José Morais, *L'art de lire*, Paris, Éd. Odile Jacob, 1994. 358 p.
Extrait du chapitre 1 - *Le langage et l'alphabet*, pages 85 à 90.

Qu'est-ce que la parole ? Nous articulons des mots par un jeu très complexe de combinaisons de mouvements des mâchoires, des joues, des lèvres, de la langue, du vélum et du larynx. Ces mouvements permettent de transformer en son de parole le souffle d'air qui provient des poumons et qui est mis en vibration par les cordes vocales. Le son de parole peut être visualisé par un sonagramme, qui est une représentation de l'information acoustique. Seules des personnes hautement entraînées pourraient identifier les mots représentés dans les sonagrammes. Avant les années quarante, on croyait que les phonèmes ont une réalité acoustique et qu'il serait possible, en les rendant visibles, de produire des signaux permettant aux sourds profonds de " lire " la parole et de pouvoir ainsi utiliser le téléphone Les premiers sonagrammes, obtenus aux laboratoires de la Bell Telephone, ont sonné le glas de ce rêve. On peut voir ci-dessous, par exemple, les sonagrammes des mots " chimpanzé " et " camembert " [*voir figure 1.14*].

Le déroulement temporel du son est donné en abscisse [l'axe horizontal], et la fréquence acoustique est représentée en ordonnée [l'axe vertical]. Les variations d'intensité du signal sont grossièrement représentées par le noircissement du papier. Les bandes horizontales d'ombre épaisse s'appellent les formants. Quelle est l'origine des formants ? La cavité qui commence au larynx et s'ouvre à l'extérieur au niveau des lèvres constitue une chambre de résonance d'une forme complexe. L'air dans ce conduit est mis en vibration d'une façon quasi périodique Le taux de vibration des cordes vocales détermine la fréquence fondamentale, ou hauteur tonale de la voix.

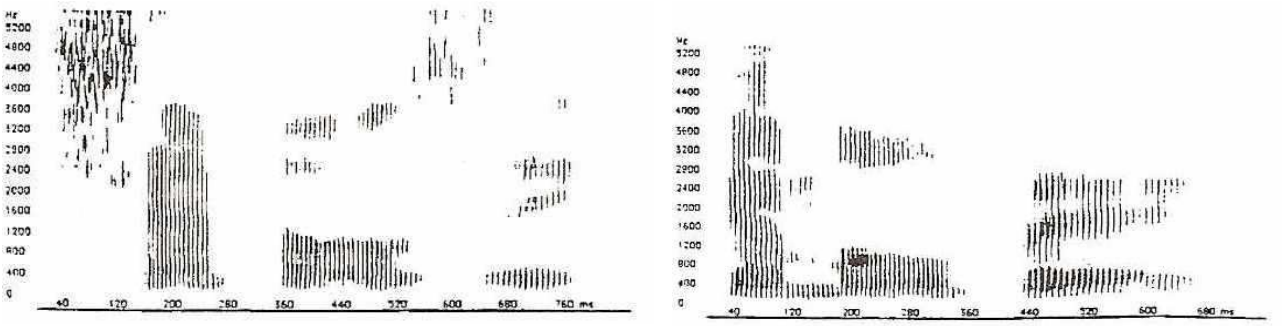

Figure 1.14 –
Sonagrammes des mots « chimpanzé » (à gauche) et « camembert » (à droite).

C'est le formant le plus bas (on le désigne par F zéro), il n'est pas apparent sur les sonagrammes reproduits. Les autres formants sont liés à la forme particulière que prend le conduit vocal pendant l'émission du son. On les appelle premier, deuxième, etc., à partir du bas.

« Dans " chimpanzé ", la voyelle nasalisée " im " ([ɛ̃]) est centrée autour de 200 msec à partir du début du signal, la voyelle nasalisée " an " ([ã]) commence vers 360 msec et s'étend jusqu'au-delà de 500 msec, et la voyelle finale " é " ([e]) est visible entre 680 et 760 msec. Quand nous parlons, la forme du conduit vocal change presque continuellement, de telle sorte que les formants sont rarement stables. Ils peuvent présenter des montées et des descentes, au début et/ou à la fin. C'est particulièrement apparent pour le deuxième formant du [a] de " camembert ", qui est descendant du début à la fin. Par ailleurs, le sonagramme montre aussi des bruits, des sons non périodiques, tels que le bruit du prévoisement qui précède la consonne [b] de " camembert " (de 330 à 440 msec), et le bruit (provoqué par le resserrement du conduit vocal) de la friction correspondant aux consonnes [ʃ] et [z], de " chimpanzé ".

Lorsque nous, individus lettrés, écoutons de la parole, nous avons l'impression d'entendre une suite de sons élémentaires, appelés phones ou segments phonétiques. Dans le mot "camembert",

par exemple, nous avons l'impression d'entendre d'abord [k], puis [a], puis [m], etc., et nous nous disons que le locuteur les a prononcés dans le même ordre. Pourtant, c'est faux. Le travail pionnier de Alvin Liberman et de ses collègues des Laboratoires Haskins, dans les années soixante du XX^{e} siècle, l'a démontré très clairement[1]. Prenons des sons de parole artificielle, synthétisée, qui sont perçus par les auditeurs comme étant les syllabes [di] et [du] [*voir figure 1.15*]. Si l'on ne fait écouter que les parties stables des formants, on a l'impression d'entendre respectivement [i] et [u].

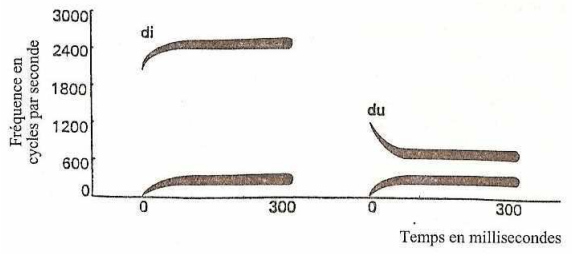

Figure 1.15 —
Sonagrammes de patrons acoustiques suffisants pour la synthèse de [di] et de [du][2].

Maintenant, supprimons cette partie stable et écoutons seulement les parties montantes et descendantes des formants, qu'on appelle des transitions de formant. On espère entendre [d] dans les deux cas, et pourtant tout ce qu'on entend est une espèce de grésillement.

On pouvait se douter que quelque chose d'étrange se produirait, puisqu'on a l'impression d'entendre le même [d] dans [di] et dans [du], et pourtant la forme physique du son au début présente une grande différence : la transition du deuxième formant est montante dans le cas du [di], mais descendante dans le cas du [du]. Comment, avec des sons si différents, aurait-on pu entendre le même phone [*l'unité perceptive la plus petite*] ? On pourrait se dire : on a trop coupé, laissons un peu de la partie stable pour essayer d'entendre le [d]. Cependant, si l'on écoute les transitions avec un peu de la partie stable, on n'obtient pas plus de succès. On entend bien le [d], mais dans chaque cas avec quelque chose de plus, c'est-à-dire on entend toujours [di] et [du], les voyelles paraissant maintenant beaucoup plus brèves. Comme on est patient, on essaie encore une fois en gardant un peu moins de partie stable. Peine perdue ! On a beau aller vers la droite, aller vers la gauche, on n'entend jamais ce qu'on espérait entendre, un beau (même un vilain) [d], rien d'autre qu'un [d]. C'est comme si le [d] n'existait pas ! On ne peut pas prononcer un [d] isolément. Nos efforts pour le produire sans y ajouter une voyelle sont inexorablement voués à l'échec. Lorsque nous essayons de prononcer la valeur phonémique de la lettre " d ", ce que nous prononçons est une syllabe ([də]), dans laquelle la voyelle contient peu d'énergie acoustique.

Pourquoi la consonne n'apparaît-elle pas en tant que telle dans le signal acoustique ? Ce qui se passe, c'est que le locuteur prépare l'articulation de la consonne et de la voyelle en même temps Les mouvements articulatoires nécessaires pour produire l'une et l'autre se combinent, et évidemment leurs effets acoustiques se combinent aussi, de telle sorte que la partie initiale du signal acoustique reflète les deux phones. Dès que l'on fait écouter assez de signal à partir du début pour pouvoir entendre la consonne, et non pas un simple grésillement, on entend aussi la voyelle. En d'autres termes, ce qui est perçu comme étant la même consonne présente, dans des contextes vocaliques différents, de très grandes différences du point de vue acoustique. On dit, en conséquence, que les consonnes, surtout les consonnes occlusives, présentent un haut degré d'encodage dans le courant acoustique de la parole. Puisque l'expression acoustique de la consonne dépend de la voyelle, on dit aussi qu'elle manque d'invariance acoustique.

Le manque d'invariance acoustique est une conséquence de la co-articulation des phones que nous avons l'impression d'entendre en succession, mais que le locuteur articule simultanément, et

[1] A.M. Liberman, F.S. Cooper, D. Shankweiler et M. Studdert-Kennedy, «Perception of the Speech Code», *Psychological Review*, 1967, 24, 431-461.
[2] A.M. Liberman & *alii*, *op. cit.* Reproduction autorisée par l'American Psychological Association.

que très vraisemblablement notre système perceptif traite aussi simultanément. Derrière le manque d'invariance acoustique il y a, pourtant, une constance articulatoire. Considérons toutes les syllabes qui commencent par [d]. On verra, en mettant côte à côte les différents sonagrammes, que les transitions du deuxième formant pointent toutes vers un point en arrière dans le temps [*figure 1.16*]. Cette fréquence unique, qui correspond à environ 1 800 cycles par seconde, peut être appelée le " site " du deuxième formant du [d]. Pour les autres consonnes occlusives qui diffèrent du [d] par le lieu d'articulation (c'est-à-dire l'endroit où le conduit vocal est fermé), le site du deuxième formant est aussi différent.

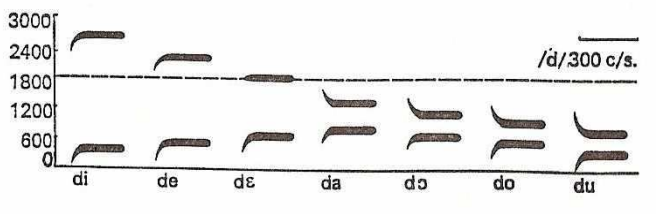

Figure 1.16 –
Sonagrammes de patrons acoustiques suffisants pour la synthèse des syllabes consonne-voyelle initiées par [d]
(Cf. note 1, reproduction autorisée par l'American Psychological Association).

Pour le [b], le site est plus bas dans l'échelle de fréquences, et pour le [g] il est plus haut. C'est normal. Pour prononcer un [b], on ferme le conduit au niveau des lèvres, et par conséquent celui-ci est plus long et la fréquence de résonance est plus grave, que si on le ferme au niveau des dents (pour produire un [d]) ou au niveau du vélum (pour produire un [g]).

La position du site nous informe donc sur le lieu d'articulation. Le site lui-même n'est pas représenté dans le sonagramme, parce qu'avant qu'il n'y ait un certain degré d'ouverture du conduit vocal il n'y a pas d'émission de son. C'est pourquoi, si on juxtaposait toutes les transitions du deuxième formant associées à une même consonne, on verrait que, bien qu'elles pointent toutes vers un même point, elles ne convergent pas. Si on créait des sons où la transition commencerait à partir du site, on risquerait de produire une autre consonne, car le système perceptif serait trompé : lui sait qu'entre le site et le début de la transition il y a un temps de silence.

Les phones ne sont donc pas des segments acoustiques. Notre système perceptif doit faire des calculs savants pour les extraire du courant acoustique. L'auditeur a l'impression de percevoir chaque mot que le locuteur prononce, à peu près à l'instant où celui-ci le prononce, et pourtant entre l'instant où le locuteur prononce un mot et l'instant où l'auditeur le perçoit il y a un grand nombre d'opérations mentales. Ces opérations mentales analysent la représentation acoustique, qui correspond grosso modo à ce que l'on voit dans un sonagramme, à partir des connaissances que l'auditeur (ou son système de perception de la parole) possède sur les effets acoustiques de la co-articulation des phones. Autrement dit, les phones sont liés aux sons par un code spécifique qui renvoie l'auditeur aux conditions de production de la parole.

Peu de gens ont jamais pensé qu'ils ont dans leur tête un dispositif qui interprète les sons de parole entendus en se référant aux conditions de production de ces sons. Pourtant, nous devons tous l'avoir. Pour nous en convaincre, prenons un son tel que celui de la syllabe [sa], comme quand nous disons " c'est quoi, ça ? ". On peut le digitaliser dans notre ordinateur, et introduire 50 msec de silence entre la friction du [s] et le début des transitions vers la voyelle [a] [*voir figure 1.17*]. On pourrait imaginer entendre [s] — un silence — [a] En fait, non ! Nous entendons [sta]. Donc, nous introduisons un silence, et nous entendons une consonne. Merveilleux !

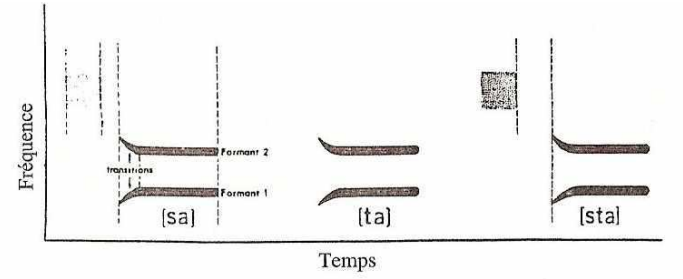

Figure 1.17
Sonagrammes de patrons suffisants pour la synthèse de [sa] et de [ta]. L'introduction d'un intervalle de silence entre la friction et la voyelle crée la syllabe [sta][3]

Ce ne sont pas des voix off et nous n'avons pas d'hallucination. Il se passe, tout simplement, que normalement le silence apporte de l'information sur le geste articulatoire de fermeture du conduit vocal qui caractérise la production de toute consonne occlusive. Notre système perceptif, qui connaît les conditions de production de la parole, nous dit qu'il faut percevoir une occlusive entre le [s] et le [a]. Pourquoi [t] et pas [p] ou [k] ? Parce que la dentale est plus cohérente avec les transitions présentes dans la syllabe [sa].

[3] A.M. Liberman et M. Studdert-Kennedy, « Phonetic Perception », dans R. Held, H. Leibowitz et H.-L. Teuber, *Handbook of Sensory Physiology, vol. VIII, Perception*, Heidelberg, Spring-Verlag, 1978. Reproduction autorisée par Springer-Verlag.