déjà difficile adéquation aux textes soumis aux synthétiseurs qui la miment.

La tâche essentielle dans ce passionnant enfer de chercheur consiste à extraire des paramètres physiques significatifs à partir d'enregistrements de parole naturelle. Les diphones en font partie, mais ils ne constituent que la matière brute, qui va être modulée en débit, intensité, tonalité, agrémentée de pauses, et colorée par un timbre susceptible d'évoluer.

On appelle prosodie l'ensemble des caractéristiques de ce que l'on pourrait appeler la « musique vocale », timbre excepté. La prosodie n'est, de fait, pas indépendante de la structure grammaticale d'un texte. Par exemple, pauses et variations d'intensité sont liées à la ponctuation. En français, par exemple, le ton d'une phrase va descendre à la fin d'une phrase affirmative et monter à la fin d'une interrogative. Néanmoins, la liberté et la personnalité du locuteur l'influencent considérablement, et l'établissement de règles prosodiques, en usage dans les synthétiseurs à diphones, doit donc s'enrichir de paramètres prescriptifs qui vont au-delà de la seule syntaxe littérale.

Par ailleurs, le sens porté par les phrases d'un texte, auquel vient se rajouter l'interprétation qu'en fait un locuteur particulier, modifient considérablement une hypothétique prosodie « neutre », collant à un non moins hypothétique sens littéral du texte.

« If we know roughly what a sentence means and what its grammar is, and if we also know how the individual words are pronounced in isolation, we are in a position to look at the sentence as a whole and work out its prosodics. […] This means establishing a rhythm for speaking the text aloud and working out a suitable intonation. The intonation will determine at what points the voice goes up and down in pitch. For example if we take the sentence *John went home* we can pronounce this with a pitch that starts as an average level and progressively falls, indicating that we are making a simple statement of fact. But if the sentence is pronounced with the voice progressively rising in pitch throughout then we know that we are asking the question *John went home ?* –the changing pitch altering the meaning rather than any rearrangement of the words themselves. The situation is not all that simple –falling intonation for statements of fact and rising intonation for questions- because often questions have the falling pitch contour, as in *why did John go home* ? With sentences of more than just three or four words the prosodics model is however extremely complicated. Human beings are constantly changing the rhythm and intonation even during the course of a sentence to emphasize the important words or sometimes just to add variety"[35].

La notion de variabilité résume assez bien ce qui sépare encore la voix en sortie de synthétiseur de sa réalisation humaine. Ce qui caractérise une élocution naturelle, de façon générale, n'est autre que sa perpétuelle évolution, sa liberté mouvante, qui la fait prendre des distances plus ou moins considérables relativement à la seule syntaxe d'un texte. Ce constat ne saurait cependant constituer autre chose qu'une amorce de solution. Certaines tentatives techniques s'en satisfirent pourtant, en introduisant des variations aléatoires artificielles dans le débit ou l'intonation usuelle d'un synthétiseur, de façon à créer une illusion de spontanéité. A l'écoute, le résultat fut médiocre. Au bout de quelques secondes, de tels systèmes révèlent immanquablement leur caractère artificiel : ce qui change dans la parole humaine, pour être extrêmement ardu à retranscrire, n'en est pas moins dénué de sens. Le recours à l'aléatoire est donc une impasse. Il reste alors à ancrer la variabilité naturelle de la parole dans des réalités physiques mesurables et reproductibles, condition indispensable à sa reproduction dans un système de synthèse. Et c'est bien là que réside toute la difficulté des travaux de recherche actuels.

Le problème est double, et chacune de ses deux dimensions est inextricablement liée à l'autre. Il faut en effet à la fois déterminer les facteurs qui provoquent la variabilité de la

[35] Christopher Baber and Janet M. Noyes, *op.cit.*, p.29

phonation, et réussir, dans le même temps, à transcrire ces facteurs en des grandeurs physiques distinctes et reproductibles. Or les variations observées dans la parole naturelle tiennent essentiellement à des événements intangibles, tels que l'humeur du locuteur, l'objectif qu'il poursuit en parlant, ou les réactions qu'il exprime, au travers des caractéristiques de son propre débit vocal, en réponse à ce qu'il entend lorsqu'il est dans une situation de dialogue. Comment, dans ce cas, parvenir à fixer de l'extérieur, ce qui semble ne se dérouler que dans le cerveau d'un locuteur au seul moment où il s'exprime ?

C.2.3. De la synthèse « littérale » à la synthèse « expressive » ?

Si on est encore loin d'une voix de synthèse capable de créer une illusion persistante de vie, au point de se laisser confondre avec celle d'un locuteur humain véritable, des solutions permettent de s'en rapprocher. Et, puisque, comme le disait Buffon, « le style est l'Homme », toutes passent par l'étude minutieuse de très volumineux corpus de voix naturelles enregistrées. En multipliant les locuteurs et les conditions de production de la voix, les chercheurs tentent de se doter d'un maximum de cas significatifs. Observer des tendances, discerner des particularités, expliciter des comportements spécifiques, et les traduire en autant de paramètres applicables à un synthétiseur, telle est leur tâche.

Il y a une dizaine d'années, les traits proprement humains de la production de parole furent abordés au travers de deux de ses propriétés spécifiques, dont l'une, que nous venons d'évoquer, n'est autre que la variabilité et l'autre un « effet pragmatique » devant se superposer aux voix de synthèse « neutres » entendues en sortie de synthétiseur . Pour floues qu'elles paraissent de prime abord, ces deux propriétés se concrétisèrent dans plusieurs systèmes effectifs de production de parole artificielle à partir du texte, en s'appuyant sur un certain nombre de paramètres particuliers, isolés en comparant les voix de plusieurs locuteurs différents.

En premier lieu, cette approche distinguait la *qualité* vocale du *style* vocal. La première renvoie, du point de vue physique, au timbre, autrement dit à la composition fréquentielle du son vocal. Le second se rattache, lui, à l'intonation, c'est-à-dire aux variations de tonalité moyenne de la voix, et à ses caractéristiques temporelles : durée des sons, occurrences et longueurs des silences.

Du point de vue technique, il était assez simple de mettre au point des échantillons de voix différents, illustrant, pour une phrase identique, différentes variantes de timbre, de durées ou d'intonations. Restait à soumettre la séparation théorique opérée, et réalisée par un synthétiseur, à l'appréciation d'un panel d'auditeur, de façon à relier chaque combinaison de facteurs à un rendu plus ou moins « naturel » de la voix synthétique. Rien

de moins évident : comme nous l'avons signalé plus haut, l'extraction de facteurs prosodiques et de colorations de timbres susceptibles de refléter les caractéristiques d'une variabilité d'élocution typiquement humaine passent par un jugement… humain. Le centre du problème se déplace alors des enrichissements à apporter aux paramètres permettant de contrôler un synthétiseur vers l'évaluation de la voix produite par le synthétiseur après les modifications apportées. La voix est-elle plus brillante ? Plus sourde ? Plus nasale ? Plus joyeuse ? Plus dure ? Plus féminine ? Plus agressive ? Et même à être capable de répondre de façon assurée à l'une de ses questions, comment savoir alors lequel des paramètres physiques du synthétiseur provoque une impression si flagrante ?

Aujourd'hui, il n'existe encore aucune réponse définitive à ces interrogations. La multiplication des corpus étudiés, les recoupements opérés entre plusieurs expériences et les progrès enregistrés dans l'architecture informatique des synthétiseurs ont néanmoins permis certaines avancées.

Le groupe de recherche « Perception Située », section du département Communication Homme-Machine, au Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI), entité du CNRS, travaille par exemple sur une tentative de définition de l' « espace perceptif du timbre ». Convaincus que l'analyse , la synthèse et la perception de la voix entretiennent une relation de symbiose, les chercheurs de ce groupe travaillent, entre autres, à la construction d'un modèle illustrant la façon dont est perçu le timbre d'un locuteur. Dans la tentative d'élaborer une voix de synthèse indiscernable d'une voix naturelle, il est en effet indispensable de saisir la manière dont le timbre d'une personne donnée évolue, et dans quelle mesure il affecte la représentation que l'on s'en fait à son écoute.

Les premières conclusions montrent qu'il est nécessaire de définir le timbre comme une caractéristique qui ne s'appréhende qu'au travers d'échelles de temps variables. Ainsi, le timbre d'une personne observé à l'échelle de temps de l'énonciation des syllabes se rapproche-t-il, dans la façon dont il est perçu, de la notion de tessiture perçue dans le cas d'un instrument de musique. En revanche, considéré sur l'ensemble d'un discours, il tend plutôt à refléter des habitudes personnelles de prononciation, des styles mélodiques individuels.

D'autres travaux sont consacrés, au LIMSI toujours, à la caractérisation de la variabilité de l'intonation. Là aussi, il s'agit d'imaginer les moyens pertinents, du point de vue de la perception, permettant de quantifier la manière dont les évolutions des hauteurs de tons, pour un locuteur donné et pour des locuteurs différents, entraînent des impressions variables pour l'auditeur.

Si de telles recherches ne se traduiront pas avant encore longtemps en applications courantes, les systèmes de synthèse vocale qui en exploitent d'ores et déjà les résultats dans les laboratoires ouvrent des perspectives enthousiasmantes. Au LIMSI, des auditeurs « profanes » furent ainsi invités à écouter deux enregistrements. Le premier était une lecture naturelle, le second la même lecture dont la prosodie avait été modifiée selon les règles régissant d'ordinaire la prosodie artificielle d'un synthétiseur. Curieusement, les auditeurs trouvèrent la voix modifiée beaucoup plus agréable à entendre que l'originale : le débit, l'intonation et les pauses imposées par la machine eurent plus de succès que leurs homologues d'origine humaine.

Cette réussite, pour significative qu'elle soit, n'en demeure pas moins partielle. Une véritable victoire verrait le jour si, encore une fois, il devenait véritablement impossible de discerner l'artefact du naturel.

Nous avons vu que, dans une telle perspective, l'étude de la perception occupe une situation centrale. Car être capable de rattacher les représentations attribuées à différents styles et qualités vocales dans la parole naturelle à un système de caractéristiques acoustiques fixe, c'est se donner les moyens d'imiter la voix dans ce qu'elle a d'instable. Il ne s'agit pas, bien sûr, de vouloir faire évoluer une voix de synthèse, au fil de la lecture automatique d'un texte, en toute liberté : la problématique serait alors celle d'une véritable intelligence artificielle, capable de s'autodéterminer au travers de son expression vocale. Mais il s'agit de réussir à faire parler un synthétiseur «à la manière » d'un locuteur qui, pour n'être pas forcément unique, emprunte à plusieurs modèles humains tout ou parties de leurs spécificités vocales.

En premier lieu, comme nous l'avons évoqué notamment dans le cadre des travaux réalisés au LIMSI, cela implique de définir des paramètres acoustiques essentiels à la construction d'une voix singulière. Si les paramètres sont assez nombreux et maîtrisés, il devient alors possible d'envisager de les faire évoluer avec suffisamment de finesse et d'amplitude pour espérer « mimer » un comportement vocal naturel.

Dans le même temps, il est indispensable de parvenir à affecter un sens, susceptible d'évoluer, aux paramètres désormais munis d'un nombre suffisant de degré de liberté. Sans quoi, à l'instar des débits et intonations aléatoires proposés un temps comme solution, que nous avons mentionnés plus haut, la voix de synthèse, pour disposer d'une grande variabilité, n'en demeurera pas moins un « pantin désarticulé » qui ne ressemble à rien d'autre.. qu'une voix de synthèse.

Nous n'entrerons pas plus, ici, dans le détail de la construction des paramètres du point de vue de la synthèse. Avoir indiqué certaines des voies de recherche actuelles,

notamment dans l'élaboration d'un modèle de mesure quantitatif du timbre, nous semble suffisamment évocateur. Il n'est pas sûr qu'aller plus loin –exposer les méthodes de mesure de la force vocale, par exemple- éclairerait notre propos.

Il nous semble cependant particulièrement intéressant de nous pencher à présent sur une approche théorique tout à fait singulière, qui propose une façon originale de considérer la relation qu'entretiennent ces mêmes paramètres avec la manière dont ils sont perçus, et qui conditionne l'impératif de sens introduit ci-avant. Il s'agit de la théorie psycho-phonétique, telle que développée par Yvan Fonagý.

C.3. Une piste de solution pour faire vivre la voix de synthèse : la théorie psycho-acoustique d'Yvan Fonagý et ses prolongements potentiels

Nous avons pris le parti de ramener la réussite de la mise au point d'une voix de synthèse à l'obtention d'une voix qui sonnerait « vraie », « vivante » pour l'auditeur. Nous avons étayé cette prise de position par l'exemple, en mettant en avant l'imperfection flagrante des systèmes d'information embarqués, installés à bord de certains véhicules. Certains développeurs, oeuvrant pour les constructeurs automobiles, nous trouverons peut-être injustes. Pourtant, cette conception des limites actuelles de la voix de synthèse, qui s'inscrit au centre des préoccupations de la recherche fondamentale en synthèse vocale, trouve bel et bien des résonances dans les applications courantes des technologies vocales. De façon très générale, maîtriser les modalités de la perception de la voix conditionne véritablement le succès des voix de synthèse. En corollaire, l'étape cruciale à franchir consiste donc à conquérir les moyens d'une synthèse vocale « expressive ». Et si cette étape oppose tant de résistances, c'est qu'elle emprunte largement aux théories élaborées dans le champ des sciences sociales, dont on conçoit qu'elles se plient très difficilement à leur mise en œuvre pragmatique :

« La modélisation du processus de compréhension, dont une grande partie est fondée sur les approches symboliques, est assez souvent motivée par la saisie des processus cognitifs. Par conséquent, les théories sous-jacentes sont celles des linguistes et des psychologues. Une des conséquences de cela est que, pendant de nombreuses années, les technologies vocales ont été assez peu intégrées aussi bien en milieu professionnel que dans

le grand public »[36].

Les passerelles entre les théories développées par les linguistes et les psychologues d'un côté, et les laboratoires où s'élaborent les modèles les plus avancés de synthétiseurs vocaux de l'autre, commencent néanmoins à s'établir, à travers certaines approches théoriques aptes à fructifier dans un cadre pragmatique.

Ainsi en est-il de la parole vue comme un «acte de langage ». Cette conception lui reconnaît plusieurs composantes. Selon la terminologie d'Austin, figure de la philosophie analytique, l' « acte locutoire » renvoie ainsi au sens littéral de ce qui est dit. L' « acte illocutoire » consiste, de son côté, à rendre manifeste la manière dont la parole doit être comprise, pendant que l' « acte perlocutoire » est susceptible de provoquer des effets individuels pour l'interlocuteur (modifications de l'état d'esprit, émotions…)[37].

En allant vite, les synthétiseurs vocaux actuels parviennent à retranscrire le contenu des actes locutoires de la parole, pendant qu'ils en sont à leurs balbutiements quant à s'attacher à transmettre ses deux autres composantes. Cela semble compréhensible, si l'on précise plus avant la nature d'un acte illocutoire. Celui-ci, tel qu'il est défini, informe notamment les interlocuteurs sur les obligations qu'ils contractent, implicitement ou non, en écoutant une parole donnée et reflète l'état psychologique du locuteur. Il peut donc être assertif, directif, promissif, expressif, déclaratif…

Autrement dit, l'acte illocutoire forme la part de la parole qui fait que celle-ci n'est pas neutre, qu'elle entraîne des conséquences chez qui l'entend. Elle est action :

« [Vanderveken] a développé la logique illocutoire en vue de construire une sémantique formelle du langage naturel, capable de caractériser les aspects à la fois véri-conditionnels et illocutoires de la signification des énoncés. Par opposition aux doctrines philosophiques récentes, qui tentent de réduire la signification des énoncés au sens, Vanderveken considère que la signification et l'usage sont logiquement liés dans la structure du langage, si bien qu'il n'est pas possible d'analyser la signification linguistique d'un énoncé sans étudier les actes illocutoires qui peuvent être accomplis par ses énonciations littérales dans des contextes d'emploi appropriés »[38].

Un tel postulat théorique, malgré son intérêt, ne saurait cependant être d'un grand secours dès lors qu'il s'agit de le mettre en pratique dans la mise au point d'un synthétiseur vocal. Certes, il implique, à nos yeux tout au moins, qu'il est possible de séparer

---

[36] Wolfgang Minker, *Parole et dialogue homme-machine*, Editions du CNRS, coll. Sciences et Techniques de l'Ingénieur, Paris, 2001, p.16
[37] Wolfgang Minker, *op. cit.*, p.109
[38] *Ibid*.

objectivement et même, logiquement, certaines caractéristiques de la parole, caractéristiques qui seraient liées à l'association d'un contexte et d'une volonté. Malheureusement, la machine qui parle est incapable d'agir, de diriger sa voix selon des finalités qui lui sont propres : comment aurait-elle connaissance d'un quelconque contexte et de sa propre volonté, afin d'user en conséquence des bons paramètres acoustiques qui rendraient sa voix vivante…puisqu'elle n'est que machine ?

Les passerelles que nous croyions discerner se seraient ainsi effondrées ? Pas tout à fait. Une autre théorie, élaborée par Yvan Fonagý et désignée comme «théorie psycho-phonétique », pour être différente de celles d'Austin ou de Vanderveken, partage suffisamment de points communs avec elles et s'agrémente de débouchés pragmatiques assez nombreux pour nous autoriser à imaginer qu'elle puisse, un jour, se traduire dans les formalismes informatiques d'un synthétiseur. Et si tel n'est pas le cas, probablement aura-t-elle ouvert, croyons-nous, suffisamment de portes pour qu'un de ses futurs avatars le permette.

Yvan Fonagý, phonéticien, rejoint la notion d'acte de langage en proposant, de son côté, celle de « geste vocal ». Lui aussi considère que lorsqu'un locuteur s'exprime, l'information que véhicule sa voix dépasse largement la seule signification littérale des mots qu'il emploie. Il évoque ainsi, en traçant un parallèle avec la théorie de l'information, un « double encodage » de la voix. Les mots prononcés, choisis et ordonnés selon le vocabulaire et les règles de syntaxe d'une langue commune à des interlocuteurs en situation de communication orale, apportent un premier niveau d'information : le sens littéral qui leur est attribué conventionnellement. Au-delà de ce premier niveau, toute une gamme de variations dans le ton, le débit, l'usage des silences, renseignent non seulement sur l'état du locuteur, mais aussi sur les intentions qu'il prête à ses paroles.

L'intérêt de l'approche de Fonagý est double, car il enrichit ce découpage théorique, plus long à formuler qu'à saisir, de multiples expériences pragmatiques, où l'analyse phonétique occupe le premier plan.

L'originalité des conclusions qu'il en tire est intéressante elle aussi pour deux raisons. D'une part, il considère que le «double encodage » du langage trouve son origine dans l'inconscient personnel du locuteur aussi bien que dans l'inconscient collectif de la communauté humaine. D'autre part, il propose d'établir une série de caractéristiques acoustiques de ce « double encodage »… propres, cette fois, à participer à la conception de modèles de synthétiseurs vocaux « expressifs ».

Avant de détailler sa démarche, il importe de démarquer les travaux d'Y. Fonagý de ceux que nous avons évoqués à propos du LIMSI. Là où les concepteurs des synthétiseurs

de demain mettent l'accent sur l'analyse du signal acoustique de la voix, Fonagý étend son étude à la gestuelle de l'appareil phonatoire : mouvements de la langue, déplacements des lèvres, tension des muscles du larynx... Ces paramètres physiologiques tiennent d'ailleurs une place largement aussi importante, voire plus déterminante, dans la construction de sa théorie que les grandeurs acoustiques que sont le ton, le spectre de fréquence ou l'intensité. Cela pourrait sembler handicaper toute volonté d'exploiter les conclusions qu'il tire de ses observations dans le cas d'un système artificiel informatique de production de la voix. Les circuits électroniques traitent en effet le signal électrique qui deviendra voix, appliqué à des hauts parleurs, avant tout en terme de fréquences, de temps et d'intensité. Pour autant, cet écart apparent semble pouvoir être dépassé, en considérant qu'il est relativement plus facile de relier les caractéristiques physiologiques instantanées de la phonation à leur résultat acoustique, que ce dernier aux phénomènes psychologiques qui les précèdent et qui, eux, forment l'objectif d'explicitation de Fonagý.

Pour être phonéticien, Y. Fonagý n'appuie pas tant sa théorie psycho-phonétique sur les caractéristiques intrinsèques du son vocal que sur leur origine psychanalytique. Il relie d'ailleurs les traits empiriques de la phonation, qui sembleraient, de prime abord, ne pouvoir donner prise qu'à des considérations mécaniques, aux stades décrits par Freud dans sa méthode psychanalytique.

La distance qui, d'ordinaire, sépare les domaines d'analyse qu'il rapproche, appelle a priori à la prudence. La manière dont il la réduit se base cependant sur un certain nombre de préalables.

Fonagý insiste en particulier sur le rôle principal tenu par les métaphores, qui, quelles qu'elles soient, « ne sont jamais gratuites ». Relativement faciles à identifier dans la langue écrite, elles demandent cependant un certain effort avant de se révéler dans la « vive voix », au delà des seuls usages qui font que l'on dise d'une voix qu'elle est « lourde », « plate », « blanche », « vulgaire », « distinguée »…

L'expressivité de la voix provient ainsi, selon le phonéticien, de métaphores cachées dans le son, conscientes et inconscientes, qui enrichissent la parole d'informations relatives au sujet qui parle, et que l'auditeur comprend grâce à un ensemble de conventions tacites.

Ces conventions prennent elles-même leurs racines dans l'évolution historique du langage et l'origine des émotions. Celles-ci constituent d'ailleurs la matière des appuis expérimentaux de la théorie psycho-phonétique, et Fonagý emprunte à Darwin le contexte de leur définition. « C'est à Charles Darwin […] que nous devons la reconquête de ce domaine où semblent dominer l'aléatoire, l'irrationnel. Darwin restitue le contexte

d'apparition des attitudes émotives : le monde qui entourait l'homme au temps de nos premiers ancêtres. En replaçant les émotions dans leur nature originelle, Darwin les transforme en activités dirigées vers des buts qu'il considère comme des activités complexes, utiles dans certaines circonstances, et qui sont automatiquement déclenchées en situation analogue »[39]. Les mains rendues moites par la peur renverraient ainsi à la peur ancestrale du prédateur, comme un facteur favorable à la fuite.

Fonagý prolonge cette conception dans l'analyse phonétique de la voix, et estime que les symptômes ancestraux (en situation de fuite, d'attaque, de quête du partenaire) rejaillissent, en vestiges, sur la phonation, au point que les organes de l'appareil phonatoire lui paraissent pouvoir légitimement représenter d'autres objets animés ou inanimés, par analogie fonctionnelle : « la langue peut représenter le bras, le déplacement de la langue vers l'avant et le haut peut représenter le bras qui pointe vers l'avant et le haut […] »[40].

Ainsi certaines attitudes vocales seraient-elles le reflet perpétué de comportements ancestraux, et cet ancrage pré-historique permettrait-il de les faire comprendre au-delà des limites imposées par des langues différentes.

L'exploitation de la théorie psychanalytique vient compléter la théorie originale de Fonagý, en renforçant le caractère conventionnel et, par là, susceptible d'explicitation, du deuxième niveau d'encodage de la parole, qui donne à la voix son « grain », son individualité.

«Le geste vocalique est donc l'écart entre l'articulation idéale et celle qui sous-tend le son concret : le geste vocal est donc un geste-écart virtuel. Ce geste-écart jouit d'une certaine indépendance vis-à-vis des restrictions phonétiques imposées par la langue. Quels que soient les traits articulatoires imposés par la langue, la tension psychique (agressive, angoissée ou joyeuse) se manifeste invariablement par une articulation plus tendue que celle exigée normalement pour le phonème donné, dans un contexte phonétique donné. La langue exige pour la voyelle /i/ une articulation antérieure, une position linguale élevée. Ces prescriptions devront être respectées même dans une profonde tristesse qui s'exprimerait spontanément par un affaissement des muscles labiaux et linguaux, par une position rétractée de la langue. […] Ce qui vaut pour la communication préconsciente de certaines attitudes émotives vaut également pour l'expression inconsciente des contenus pulsionnels. La pulsion qui relève de la deuxième phase du stade anal tendra toujours à empêcher le flux libre de l'air sonore, à créer des obstacles pour mieux répondre au niveau

---

[39] Yvan Fonagý, *La vive voix, essai de psycho-phonétique*, éditions Payot, Paris, 1991, p. ?
[40] Yvan Fonagý, *La vive voix, essai de psycho-phonétique*, éditions Payot, Paris, 1991, p. ?

laryngo-buccal au modèle de la rétention anale. Elle s'exprimera donc nécessairement dans le renforcement des consonnes occlusives, en augmentant la superficie de l'occlusion ainsi que sa durée ; elle accroîtra la friction en réduisant le passage de l'air dans les consonnes fricatives, en réduisant la durée relative des voyelles et des sonnantes »[41].

Il est clair que la théorie psycho-phonétique fait appel à de nombreux soubassements théoriques antérieurs, et qu'elle les ordonne suivant une logique qui, pour être efficace en apparence, n'en demeure pas moins discutable. Néanmoins, la vérification expérimentale vient étayer le propos de son auteur, dont les travaux de recherche relatent nombre d'exemples convaincants :

« Trois jeunes actrices ont prêté gracieusement leur concours à une expérience ayant pour but l'analyse de la projection sonore des attitudes agressives et non agressives, la colère, la haine, l'ironie d'une part, la tendresse, le désir, la joie d'autre part. Il est apparu que les émotions à base sadique-anale reproduisent la rétention au niveau vocal. Ainsi les émotions tendres favorisent (allongent) les voyelles, par contre, la haine, la colère et, à un degré moindre, l'ironie, augmentent la durée des consonnes, et allongent surtout la durée des occlusives sourdes, et réduisent la durée relative des voyelles »[42].

D'autres exemples permettraient de détailler les «preuves » par lesquelles Fonagý montre que tel ou tel type d'émotion est, de façon plus ou moins directe, traduite en tel ou tel trait –style ?- vocal. Nous nous contenterons ici de retenir certaines des conclusions dégagées à partir de nombreuses expériences réalisées en langues française et hongroise, et relatant plus particulièrement les caractéristiques dynamiques des organes de la phonation exprimant des émotions particulières :

« La colère se manifeste en français comme en hongrois :

a) par des mouvements saccadés de la langue : des périodes extrêmement rapides sont suivies de périodes de figement, prolongeant la période de la tenue dans les positions extrêmes : la plus fermée pour les voyelles fermées, la plus ouverte pour les voyelles ouvertes ;
b) par la tension musculaire linguale, labiale, pharyngée : la langue est fortement pressée contre le palais pour le /k/ […] ;
c) les voyelles /i/, /y/, /u/ fermées sont plus fermées et leurs analogues ouvertes nettement plus ouvertes ; ainsi l'angle maxillaire augmente et la distance entre les lèvres s'accroît dans le cas de /a / autour d'une moyenne de 25 mm ;

---

[41] Yvan Fonagý, *La vive voix, essai de psycho-phonétique*, éditions Payot, Paris, 1991, p. 33
[42] *Ibid*.

d) la langue est plus rétractée pour les voyelles comme pour les consonnes par rapport à la parole neutre, ce qui prête à la voix un timbre plus sombre ;

[…]

Dans la tendresse simulée :

a) l'articulation est déliée ; les transitions sont plus lentes, plus graduelles, et les tenues relativement courtes ;
b) l'articulation labiale et linguale est plus relâchée même par rapport à la parole neutre, le voile du palais est détendu ;
c) l'articulation est relativement avancée, ce qui prête, dans l'ensemble, aux sons un timbre plus clair ;
d) les /i/ et les /e / sont souvent labialisés […] »[43].

De telles observations nécessitent un appareillage spécifique. Fonagý a ainsi recours à la radiocinématographie, à l'électroglottographie, techniques qui permettent de fixer sur le papier ou sur un écran les mouvements de la langue, des lèvres, du pharynx, aussi bien que l'ouverture de la glotte, la tension des muscles du larynx… et de les comparer dans des situations de locution différentes. Il va sans dire que les protocoles expérimentaux qu'implique un tel équipement ne sont pas envisageables dans les situations courantes, naturelles, de communication. C'est la première raison pour laquelle Fonagý fait appel, pour ses expériences, à des comédiennes et comédiens. Ce qui l'oblige à parler, par exemple, de « tendresse *simulée* ». Mais le recours à des volontaires rompus à l'art dramatique présente un avantage plus grand peut-être que la seule commodité technique. Non seulement les artistes peuvent réciter un texte selon des attitudes « commandées », mais, et c'est peut-être le plus important, le phrasé théâtral densifie, concentre, le poids des attitudes et des émotions, en regard des situations naturelles. Il est alors d'autant plus simple d'en extraire des caractéristiques significatives.

En enrichissant la palette d'attitudes soumise à l'expérimentation, la théorie psycho-phonétique dégage ainsi un panel de postures vocales, relative chacune à une «catégorie » d'émotions donnée. Les fondements psychologiques qu'elle attribue à chacune de ces postures sont déterminants, quand il s'agit d'affirmer que les caractéristiques de la phonation relèvent de conventions sous-jacentes, compréhensibles au-delà des contraintes

[43] Yvan Fonagý, *La vive voix, essai de psycho-phonétique*, éditions Payot, Paris, 1991, p. ?

de la langue d'usage des locuteurs, et que, comme c'est le but de Fonagý, on cherche à en brosser le complexe tableau.

Dans notre recherche des conditions d'existence d'une voix de synthèse « expressive », en revanche, c'est la première corrélation, celle qui relie les postures vocales, en tant que caractéristiques acoustiques et physiologiques objectives, aux émotions qui retient avant tout notre attention.

En effet, il devient alors envisageable de configurer un synthétiseur vocal avec des paramètres explicites dont on va pouvoir déterminer, du point de vue perceptif, les émotions qu'ils traduisent. La machine parlera avec une voix « gaie », une voix «triste », ou affichera sa « colère ».

Si l'on suit Fonagý, il paraît possible d'aller au-delà, en donnant même à une voix une coloration personnelle, individuelle, sans se limiter à l'expression momentanée d'une attitude émotive. D'après le phonéticien, en effet, les traits acoustiques qui identifient une émotion donnée prennent aussi part, quand ils deviennent permanents, à la constitution d'une « personnalité » vocale :

« Dans la communication quotidienne, on se contente généralement de ramener un complexe de gestes-écarts vocaliques à telle ou telle attitude (émotive ou intellectuelle). Il arrive cependant que les traits qui expriment normalement une attitude, par exemple une attitude agressive –les accents fréquents et vigoureux, la contraction laryngée et pharyngée, etc.- deviennent dans la parole d'une personne extrêmement fréquents, quasi permanents. Ces particularités vocales seront, par conséquent, de moins en moins perçues par son entourage. Il semble que la loi d'accommodation, qui fait qu'un stimulus constant élicite des réponses de plus en plus faibles, vaut également au niveau de l'interprétation psychologique des signaux sonores. A une différence près : tandis qu'une sensation tactile (le contact d'une chemise qu'on porte) peut complètement disparaître, sans laisser de trace, les gestes vocaux qui constituent le style vocal se transforment : ils disparaissent en tant que gestes phonatoires, porteurs de messages, pour réapparaître comme manière de parler individuelle. Cela revient à dire que les gestes vocaux se détachent de l'attitude émotive qu'ils reflètent et seront directement rattachés, sans analyse sémantique préalable, à la personne du locuteur, pour faire partie de son signalement, au même titre que la couleur de ses cheveux, sa taille, son nom »[44].

[44] Yvan Fonagý, *La vive voix, essai de psycho-phonétique*, éditions Payot, Paris, 1991, p. ?

A ce stade, il semblerait que le «Graal » de la voix de synthèse puisse être atteint. Résumons. Le principe des synthétiseurs à formant permet, dans un premier temps, de faire lire à la machine n'importe quel texte en continu, en respectant des règles prosodiques les plus fidèles possibles à la syntaxe de la langue dans laquelle on souhaite travailler. Reste à apporter au « geste phonatoire neutre » de la machine une coloration expressive et, si possible, personnelle. C'est, comme nous venons de le voir, ce que semble promettre la théorie psycho-acoustique de Fonagý. Au prix d'une traduction, sans doute longue et laborieuse, mais envisageable, des observations agrégées par le phonéticien, dans un modèle informatique conformant les caractéristiques physiologiques et phonétiques de l'expression des attitudes émotionnelles, il devient possible de « faire vivre » le synthétiseur. De la même façon que de nombreux corpus d'enregistrements de locuteurs humains ont permis d'élaborer des bibliothèques de diphones et d'affiner les règles prosodiques rendant compte de la syntaxe et de la ponctuation d'un texte, on peut imaginer constituer des corpus de gestes vocaux rendant compte d'une palette d'attitudes expressives. Et pourquoi ne pas aller plus loin, et exploiter la théorie psycho-acoustique jusqu'au bout, en conférant à la machine un passé virtuel, une chimère de profil psychologique, puisque les conclusions de Fonagý paraissent autoriser un rapprochement explicite entre l'acte phonatoire et le fonds psychanalytique d'un individu ?

Au delà de ses seules contingences techniques, dont on pourrait espérer s'affranchir en s'en remettant au progrès technologique porteur de solutions providentielles, un tel projet, aussi séduisant soit-il, semble malgré tout très loin de voir le jour. Et la voix de synthèse, fût-elle expressive, ne paraît pas pouvoir jamais créer l'illusion de vie qu'on souhaiterait tant lui prêter.

D. En quête d'inouï

D.1. Une autre voie pour la voix de synthèse : la machine au service de la création à l'IRCAM

Exploiter la théorie psycho-acoustique sous un angle technique, comme nous venons de le proposer en tant que projet de voix de synthèse «idéale », c'est-à-dire aussi proche que possible d'une voix naturelle, s'apparente, basiquement, à construire une banque d'attitudes émotives sonores et de se doter les moyens de les appliquer en toute liberté à la lecture artificielle d'un texte. Pour cela, les observations personnelles de Fonagý, complétées des recherches effectuées par d'autres chercheurs reprenant, au moins sur le plan pragmatique, la démarche du fondateur de la théorie psycho-acoustique, forment la matière dans laquelle puiser.

Cependant, cette ressource apparemment prometteuse risque fort de ne pas produire les résultats attendus, pour deux raisons distinctes.

En premier lieu, les corrélations établies entre la manifestation vocale d'une émotion et son existence intrinsèque, au sein du psychisme d'un individu, ne sauraient relever d'autre chose que d'une interprétation subjective. Qui garantirait, en effet, que telle ou telle émotion sera définie, partout et toujours, de la même façon ? Qui, même, oserait prétendre qu'il est raisonnable d'enfermer définitivement les émotions dans une nomenclature déterministe ? De tels classements existent pourtant, mais leur seule hétérogénéité suffit à disqualifier leur éventuelle prétention à la construction d'un sens universellement intelligible. Suivant les auteurs, les émotions se trouvent réparties dans des typologies qui comptent entre 7 et 700 catégories… Comment choisir ? Laquelle de ces typologies se prête-t-elle la mieux à sa mise en œuvre dans un système de règles prosodiques, voire de styles de timbres, capable de restituer une imitation de vivacité vocale ?

A prendre la moins adaptée, le synthétiseur risque de ne produire qu'une désagréable caricature. A opter, par chance, pour celle qui s'avèrera la meilleure, il reste l'obstacle de la variabilité inhérente à la phonation humaine, que les apports de la théorie psycho-acoustique ne permettent pas de franchir.

En effet, toute typologie d'émotions, aussi raffinée et, dans la mesure du possible, réaliste soit-elle, ne peut être utile que si elle est intégrée de façon évolutive au synthétiseur qui l'exploite. Autrement dit, les émotions agrémentant la lecture d'un texte par la machine doivent être intégrées, dans l'idéal, d'une manière qui n'a rien d'automatique, au risque de

réduire leur contribution à la part illocutoire du langage à une surimpression purement déterministe. Le double encodage de la parole, permis par diverses tentatives de caractérisation acoustique des émotions, telle celle que propose la théorie psycho-acoustique, demeure rigide et révèle, immanquablement, la nature artificielle de la voix produite.

Un deuxième écueil, dérivatif de l'irréductible variabilité de la voix humaine, s'érige dans le chemin que la voix de synthèse essaie de se frayer vers la voix naturelle. Il s'agit de ce l'on pourrait appeler l'infaillibilité de l'oreille humaine. Celle-ci est dotée d'une grande sensibilité, facilitant la reconnaissance des sons voisés par rapport aux timbres d'autres instruments. La démarche d'A. Tomatis se réclamait d'ailleurs de ce constat : la moindre lacune sur les variations spectrales des sons vocaux est immédiatement détectée par l'oreille. Or, utiliser une typologie d'émotions pour l'imposer à un système de synthèse vocale, c'est, nécessairement, emprunter des raccourcis, faire des impasses, effectuer des choix réalistes eu égard aux contraintes technologiques du traitement informatique du son. Par conséquent, les défauts occasionnés dans la voix produite –le passage insuffisamment graduel ou, au contraire, trop progressif de l'expression acoustique de la colère appliqué à un passage d'un texte lu à une locution neutre dans la suite du même texte- seront immédiatement perçus par l'auditeur humain.

Une question viendra probablement au lecteur : pourquoi avoir insisté autant sur le fait qu'il existe des démarches de recherche tournées vers l'explicitation des phénomènes phonatoires conditionnés par l'émotion, si le résultat de telles démarches ne saurait conduire, quoi qu'il en soit, vers l'obtention d'une voix de synthèse «clone » de la voix naturelle ? Question légitime, qui valait d'être posée, d'ailleurs, aux chercheurs concernés : quelle est donc cette voix qu'ils espèrent faire jaillir de leurs synthétiseurs, puisqu'elle semble devoir rester à jamais in-humaine ?

Nous allons tenter d'y apporter une réponse, à partir des propos échangés avec Xavier Rodet[45], directeur de recherche à l'Institut de Recherche et Coordination Acoustique / Musique (IRCAM), et plus spécifiquement responsable de l'équipe de recherche « Analyse Synthèse ».

C'est à Xavier Rodet et son équipe que Gérard Corbiau, réalisateur du film *Farinelli,* fit appel pour réaliser la voix du castrat éponyme, véritable «star » de l'époque Baroque, dans l'Italie du XVIII^{ème} siècle . En 1994, année de tournage du film, il était

[45] Entretien avec X. Rodet, IRCAM, juillet 2002

impossible de recruter un chanteur castrat pour enregistrer les 39 minutes de chant prévues au total par le scénario. Les castrats sont en effet interdits depuis le siècle dernier ; le dernier représentant de cette catégorie très spéciale de chanteurs s'est éteint en 1922, et le recours à une voix virtuelle était donc incontournable.

Reproduire une voix de castrat ressemblait, au départ, à un véritable défi. Les quelques rares enregistrements préservés des derniers castrats s'avérèrent rapidement inexploitables, car les supports sur lesquels ils avaient été fixés avaient été largement endommagés. Il fallait donc imaginer ce qu'avait pu être la véritable voix d'un castrat, et non pas reproduire quelque chose de connu, d'où il eût été possible d'extraire des paramètres acoustiques utiles à la synthèse.

Les castrats, jeunes chanteurs émasculés avant la puberté pour empêcher leur voix de muer, alliaient la capacité pulmonaire, l'endurance et la force physique d'un homme, à un larynx proche de celui d'une femme, donc plus souple et de taille plus petite. En conséquence, ils étaient capables de chanter avec une puissance incomparable sur plus de trois octaves, tenant les notes sur plus d'une minute sans avoir besoin de reprendre leur souffle.

Le travail de synthèse visant à approcher ce qu'avait du être la voix de Farinelli consista, pour l'équipe de Xavier Rodet à « fusionner » les voix, bien réelles, d'une soprano et d'un contre-ténor. Pour cela, des enregistrements des deux interprètes furent réalisés numériquement, lors de sessions avec orchestre. Puis la voix de la soprano fut soumise à une procédure de *morphing*, visant à rapprocher son timbre de celui de la voix du contre-ténor. Cette voix modifiée une première fois fut altérée à nouveau, pour lui conférer un rendu plus juvénile, par l'atténuation de certaines bandes de fréquence de son spectre. Enfin, l'enveloppe spectrale dans son ensemble fut retouchée, afin de donner un aspect plus brillant à la voix finale.

Le résultat est édifiant, et seules des oreilles exercées parviennent à identifier certaines discontinuités dans le chant de synthèse, trahissant sa nature artificielle.

Mais quelle est cette voix ? Celle de l'un des deux interprètes ? Celle de Farinelli ? En réalité, aucune : ce n'est rien d'autre qu'une chimère vocale.

Pour Xavier Rodet, elle n'en demeure pas moins une réussite. En effet, elle crée une illusion de vie qui n'essaie pas de mimer un exemple –l'exemple en question étant quoi qu'il en soit inaccessible-, mais crée en revanche une représentation vocale réaliste, tout aussi capable de susciter enthousiasme et émotion. La voix de synthèse est, bel et bien, création et non reproduction, clonage.

Jusqu'ici, nous l'avions pourtant définie implicitement par la tentative de reproduction de la parole naturelle. Nous avions donc choisi d'aborder les éléments qui composent les modèles susceptibles de conduire à cet objectif. Les liens entre les conditions humaines de la phonation et les paramètres physiques mis en œuvre au sein des synthétiseurs vocaux semblaient devoir s'affiner, au fil des théories présentées, pour aboutir, idéalement, à des systèmes informatiques copies conformes, dans leur logique comme dans leur résultat sonore, à la voix naturelle.

Il apparaît en fait que les développements les plus récents des modèles numériques de voix, s'ils ne cessent de s'enrichir en effet de l'analyse de la voix naturelle, tendent à servir des finalités qui dépassent, par défaut en partie, par l'existence d'autres motivations surtout, la quête de l'imitation parfaite.

Comment définir, alors, la voix de synthèse ? Peut-être une analogie avec la synthèse chimique, proposée par Tassoula Georgaki, en donne-t-elle une image originale, ouverte sur la création plutôt que préoccupée par la compréhension exhaustive d'un phénomène : « La synthèse chimique, telle que l'a conçue Berthelot, n'est pas seulement la reconstitution de corps déjà connus et analysés, mais un procédé d'investigation direct, au moyen de la production artificielle de composés nouveaux qui n'ont jamais été rencontrés dans la nature, et n'ont pas pu, par conséquence, être soumis à l'analyse »[46].

Ainsi l'aller-retour incessant entre l'analyse et la synthèse donne-t-il autant à comprendre qu'il permet de créer. L'analyse première de ce qui existe dans la nature –la voix humaine, sa physiologie, la prosodie, l'expressivité- permet la synthèse artificielle dans un premier temps par imitation –les synthétiseurs vocaux, malgré leurs imperfections- puis autorise la création, en ordonnant les éléments séparés par l'analyse d'une façon radicalement nouvelle.

On entre ainsi dans le paradigme de la nouveauté, sous-jacent aux échanges qui ont lieu à l'IRCAM, notamment, entre chercheurs et compositeurs. De la même façon que les instruments de musique artificiels ouvrent de nouveaux horizons à la composition, les synthétiseurs vocaux permettent d'imaginer des interprétations chimériques, aux frontières de l'humain, laissant libre cours à une créativité autrement limitée par les performances vocales des interprètes de chair et de sang.

---

[46] Tassoula Georgaki, *Problèmes techniques et enjeux esthétiques de la voix de synthèse dans la recherche et création musicales*, thèse EHESS/IRCAM/CID-CNRS, sous la direction de H. Dufourt, Paris, 1998, p. 462

De fait, les voix de synthèse développées à l'IRCAM n'ont pas d'interprètes. Elles sont désincarnées, elles n'ont pas d'âme, pas de corps, et pourtant elles existent. Leur perception elle-même est un territoire inexploré.

Nous serions tentés d'affirmer que cette philosophie de la synthèse, celle de produire de l'inouï, est sans doute beaucoup moins restrictive et plus enrichissante que la quête, perdue d'avance, de la voix humaine enfermée dans la machine.

En définitive, la synthèse vocale est un mode d'investigation de la voix telle qu'elle est, tout autant qu'une projection imaginaire de ce qu'elle ne peut pas, naturellement, atteindre.

Nous allons voir par la suite que la voix humaine naturelle se prête aussi, elle-même, à un travail visant à démultiplier ses horizons, à faire jaillir l'inouï. Et c'est, nous semble-t-il, heureux, car sinon il faudrait conclure ici que l'homme, pour jouir de la liberté de découvrir d'autres territoires vocaux, n'aurait d'autre recours que la machine informatique. En réalité, sa propre mécanique naturelle dispose de bien des ressources…

D.2. L'inouï au théâtre : libérer la voix

Quand nous avons abordé le travail de la voix dans le théâtre, après avoir entr'aperçu, via l'exemple du chant diphonique, la malléabilité de l'organe phonatoire, nous avons cité Sarah Bernhardt, qui disait que «la voix est l'instrument le plus nécessaire à l'artiste dramatique »[47]. La comédienne insistait par là sur la nature d'*instrument* de la voix. Nous souhaitions alors relever cette façon de considérer l'appareil phonatoire, pour mieux signifier qu'il constituait, pour les artistes, un outil de travail perfectible. Dans la foulée, nous avons évoqué un certain nombre de techniques et de stratégies pédagogiques, au travers des arts dramatiques et lyriques, détaillant les moyens d'appréhender l'outil vocal dans une logique performative.

Nous allons à présent tenter de dépasser ce regard utilitariste sur la voix, en ouvrant d'autres perspectives à la place qu'elle occupe dans l'art dramatique.

A l'instar de la voix de synthèse élargissant ses perspectives vers des horizons qui dépassent la dichotomie de l'analyse et de l'imitation, la voix du comédien se prête à une approche où la clarté de l'élocution, la force de projection –le volume- suffisante à atteindre les derniers rangs de spectateurs, l'intensité dramatique nourrissant le timbre passent au second plan. Bien sûr, la qualité vocale de l'artiste dramatique reste tributaire des connaissances qu'il a pu acquérir des caractéristiques physiologiques des organes de la phonation, et des capacités qu'il a, en conséquence, de les maîtriser. Cependant, et cela est d'autant plus vrai dans le théâtre contemporain, les canons classiques d'une «belle voix de théâtre » ont tendance à céder la place à un champ de liberté, où les écarts aux chemins empruntés par les grands comédiens sont non seulement tolérés, mais encouragés comme un support au renouveau, à la création.

Art vivant par excellence, le théâtre doit ainsi chercher de nouvelles inspirations, un nouveau souffle, susceptible de résonner avec les préoccupations d'un siècle incertain, changeant, dangereux. Certains directeurs de troupes et metteurs en scène avouent d'ailleurs ne plus supporter ce que nous serions tentés d'appeler des «voix de théâtre », évoquant par ces termes les archétypes vocaux qui ont pu régner trop longtemps sur scène, et dont l'usage ne correspond plus à l'époque contemporaine. Claude Régy, metteur en scène et directeur de la compagnie *Les Ateliers Contemporains* est un de ceux-là : «Les

[47] Sarah Bernhardt, *L'art du théâtre. La voix, le geste, la prononciation*, L'Harmattan, coll. Les introuvables, Paris, 1993, p.41

gens de théâtre sont très primaires, très primitifs, et surtout fonctionnent avec des habitudes ancestrales, qu'on retrouve, justement, chez les jeunes gens qui veulent faire du théâtre. On entend des voix de naturalisme, des voix de boulevard, chez des gens qui sont en général peu allé au théâtre, qui n'ont rien vu, rien entendu, et qui ont très peu lu. Il y a des gens qui ont, à dix-neuf ans, une parole conventionnelle, qui parlent, d'instinct, une parole de bois, qui est certainement imitée de quelque part, et pas seulement de ce qu'ils entendent autour d'eux […] »[48].

Un certain théâtre, contemporain, essaie donc de s'affranchir des habitudes vocales propres au théâtre classique. Pour cela, il doit repenser la manière dont le travail vocal est abordé par les jeunes comédiens. Il ne s'agit pas d'éliminer les ateliers vocaux, ou de nier qu'il y ait un quelconque intérêt à se préoccuper de la voix, mais d'inventer de nouvelles manières de l'appréhender. Cela semble d'autant plus important que « la voix est le metteur en scène et l'acteur. C'est elle, plus que le visible, qui est l'essentiel du théâtre. Il y a des pièces radiophoniques. Ce n'est pas la voix qui est dans le théâtre, c'est le théâtre qui est dans la voix. On pourrait dire que le théâtre est complètement théâtre quand c'est la voix qui donne à voir, et le visible à entendre, tous deux inséparablement »[49].

Comment alors libérer la voix des stéréotypes qui l'enferment dans une reproduction perpétuelle de ce qui a déjà été fait ? La position de Claude Régy fournit un point de départ :

« Au théâtre, en général, il me semble qu'on travaille à l'envers. On entend encore parler de travailler l' « intonation », comme s'il y en avait une, et, surtout, on veut comprendre chaque phrase. En même temps, on rajoute du jeu, du sentiment, on englue, on fait pléonasme avec le sens qui déjà est épuisé très vite ; donc on ne fait pas entendre l'écriture, on s'agite, on fait des effets en tout genre, et l'écriture est ensevelie, elle est enterrée. […] D'un point de vue général, tous ces termes qu'on emploie quand on parle de « personnage », d' « incarnation », de «psychologie du personnage », d' « être ou de ne pas être le personnage », de « travailler le personnage », tout cela me paraît vraiment inutile. Quand, d'autre part, je vois des gens se consacrer à ce qu'on appelle le travail vocal, et d'autres se spécialiser dans le travail corporel et la gestuelle —on voit des gens qui passent trois ans à travailler leur voix et des gens qui passent trois ans à travailler leur corps— ça me paraît une aberration. Il est évident que la voix, au théâtre (pour chanter, il

[48] Claude Régy, entretien avec Gérard Dessons, « Le champ de la voix », *Penser la voix* (textes réunis par Gérard Dessons), éditions La licorne, UFR Langues Littératures, Poitiers, 1997, p. 46
[49] Henri Meschonnic, « le théâtre dans la voix», *Penser la voix* (textes réunis par Gérard Dessons), éditions La licorne, UFR Langues Littératures, Poitiers, 1997, p.39

faut certainement acquérir au moins une technique, même si, comme je le crois, cette technique n'est pas suffisante : on devrait, dans les écoles de chant, apprendre aussi autre chose), la voix ne peut pas être travaillée isolément. Séparer la voix du corps, c'est une vivisection. L'origine du geste et l'origine de la vibration vocale viennent sans doute du même centre en nous, de même qu'il y a sans doute un seul centre d'où émanent tous les actes de création ; Quand on travaille, on se tait d'abord, on passe par le silence, de même qu'on passe par l'immobilité, on se met en état d'écoute et de réceptivité, […]. A partir de là, on peut trouver, en écoutant, tout ce qui est dans le texte, et pas simplement le sens du texte : on peut trouver la voix pour le dire, et on peut trouver le geste »[50].

Autrement dit, la conquête d'une voix en accord avec le texte d'une pièce de théâtre ne peut pas se mener de façon fructueuse si l'on se contente de raccorder à l'analyse sémantique d'un texte des modèles expressifs d'émotion, des intonations a priori idoines. De la même façon qu'un synthétiseur vocal ne peut reproduire une voix vivante en mimant d'hypothétiques modèles émotionnels acoustiques, la voix d'un comédien ne peut se couler dans un texte qu'en n'ayant recours qu'à sa propre sensibilité… « la voix est un sens […], le sens de l'affect le plus grand qui soit, dans toutes ses variations, l'affect de dire le vivre. Elle en porte et elle en transmet toute l'animalité, toute l'historicité »[51].

Pour interpréter un texte, le comédien doit donc s'impliquer dans l'écriture de l'auteur, s'imprégner pour faire jaillir quelque chose de nouveau, qui est à la fois ce qui émane de l'auteur et ce qui est le propre singulier de l'interprète : « Plus il y a d'affect dans la voix, plus on a du sujet dans la voix, dans *sa* voix ; plus l'écriture est subjectivée, plus elle peut se dire la voix du sujet. Plus l'écriture est écriture, plus elle est la voix. Invention, non inscription »[52].

La voix du comédien échappe donc à une prédétermination par le texte. Elle doit se révéler à elle-même grâce au texte, mais ne lui est pas enchaînée. Reste à ancrer cette conception dans la réalité. Comment se traduit une telle philosophie du travail vocal dans les faits ? Quels objectifs donner à des ateliers vocaux ? Le travail réalisé au Roy Hart Théâtre, puis au sein de la troupe Pantheatre, offre un aperçu de ce que peut être la création vocale, de cette façon d'aborder la voix non pas comme un instrument docile à soumettre à des règles établies, mais comme un facteur d'ouverture.

---

[50] Claude Régy, *loc. cit.*, p.45
[51] Henri Meschonnic, *loc. cit.*, p. 25
[52] *Ibid*.

L'originalité de la démarche créative du Roy Hart Théâtre s'inscrit dans le courant théâtral des années 60-70 et présente, par conséquent, un certain nombre de points de confluences, tant du point de vue des objectifs que des moyens de mise en œuvre. Elle dépend surtout de son fondateur, Roy Hart et, avant lui, de la personnalité singulière d'Alfred Wolfsohn. Celui-ci fut le précurseur d'une approche où la voix, avant d'être un moyen d'expression artistique, s'avère avant tout capable d'une part de témoigner de la psychologie, passée et présente, d'un individu, d'autre part d'offrir une prise à vocation thérapeutique sur sa vie intérieure, son psychisme, son rapport au monde.

Alfred Wolfsohn, né à la fin du XIX^{ème} siècle, participa aux deux guerres mondiales. Les cris des soldats mutilés, agonisant, le traumatisèrent. Il retira de cette expérience douloureuse le sentiment que la voix est capable, poussée dans ses limites, de retranscrire toutes les émotions qu'un homme est susceptible d'éprouver, et qu'elle lui permet, en conséquence, de mieux les intégrer, de mieux les comprendre.

Il instaura donc une pédagogie du chant qui n'avait plus rien à voir avec les credo du « bel canto », où la recherche d'un son pur est l'objectif premier, pour travailler au contraire sur les cassures, les imperfections de la voix, largement plus révélatrices de la nature humaine.

Un de ses élèves, Roy Hart, fut marqué profondément par leur première rencontre, au point d'affirmer : « Je me suis rendu compte que j'avais à faire pour la première fois avec ce que l'on peut appeler un être humain. […] Quelque chose émanait de lui qui me mit à l'aise. Plus tard, en y repensant, j'ai compris pourquoi : il m'acceptait tel que j'étais »[53]. Roy Hart, après avoir poursuivi des études de littérature anglaise, d'histoire de la musique, de philosophie et de psychologie à l'université de Johannesburg, abandonna rapidement l'Académie Royale d'Art Dramatique de Londres, pour laquelle il avait obtenu une bourse, pour prolonger le travail d'Alfred Wolfsohn par la création du Roy Hart Théâtre. Ce théâtre tenait à la fois de la troupe au sens classique, du gymnase et de l'église. Les comédiens, chanteurs, ou curieux convaincus qui y prenaient part venaient y chercher beaucoup plus qu'un enseignement de technique vocale. «Face à une société éclatée, qui semble s'être éloignée de ses sources et dans laquelle l'individu est contraint d'adopter un masque en conflit avec son être profond, ces groupes conçoivent l'activité théâtrale comme instrument de régénération, individuelle et collective, comme moyen de libérer les « forces captives » —sans jamais pour autant renoncer à la maîtrise ni à la lucidité— et donnent pour cela la

[53] Roy Hart, "Let there be consciousness tonight and forever", interview de Jose Moleon et Ricardo Domenech à Madrid, *Malérargues*, p.1

préférence aux mythes et modèles archétypaux qui fondent notre civilisation plutôt qu'aux textes se référant aux situations concrètes de la vie. L'acteur, au terme d'un travail exigeant, de longue haleine, récupère et intègre des énergies jusqu'alors inconnues, s'ouvrant en cela à une démarche proche de la thérapie, sous les auspices du directeur-metteur en scène, souvent appelé à jouer le rôle de guide spirituel »[54].

Dans cette démarche spécifique, le Roy Hart Théâtre faisait de la voix sa «pierre philosophale », mettant en avant la conviction que le travail sur le corps et le travail sur la voix s'équivalent.

Roy Hart et les membres de sa troupe ont ainsi travaillé à dépasser l'étendue vocale ordinaire des deux octaves et demi et la spécialisation des voix du chant classique, en poussant la voix vers les sons extrêmes, graves et aigus. «[Ce n'était pas] par pure bravade, dans l'unique but de faire étalage de virtuosité. D'ailleurs, il serait erroné de voir derrière l'expression «voix de huit octaves », fréquemment utilisée par les membres du Roy Hart Théâtre, l'image quantitative d'un tout achevé, d'une sorte de géant sans faille ni faiblesse : il s'agit, au contraire, de la voix « sans fard », pleine de craquements, de fissures, de tremblements, souvent de la voix viscérale qui fait entendre les sons « cassés » ou « cordés », échappant aux nomenclatures classiques. Voix de celui qui ose s'aventurer hors du connu, voire de l'humain —puisque ces sons s'apparentent à des sons d'animaux […] ou de machines— ou plus précisément hors des sons humains connus de nos jours : rien ne dit qu'ils n'aient été présents en des temps reculés de l'humanité. En tout cas, après avoir atteint ces extrêmes et pris conscience de ces strates inconnues, réprimées ou inexprimées, tant dans le chant que dans la parole, l'individu tentera de relier ses nouvelles découvertes au déjà connu pour en faire une synthèse et s'approcher ainsi du juste milieu. Démarche étrangère au chanteur classique, lui que l'on a toujours encouragé à perfectionner sa tessiture, sans surtout sortir des limites de son registre confortable »[55].

Peter Oustinov, Yehudi Menuhin, Laurence Olivier, Aldous Huxley, John Cage, Harold Pinter fréquentèrent l' « Abraxas-Club », salle-studio où se tenaient les réunions/ateliers/thérapies organisées sous l'égide de Roy Hart. Après plusieurs années passées à donner des représentations dérangeantes à travers le monde, ne laissant jamais le public indifférent, remportant nombre de succès, les membres de la troupe du Roy Hart Théâtre se sont disséminés à la mort de son fondateur, en 1975, pour essaimer et reprendre

---

[54] Marianne Ginsbourger, *Voix de l'inouï – Le travail de la voix au Roy Hart Théâtre hier et aujourd'hui*, Le souffle d'or, collection chrysalide, Barret-le-Bas, 1996, p.13
[55] Marianne Ginsbourger, *op. cit.*, p. 46

sa démarche à travers la création de nouvelles troupes, ou en intervenant auprès de troupes existantes.

L'un des prolongateurs du Roy Hart Théâtre s'appelle Enrique Pardo. «Si Enrique Pardo a quitté l'enseignement des beaux-arts pour rejoindre le groupe [du Roy Hart Théâtre] en 1970, c'est qu'il pensait réfléchir avec lui aux liens entre la création artistique et la psyché ; il n'est venu à la voix et au théâtre qu' « incidemment ». Son travail vocal avec Roy Hart et Liza Mayer, le travail collectif sur les rêves lui ont permis d'avancer dans cette quête philosophique et artistique mais aussi sur un chemin personnel. En 1981, il crée le solo Hymne au Dieu Pan —Pan étant le dieu chanteur-danseur— puis la compagnie Panthéâtre. Il mène avec celle-ci une recherche —dans une perspective mythique ou archétypale, nourrie notamment par les travaux de James Hillman et Rafael Lopez Pedraza— sur la relation entre la voix et le mouvement, entre le chant et la danse, qui traverse ses activités de metteur en scène, d'acteur, de pédagogue et d'écrivain, spécialiste en mythologie gréco-romaine »[56].

Nous l'avons rencontré au Théâtre National de Chaillot, où il intervenait auprès de la troupe de Philippe Genty, environ 9 mois avant la première représentation d'un nouveau spectacle, soit largement avant que les comédiens ne commencent même à se préoccuper de texte ou de mise en scène.

Enrique Pardo, «aventurier de la voix », pour s'inscrire dans le prolongement de la démarche du Roy Hart Théâtre, s'en démarque sensiblement. Ainsi le pendant quasi psychothérapeutique du travail vocal ne représente pas pour lui un objectif, et il affiche une prudence mesurée à l'égard de la «légende communautaire » du groupe des années 70. Ce qu'il apporte aux comédiens, c'est avant tout une autre manière de découvrir leurs capacités vocales, une façon de libérer leur voix pour faire vivre leurs interprétations avec plus d'ambitus, plus de souplesse. Pour cela, bien sûr, le corps est une ressource essentielle, et le jeu inextricable du corps et de la voix doit permettre de susciter l'émotion, en même temps que celle-ci doit habiter le comédien qui s'exprime à haute voix. Pour autant, Enrique Pardo s'oppose clairement à une « idolâtrie de la voix », selon laquelle l'expression vocale et l'affect fusionneraient nécessairement, incitant à vouloir à tout prix ajuster les colorations vocales aux affects que l'on cherche à exprimer. Au contraire, il préfère la notion d' « emoting », autrement dit la capacité de se rendre émotif et de le montrer. Dans sa boîte à outils pédagogique, il lui arrive d'imposer pour cela aux comédiens la contrainte de « non

[56] Marianne Ginsbourger, *op. cit.*, p. 125

illustration » : le son de la voix ne doit pas coller à l'affect issu du texte, mais évoquer au contraire autre chose, de façon à faire naître des métaphores d'émotions décalées, en lieu et place de l'illustration banale.

Dans ce travail, les sons « brisés », « cassés », « rugueux » tiennent une place importante, puisqu'ils sont les frontières d'un terrain sonore rarement exploré d'ordinaire. Enrique Pardo lui-même utilise les sons « cordés », qui sont l'avatar amélodique du chant diphonique : une seule voix produit deux, trois, quatre…harmoniques distinctement audibles. A la différence du chant des Touvas, la qualité mélodique n'est cependant pas, pour le théâtre, une priorité.

Cette recherche d'inouï n'est pas sans résonance avec la conviction, dans la démarche de la troupe Panthéâtre de façon générale, que la pédagogie vocale est une casuistique. Plutôt que de fournir aux comédiens des « modèles émotionnants », il est largement préférable de leur faire découvrir par l'expérimentation l'étendue de leur propre bibliothèque vocale, susceptible de s'augmenter continûment de nouveaux rayonnages.

Lors de l'atelier vocal auquel nous avons assisté au Théâtre National de Chaillot, la réaction des six comédiens présents à leurs propres progrès s'est avérée significative de la réelle nouveauté de ce que leur proposait leur pédagogue vocal.

Le principe de la « leçon » —qui n'en était , justement, pas une au sens classique— consistait, de façon général, à choisir un leitmotiv sonore –en l'occurrence « SANTARCO »- ou simplement une succession de syllabes, et de les faire évoluer au gré des progressions tonales proposées par Enrique Pardo au piano. L'exercice se rapprochait des vocalises rituelles des écoles de chant « orthodoxes ». Les interventions du fondateur de Panthéâtre changeaient cependant radicalement la donne. Il déstabilisait les certitudes et habitudes, inconscientes ou formatées par une pratique antérieure encadrée de la voix, des comédiens et comédiennes, par des injections inopinées ou en se mettant en scène lui-même, par la voix et par le corps. Et cela, au moment où, souvent, les « élèves » parvenaient aux limites « ordinaires » de leur voix : « Secoue ta voix », « laisse-là s'abîmer », « amplifie au niveau théâtral l'animation de ce son-là : déglingue le », « ne t 'enferme pas dans un couloir, joue avec le volume, même, pleure », «J'aimerais que tu casses un peu la vaisselle, pour voir comment tu commences à approcher l'idée de son cassé »…

Le rôle du corps était prépondérant. A la statique préconisée dans le chant classique, les comédiens devaient, par exemple, préférer jouer de leur voix en se secouant les épaules, en faisant rouler leur nuque, ou en s'appuyant mutuellement sur le torse pour sentir les vibrations sonores dans le corps de l'autre et ajuster deux voix sous les encouragements

d'Enrique Pardo : « Enjoy each other presence, pour une rencontre de la texture, du timbre vocal », « I encourage you to sing to each other ».

A la nouveauté des sons découverts, re-découverts par les comédiens, s'ajoutait l'humour régnant entre tous les participants et l'entière liberté offerte à leur expérimentation.

Nous pourrions décrire plus précisément les différentes étapes du travail d'Enrique Pardo, isoler les différences observées dans l'attitude de chacun des comédiens… La vie d'un tel moment est cependant difficile à donner à entendre…

Nous retiendrons en revanche, en particulier, que si les aspects «techniques » de la voix n'étaient pas oubliés (les notions de souffle, de détente ou de contraction laryngée notamment), l'attrait et l'apport majoritaire de l'atelier résidait essentiellement dans sa logique de produire des sons hors cadre, hors définition. Libérée de schémas artistiques figées comme des attributs émotifs conventionnels, la voix s'échappe ainsi vers ce que seul l'individu qui la produit veut donner à entendre et, même, laisse entendre spontanément en s'abandonnant à une absence de contrôle sur lui-même.

Nous croyons, dans cette dernière partie, avoir rapproché légitimement les causes finales qui président à la synthèse vocale artificielle et la façon contemporaine qu'a la voix humaine naturelle d'être au centre de l'activité théâtrale, en les réunissant dans la recherche d'un inouï vocal.

Ainsi, qu'il s'agisse des progrès enregistrés dans sa réalisation technologique ou de l'investigation organique que se propose l'homme préoccupé de création, la voix échappe à tout cadre censé la définir, se défie des paramètres, s'affranchit des références du beau et du vulgaire.

La science informatique parvient, au mieux, à en faire d'impossibles chimères, pendant que la connaissance empirique pédagogique, dans ses embranchements les plus novateurs, cherche à la libérer des dizaines d'années d'un perfectionnement technique désormais vécu comme une impasse.

Nous n'avons , au fil de notre progression, cessé d'observer des parallèles entre la tentative de saisir la voix du plus loin possible de ce qui la lie à la nature profonde de l'homme —en considérant typiquement la synthèse artificielle vocale— et celle de la cerner au contraire au plus près de sa nature organique individuelle —via la démarche du Roy Hart Théâtre. Malgré certaines similitudes entre ces deux démarches extrêmes, la voix, universel anthropologique, se refuse à toute définition unificatrice. Et pour cause !

S’il est vrai qu’elle est propre à l’Homme, elle demeure, surtout l’apanage exclusif du sujet, autrement dit de l’individu. Et si la communauté des hommes existe, biologiquement au moins, il semble clair que celle des individus relève de l’utopie.

Conclusion

Vouloir maîtriser la voix de façon exhaustive relève de l'utopie. Quand on la considère comme une composante purement biologique de l'humain, elle ne se laisse conformer par aucun «entraînement » infaillible. Dès lors qu'on la pense comme un bras ou un cœur dont on voudrait, à force de répétitions et de précision dans les exercices, améliorer les performances, elle montre, en dépit d'un œrtaine marge de progrès, qu'elle dépend avant tout de la motivation nerveuse, psychique, de l'individu à qui pourtant, organiquement, elle appartient.

Si on essaie, pour mieux la vaincre, de la remplacer par la machine, le même problème resurgit. Sa manifestation sonore, loin de traduire simplement les mouvements des organes qui la créent, s'avère intraduisible, irréductible. Emotions, affects, pulsions la conditionnent anarchiquement.

Nous avons vu pourtant que, justement, une ébauche de compréhension de ces mécanismes qui n'en sont pas, un embryon d'interprétation des mouvances internes au cerveau qui, en définitive, fait la voix, semblait se faire jour. Reste à savoir si la lumière aperçue au fond du tunnel de la découverte est faible parce que le chemin est encore long, ou s'il ne s'agit, en fait, que d'un mirage.

Nous penchons, à la fin de ce travail, pour l'hypothèse de l'illusion. Comment pourrait-on en effet comprendre, en construisant des théories interprétatives, ou réaliser, en inventant un système de synthèse vocal «idéal », «vivant », ce qui nous échappe déjà en tant qu'individu ?

Car la voix, en définitive, n'existe jamais de façon unique. Elle revêt une forme, acoustique et vibratoire, pour l'un, qui l'émet, et une autre, souvent très différente, pour l'autre, qui l'entend. Autrement dit, elle se manifeste pour tous mais n'est reconnaissable par personne.

Dès lors, la quête d'un «absolu vocal », qu'elle soit celle de l'interprète avide d'un son parfait, ou celle du chercheur spécialiste de la synthèse vocale imprégné du mythe de Frankenstein, désireux de re-créer le vivant, n'a pas de sens.

Heureusement, il n'est pas besoin d'absolu pour susciter l'effort. La voix, parce qu'elle est essentiellement inobjectivable, parce que le sujet l'aliène autant qu'elle le libère, concentre d'une façon rare les travaux conjoints des sciences « dures » et « sociales ». L'utopique désir de saisir une fois pour toute ce phénomène sonore singulier mais universel les rassemble. La voix, paradoxalement, nécessite, en tant précisément que simili-objet, les supports théoriques et expérimentaux de deux façons d'envisager la science qui, souvent,

tendent à s'ignorer. Cet état de fait n'est d'ailleurs pas sans liaison avec les récents rapprochements entre psychiatrie et « neurosciences », dans la même improbable quête de localisation –d'objectivation ?- de la conscience ou de l'intelligence. Car, si un jour une voix synthétique réussissait à « tenir conversation », se faisant passer pour celle d'un homme de façon indiscernable, c'est que l'intelligence artificielle aurait vu le jour.

Bibliographie

?? C. Baber and J. M. Noyes, *Interactive speech technology. Human Factors issues in the application of speech Input/Output to computers*, Taylors & Francis Ltd, 1993

?? S. Bernhardt, *L'art du théâtre. La voix, le geste, la prononciation*, L'Harmattan, coll. « Les introuvables », Paris, 1993

?? G. Cornut, *La voix*, PUF, coll. « Que sais-je », Paris, 1998

?? G. Dessons, *Penser la voix – Chant – Communication - Linguistique clinique – Littérature – Musique – Peinture – Psychanalyse – Théâtre*, Editions La Licorne, UFR Langues Littératures, Poitiers, 1997

?? Y. Fonagý, *La vive voix, essai de psycho-phonétique*, Editions Payot, Paris, 1991

?? T. Georgaki, *Problèmes techniques et enjeux esthétiques de la voix de synthèse dans la recherche et création musicales*, thèse EHESS/IRCAM/CID-CNRS, sous la direction de H. Dufourt, Paris, 1998

?? J-S. Liénard, *Les processus de la communication parlée: introduction à l'analyse et la synthèse de la parole*, Editions Masson, Paris, 1977

?? W. Minker, *Parole et dialogue homme-machine*, Editions du CNRS, coll. « Sciences et Techniques de l'Ingénieur », Paris, 2001

?? J. et C. OTT, *La pédagogie et les techniques européennes du chant*, Editions EAP, Issy les Moulineaux, 1994

?? G. Robert, « Une chaîne à contre-courant : FIP 514 », *Cahiers d'Histoire de la Radiodiffusion*, , n°70, 2001

?? A. Tomatis, *L'oreille et le langage*, Editions du Seuil, coll. « Le rayon de la science », Paris, 1963

?? *Pour la Science*, Dossier « Le monde des sons », n°32, juillet-octobre 2001