Variabilité phonétique en production et perception de parole : stratégies individuelles

René Carré, ENST-CNRS, 46 rue Barrault, 75634 Paris cedex 13, carre@tsi.enst.fr

Jean-Marie Hombert, Laboratoire Dynamique du Langage, CNRS et Université Lyon 2, Institut des Sciences de l'Homme, 14 avenue Berthelot, 69363 Lyon cedex 07, hombert@univ-lyon2.fr

Introduction

Les caractéristiques des sons de parole portent des informations aussi bien sur les catégories linguistiques que sur l'identité des locuteurs et sur leurs états émotionnels. Mais, à ce jour, la grande majorité des études consacrées aux caractéristiques phonétiques vise à gommer les variations inter-individuelles en considérant que seules les moyennes obtenues à partir de données sur de nombreux sujets sont représentatives des caractéristiques linguistiques d'une population. La variabilité inter et intra-locuteur n'est pas considérée par la plupart des chercheurs comme pouvant participer à l'explication des mécanismes fondamentaux de la communication parlée. On considère plutôt cette variabilité comme un inconvénient qu'il faut limiter afin, sur le plan théorique, d'obtenir une représentation d'un système phonologique homogène pour l'ensemble de la communauté et dans le domaine des applications pour permettre, par exemple, par élimination des variabilités, un bon fonctionnement des systèmes de reconnaissance. Comme conséquence de cette position, les données disponibles aujourd'hui (qui doivent être statistiquement 'significatives') sur les sons de parole sont des moyennes, moyennes de formants pour les voyelles par exemple, moyennes obtenues à partir d'au moins un dizaine de locuteurs ou bien moyennes sur les limites perceptuelles entre catégories (sur les VOT[1] des occlusives par exemple). On dispose donc de données

[1] VOT : Voice Onset Time. Durée entre l'instant d'ouverture après occlusion lors de la production des consonnes occlusives (comme /b, d, g, p, t, k/) et l'instant de mise en fonctionnement des cordes vocales. Cette durée peut être positive ou négative. Elle dépend de la consonne produite et du locuteur.

avec des écarts-types assez grands, lesquels sont dus, en grande partie, aux différences inter-locuteurs. Dans le cas d'études très rares de type mono-locuteur, les écarts-types sont naturellement beaucoup plus réduits.

Dans notre étude, nous adoptons une position radicalement différente : nous mettons l'accent sur les caractéristiques individuelles (en production et en perception) de chacun des locuteurs et sur leurs rôles dans le processus de communication. Nous considérons que l'existence et le traitement des variabilités sont à la base des mécanismes de communication entre sujets possédant des caractéristiques phonétiques (en production et en perception) différentes : la prise en compte des caractéristiques et des stratégies individuelles de production et de perception est essentielle car ces caractéristiques participent directement et globalement aux processus de communication dans ses dimensions variées intégrant aussi bien l'acquisition que le changement phonologique et l'évolution (Hombert, 1984). Hombert et Puech (1984) ont montré avec 4 sujets Swahili de la même zone dialectale que les plages d'identification des voyelles dans le plan F1/F2 varient notablement d'un sujet à un autre et donc que leurs stratégies de perception varient (figure 1, pour 2 sujets). Ils émettent l'hypothèse que ces variations peuvent être liées à leur histoire linguistique. La variabilité est certainement intrinsèque à la parole elle-même pour permettre la communication entre des personnes différentes, ayant des systèmes de production et de perception différents. La variabilité est partout présente en communication parlée : variabilité inter et intra-loculeurs (portant des informations sur l'origine géographique du locuteur, ses caractéristiques propres, son état émotionnel,…) de type acoustico-phonétique, phonologique, lexical, sémantique, dans le dialogue… On se bornera ici à étudier la variabilité acoustico-phonétique en liaison avec la phonologie.

Par ailleurs, dans notre démarche, on considère que les représentations phonologiques de la parole correspondent non seulement à des segments de caractéristiques statiques mais aussi à des « gestes » au sens de la phonologie articulatoire (Browman and Goldstein, 1989). Par geste, on entend des transitions c'est à dire des événements qui se déroulent dans le temps avec des caractéristiques dynamiques contrairement aux « segments » qui, eux, sont généralement associés à des caractéristiques fixes sur une plage de temps donnée. Chacune de ces représentations phonologiques est produite et perçue selon chacun des individus qui ont des caractéristiques de production et de perception propres.

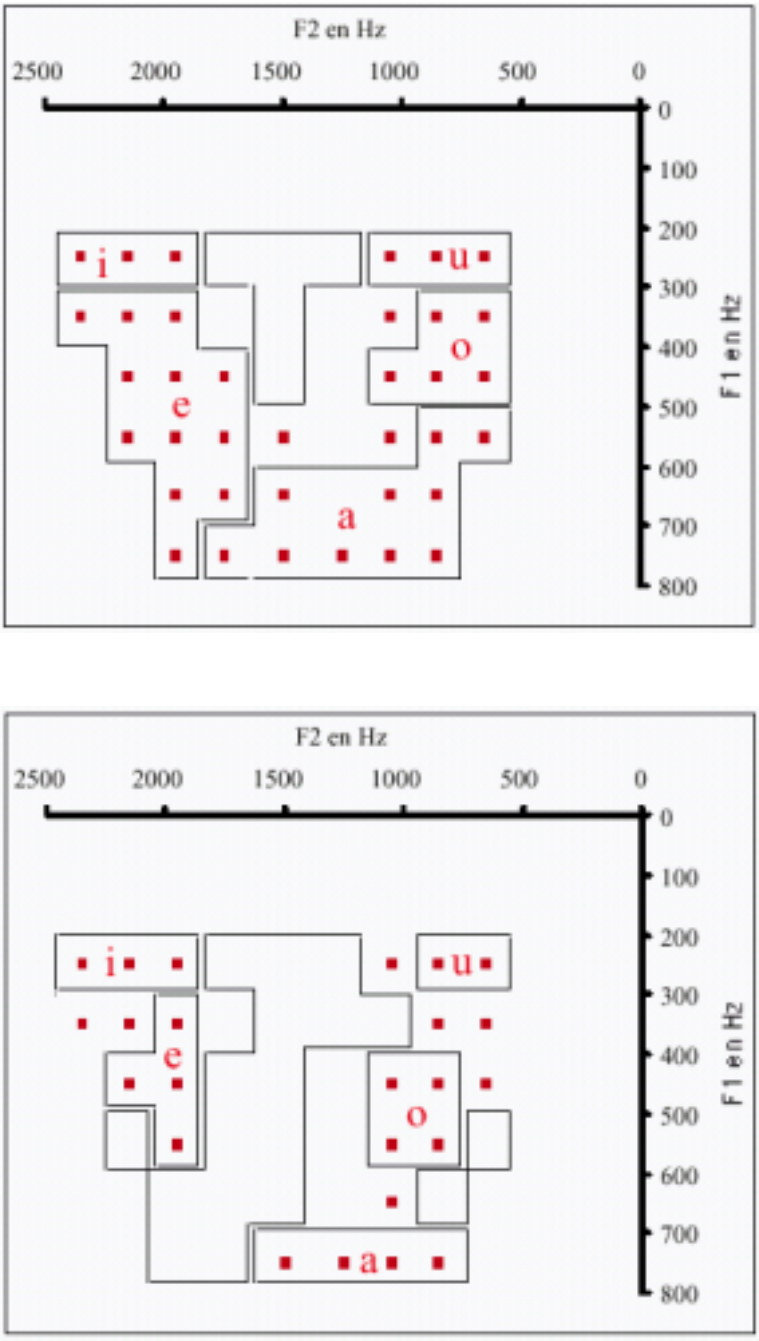


Figure 1. Plages d'identification des voyelles /u, i, e, a, o/ dans le plan F1/F2 pour 2 sujets Swahili de même zone dialectale. Les régions vides sont considérées par les sujets comme hors du système vocalique. On note une grande variabilité inter-locuteur en ce qui concerne les frontières entre voyelles.

La variabilité en production et en perception de parole

Rappelons quelques données sur des paramètres variés qui montrent que les plages de variation inter-locuteur en production sont considérables :

- le premier exemple classique est celui de la plage de variation des fréquences des formants représentés dans le plan F1-F2 pour les voyelles de l'anglais américain prononcées par différents locuteurs (Peterson and Barney, 1952) ; les plages correspondant aux voyelles se chevauchent fortement, en particulier lorsque l'on compare les données provenant de production d'hommes, de femmes et d'enfants. Les différences inter-locuteurs sont encore plus grandes si l'on s'intéresse aux formants F3, F4,… à la fréquence fondamentale, à l'intensité.
- Mais, si les positions dans le triangle vocalique des voyelles produites par plusieurs individus varient, en revanche, les positions des voyelles perçues par ces mêmes individus comme étant les plus 'prototypiques' sont plus stables et correspondent à un triangle vocalique de dimension quasi-maximale : les meilleurs prototypes perceptifs correspondent à des voyelles de production 'hyperarticulées'[2] (Johnson, et al., 1993).
- un autre exemple concerne les productions de type CV. Il montre que l'équation du locus[3], dépend du locuteur, de la langue,... (Sussman, et al., 1991). Rappelons que l'équation du locus correspond à une droite représentative des fréquences de début de transition du 2ème formant pour les productions CV (avec C étant une consonne donnée et V prenant les valeurs des différentes voyelles de la langue) en fonction de la valeur du 2ème formant pour les voyelles V considérées. Cette droite dépend du locuteur.

Les précédentes études ont été effectuées sur des données statiques comme les formants des voyelles ou bien les valeurs des fréquences de début de transition du deuxième formant pour le calcul de l'équation du locus. En ce qui concerne les

---

[2] Par voyelles hyperarticulées, on entend des voyelles prononcées lentement, le plus clairement possibles.

[3] Le terme locus a été utilisé pour la première fois par Delattre et al. Delattre, P. C., Liberman, A. M. and Cooper, F. S. (**1955**). "Acoustic loci and transitional cues for consonants," J. Acoust. Soc. Am. **27**, 769-773.. Ceux-ci ont obtenu par la synthèse des ensembles Consones-Voyelles, les consonnes /b, d,

caractéristiques dynamiques tant en production qu'en perception qui seraient dépendantes du locuteur, de son type de voix, de ses émotions, la situation est encore plus complexe. Comment caractériser, modéliser, les aspects dynamiques relatifs à un locuteur donné ? Comment les aspects dynamiques sont-ils perçus par un sujet donné ? On sait malgré tout que les aspects dynamiques sont importants :

- dans les mécanismes de perception des voyelles, pour le cas des voyelles 'réduites'[4] (Lindblom and Studdert-Kennedy, 1967), alors que les cibles acoustiques prototypiques des voyelles (lorsqu'elles sont produites isolément) ne sont pas atteintes dans des transitions CVC par exemple, la voyelle perçue est celle correspondant à l'intention du locuteur ;
- la perception des voyelles dépend de l'environnement consonantique (Strange, et al., 1976). On peut supprimer les parties du signal correspondant à la cible vocalique tout en continuant à la percevoir ;
- l'étude des mécanismes de perception des plosives montre l'importance de l'ensemble des transitions (Dorman, et al., 1977), (Kewley-Port, et al., 1983).

Si les variations inter-locuteurs sont grandes, les variations intra-locuteurs ne sont pas négligeables mais elles sont peu étudiées (Liénard, 1999 ; Liénard and Di Benedetto, 1999). Ces variations sont en particulier caractérisées par la fréquence fondamentale, l'intensité, les durées (débit), F3, F4, à coté de paramètres plus traditionnellement exploités comme F1, F2.

Nous avons reproduit l'expérience de Johnson pour le Français et développé de nouvelles expériences pour étudier les stratégies individuelles de production et de perception de parole.

Expériences

1. L'expérience de Johnson a été reproduite avec 5 locuteurs français masculins (Hombert and Carré, 1999) prononçant les 5 voyelles du français (/i/ comme dans 'lit', /e/ comme dans 'les', /a/ comme dans 'la', /o/ comme dans 'l'eau', /u/

---

g/ étant obtenues chacune avec des transitions du deuxième formant partant d'un même point virtuel (caractéristique du lieu d'articulation) qu'ils ont appelé locus.

[4] Généralement, en parole spontanée, les cibles articulatoires et acoustiques des voyelles ne sont pas atteintes. On parle alors de voyelles réduites ou bien de réduction des voyelles. Les cibles qui sont des 'prototypes' sont généralement obtenues en parole hyperarticulée.

comme dans ‘loup’, /y/ comme dans ‘lu’). On a représenté figure 2 dans le plan F1-F2 les voyelles produites par les locuteurs jh et fp.

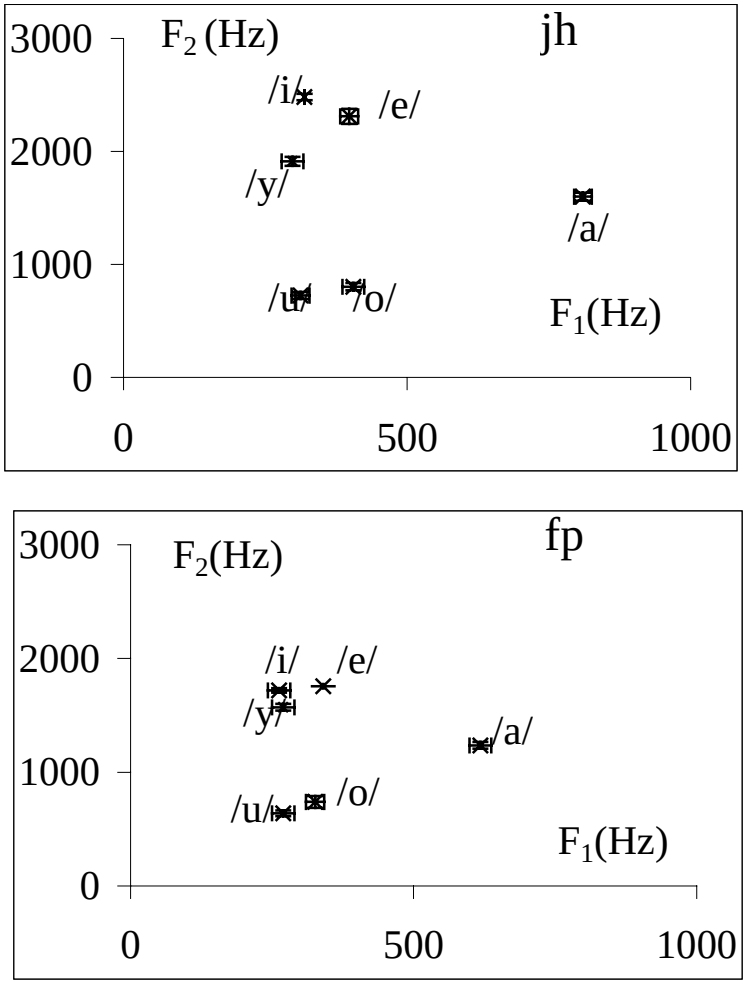


Figure 2. Représentation dans le plan F1/F2 des fréquences des formants (avec les écarts types) des voyelles produites par les locuteurs jh et fp. Le triangle vocalique du locuteur fp est nettement plus petit que celui du locuteur jh.

On voit ici clairement la grande variabilité obtenue. Le triangle vocalique du locuteur fp est beaucoup plus petit que celui de jh.

On a ensuite demandé aux mêmes sujets d’ajuster les fréquences des formants F1 et F2 d’un synthétiseur pour percevoir les meilleures voyelles précédentes. On a représenté figure 3 dans le plan F1/F2 les résultats obtenus.

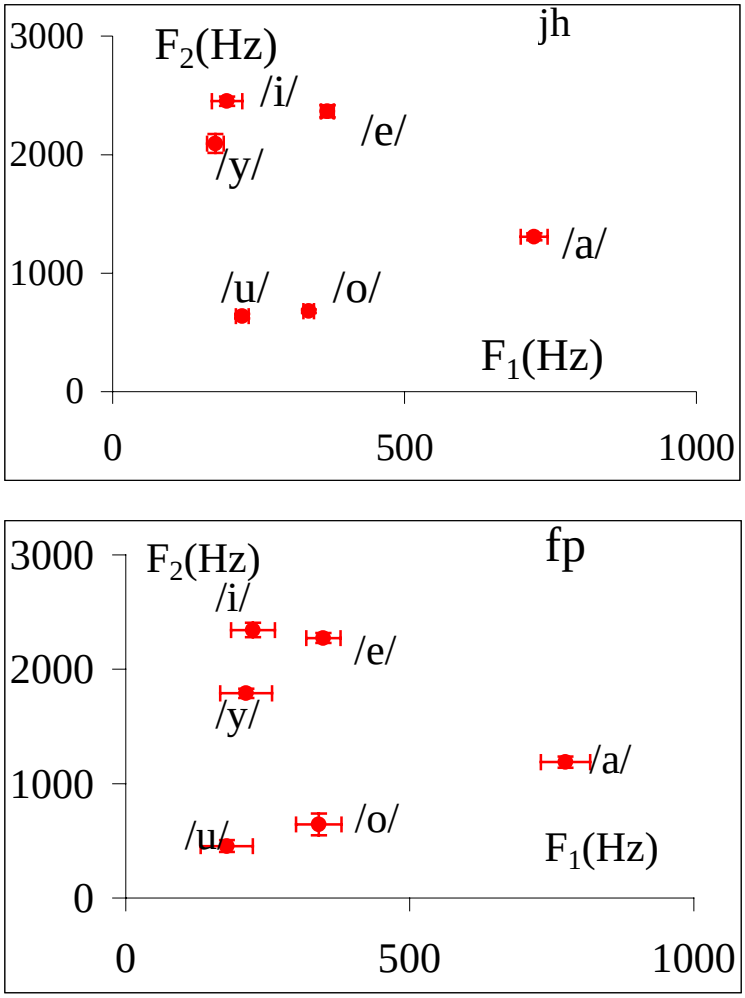
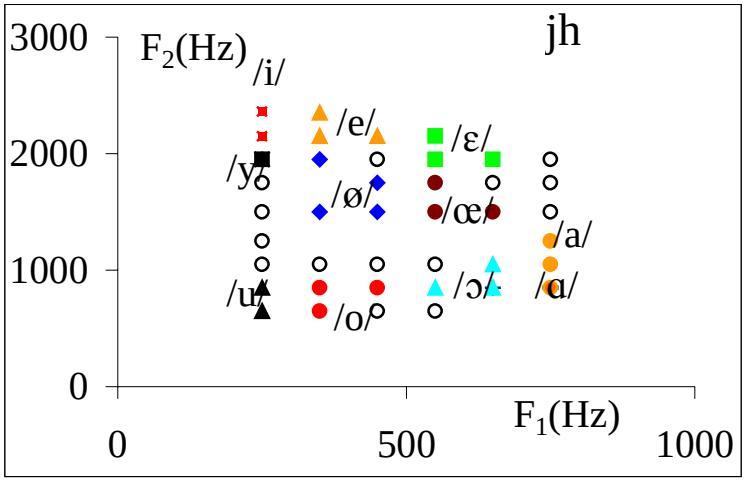


Figure 3. Représentation dans le plan F1/F2 des fréquences des formants des voyelles préférées des sujets jh et fp. Les deux triangles vocaliques sont plus ou moins de même dimension.

Les triangles vocaliques obtenus par les deux sujets en perception sont beaucoup plus proches que ceux obtenus en production.

2. L'expérience de Hombert et Puech (1984) (voir aussi Fujisaki (1969)) sur des sujets Swahili a été reproduite avec des sujets français. 53 sons différents pavant l'espace vocalique du plan F1/F2 ont été présentés et on a demandé d'identifier ces sons parmi les voyelles comme dans 'lit', 'les', 'l'air', 'la', 'las', 'l'or', 'l'eau', 'loup', 'lu','le', 'l'heure'. On constate figure 4 (comme pour la figure 1) que les plages d'identification des voyelles dans le plan F1-F2 varient avec les sujets (Hombert and Carré, 1999). Cette différence est probablement due à l'origine géographique à l'intérieur du domaine français (et par conséquent,

à « l'histoire linguistique » des sujets). La figure suivante montre ces plages pour deux individus.



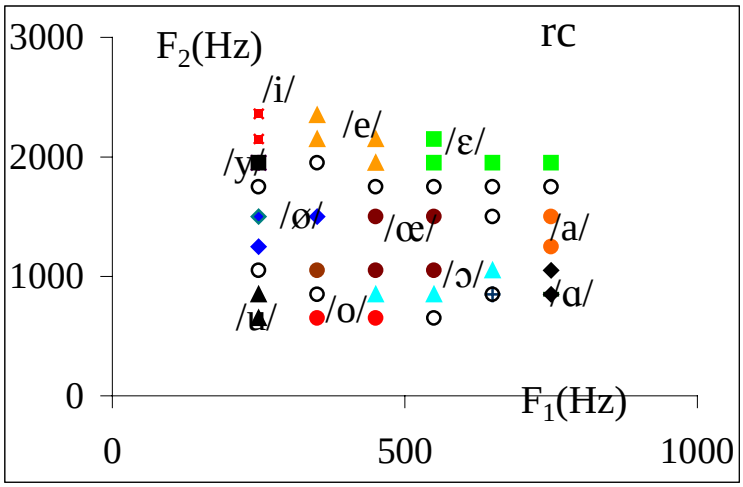
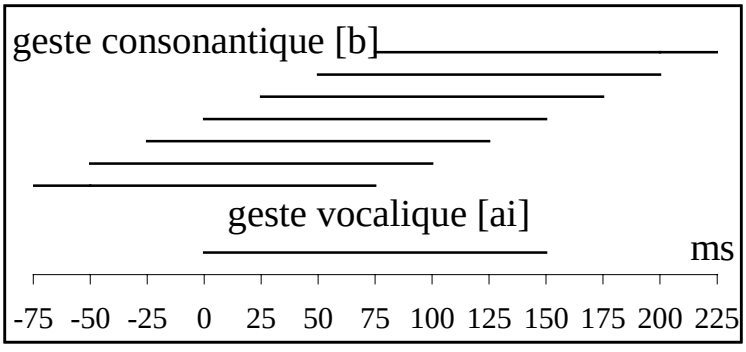


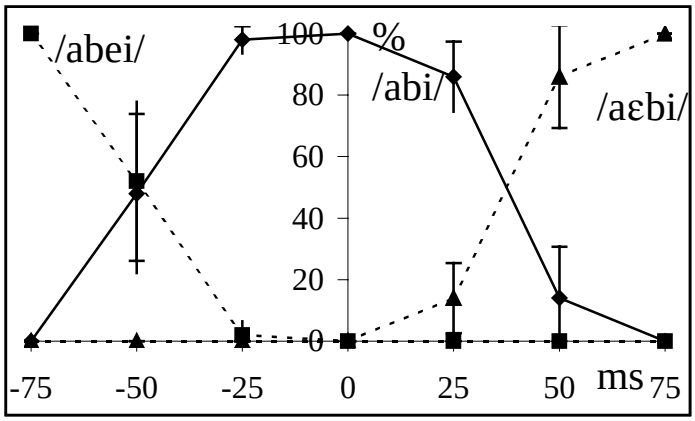
Figure 4. Plages d'identification des voyelles françaises dans le plan F1/F2 pour 2 sujets français pour 53 sons pavant le triangle vocalique. Chacune des voyelle est représentée par un symbole. Le signe 'o' indique que les sujets n'ont pas pu identifier le son présenté. On note une grande variabilité entre les deux sujets en ce qui concerne les frontières entre voyelles.

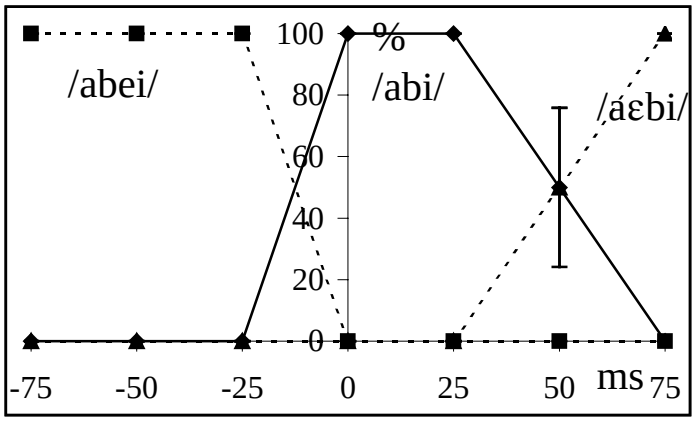
3. Dans l'expérience suivante, on a voulu tester des aspects dynamiques liés à la coarticulation dans le contexte d'ensembles VCV. L'approche théorique que nous avons retenue pour décrire ces ensembles est celle de la phonologie articulatoire qui décrit la parole en termes dynamiques par des 'gestes' se déroulant dans le temps (Browman and Goldstein, 1989). Par ailleurs, comme

conséquence de notre approche en termes de gestes donc en termes dynamiques qui implique un traitement dans le durée, nous adoptons comme unité de traitement en production ainsi qu'en perception la syllabe résultat de la co-production de plusieurs gestes (Kozhevnikov and Chistovich, 1965). En conséquence, des ensembles de type VCV seront traités comme une superposition d'un geste consonantique sur un geste vocalique. Les travaux de Browman et Goldstein s'appuient sur des données concernant les caractéristiques des gestes de parole. Dans la présente étude, nous nous appuierons sur nos propres travaux qui ont permis, à partir de la théorie acoustique et de la théorie de la communication, de déduire automatiquement des gestes de déformation de la fonction d'aire du conduit vocal (Mrayati, et al., 1988). Cette approche nous a permis de retrouver automatiquement, en particulier, les principaux lieux d'articulation généralement observés en production de voyelles et de consonnes (Carré and Mody, 1997). C'est avec de tels gestes que nous avons étudié l'effet perceptif du déphasage du geste vocalique par rapport au geste consonantique dans le cas de la production [abi] (Carré, 1999). La figure suivante (figure 5) montre : a) les différents déphasages (avance ou retard dans le temps) du début du geste consonantique correspondant à la occlusive [b] avec le début du geste vocalique produisant la transition vocalique [ai] ; b) le résultat des tests de perception pour deux sujets et pour ces différents déphasages. Le percept /abi/ est obtenu pour des variations importantes de déphasages d'environ 75 ms. En dehors de ces plages on perçoit /aɛbi/ ou /abei/ ce qui permet de faire apparaître de nouveaux sons.



a)





b)

Figure 5. a) Déphasage du geste consonantique par rapport au geste de constriction (entre –75 et 75 ms) pour la production de [abi] ; b) Résultats du test de perception correspondant pour 2 sujets français : /abi/ est perçu sur une plage de déphasage entre 50 et 75 ms et dépend du sujet. En dehors de cette plage, les sujets perçoivent /aɛbi/ et /abei/.

On a pu montrer par ailleurs qu'un déphasage spécifique de notre expérience peut être une caractéristique fixée d'un locuteur (Chennoukh, et al., 1997) et que notre modélisation explique les paramètres de l'équation du locus (Sussman, et al., 1998), ie, que l'équation du locus peut être obtenue au moyen du modèle et qu'elle varie avec le déphasage du geste consonantique par rapport au geste vocalique..

Ces premières expériences montrent donc, d'une part, que les stratégies de production et de perception des voyelles dépendent des sujets et que, d'autre part, un même locuteur a sa propre stratégie de co-production de deux (ou trois) gestes,

que de simples manipulations de durée ou de déphasage de gestes permettent de faire apparaître de nouveaux sons et que les plages de perception varient notablement d'un auditeur à un autre.

Discussions

Examinons maintenant dans quelle mesure la variation phonétique peut jouer un rôle dans l'évolution des langue. On peut en effet penser que, par analogie avec les processus biologiques, les évolutions linguistiques résultent d'une sélection opérée sur un ensemble de variantes. Cette hypothèse n'est évidemment pas nouvelle (voir par exemple Lindblom et al (1995)) mais nous souhaitons ici l'évaluer dans le cadre des stratégies individuelles proposées ci-dessus. Depuis les années 70, Ohala (voir par exemple Ohala (1981), Hombert et al (1979)) a proposé un modèle de changement phonético-phonologique s'appuyant à la fois sur l'effet des contraintes articulatoires et des contraintes perceptives. Il a montré comment les perturbations non intentionnelles du signal acoustique résultant des contraintes mécaniques du système de production de parole (en particulier, des effets de coarticulation) pouvaient aboutir à un décodage erroné ('mis-perception') de la part de l'allocutaire ('hearer'). Il faut souligner, d'une part, que dans ce modèle le décalage entre le message intentionnel du locuteur et le message réellement perçu par l'allocutaire n'est pas volontaire et ne code aucune information linguistiquement significative. D'autre part, le locuteur et l'allocutaire sont ici pris comme des individus 'moyens', ie, n'ayant aucune caractéristique linguistiquement distinctive d'une communauté donnée. En 1990, Lindblom (1990) propose le modèle hyper et hypo-speech (H&H) dans lequel la production du signal de parole prend en compte la situation interactionnelle entre locuteur et allocutaire : en s'appuyant sur les connaissances supposées de l'allocutaire, le locuteur ne produit que ce qui est nécessaire à la compréhension d'un nouveau message. Ainsi un message contenant des nouvelles informations sera hyper-articulé alors qu'un message se référant à de l'information partagée sera hypo-articulé.

Nous pensons, comme Lindblom, que le signal acoustique transmis s'adapte au contexte pragmatique de l'échange socio-langagier. Mais nous pensons aussi qu'au delà du système adaptatif proposé dans le cadre H&H, des informations sur le système linguistique d'un locuteur donné sont volontairement codées, transmises et évaluées par l'allocutaire. Les paramètres utilisés pour le codage (par le locuteur) et

pour le décodage (par l'allocutaire) sont évidemment fonction de leurs systèmes linguistiques respectifs. Ces systèmes résultent de l'histoire linguistique spécifique à chaque individu et sont implémentés par des stratégies individuelles en production et en perception (Hombert, 1984). Dans cette approche, le point de départ d'un changement linguistique dans une communauté se produit entre deux individus dont le système de production de l'un et le système de perception de l'autre seront maximalement distincts et non pas entre deux individus de caractéristiques moyennes comme cela a été proposé dans le modèle de Ohala. Les premiers résultats présentés ci-dessus sont compatibles avec nos hypothèses.

Remerciements

Ce travail est effectué avec l'appui du Programme Cognitique (Projet Langage et Cognition n° 41) du Ministère de la Recherche.

Bibliographie

Browman, C. P. and Goldstein, L. (**1989**). "Articulatory gestures as phonological units," Phonology **6**, 201-252.

Carré, R. (**1999**). "Perception of coproduced speech gestures," *Proc. of the 14th Int. Cong. of Phonetic Sciences*, (San Fransisco), pp. 643-646.

Carré, R. and Mody, M. (**1997**). "Prediction of Vowel and Consonant Place of Articulation," *Proceeding of the Third Meeting of the ACL Special Interest Group in Computational Phonology, SIGPHON 97*, (Madrid), pp. 26-32.

Chennoukh, S., Carré, R. and Lindblom, B. (**1997**). "Locus equations in the light of articulatory modeling," J. Acoust. Soc. Am. **102**, 2380-2389.

Delattre, P. C., Liberman, A. M. and Cooper, F. S. (**1955**). "Acoustic loci and transitional cues for consonants," J. Acoust. Soc. Am. **27**, 769-773.

Dorman, M. F., Studdert-Kennedy, M. and Raphael, L. J. (**1977**). "Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues," Perception and Psychophysics **22**, 109-122.

Fujisaki, H., Tomisawa, M. and Sato, T. (**1969**). "Speech audiometry by synthetic Japanese vowels," Sogoshikenjo-Nenpo **28**, 74-79.

Hombert, J. M. (**1984**). "Réflexion sur le mécanisme des changements phonétiques," Pholia **1**, 87-112.

Hombert, J. M. and Carré, R. (**1999**). "Correlation between vowel production and perception for a given speaker," J. Acoust. Soc. Am. **106**, S2152.

Hombert, J. M., Ohala, J. J. and Ewan, W. G. (**1979**). "Phonetic explanations for the development of tone," Language **55**, 37-58.

Hombert, J. M. and Puech, G. (**1984**). "Espace vocalique et structuration perceptuelle : Application au Swahili," Pholia **1**, 199-208.

Johnson, K., Flemming, E. and Wright, R. (**1993**). "The hyperspace effect: Phonetic targets are hyperarticulated," Language **69**, 505-528.

Kewley-Port, D., Pisoni, D. B. and Studder-Kennedy, M. (**1983**). "Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants," J. Acoust. Soc. Am. **73**, 1779-1793.

Kozhevnikov, V. A. and Chistovich, L. A. (**1965**) "Speech, articulation, and perception," JPRS-30543. NTIS, US Dept. of Commerce.

Liénard, J. S. (**1999**). "Variability, ambiguity and attention: a perception model based on analog induction," in *Human and Machine Perception; Emergence, Attention and Creativity*, edited by V. Cantoni (Kluwer Academic, New York) pp. 87-98.

Liénard, J. S. and Di Benedetto, M. G. (**1999**). "Effect of vocal effort on spectral properties of vowels," J. Acoust. Soc. Am. **106**, 411-422.

Lindblom, B. (**1990**). "Explaining phonetic variation: a sketch of the H and H theory," in *Speech Production and Speech Modelling, NATO ASI Series*, edited by A. Marchal and W. J. Hardcastle (Kluwer Academic Publishers, Dordrecht) pp. 403-439.

Lindblom, B., Guion, S., Hura, S., Moon, S.-J. and Willerman, R. (**1995**). "Is sound change adaptive?," Revista di Linguistica **7**, 5-37.

Lindblom, B. and Studdert-Kennedy, M. (**1967**). "On the role of formant transitions in vowel perception," J. Acoust. Soc. Am. **42**, 830-843.

Mrayati, M., Carré, R. and Guérin, B. (**1988**). "Distinctive region and modes: A new theory of speech production," Speech Communication **7**, 257-286.

Ohala, J. J. (**1981**). "The listener as a source of sound change," in *Papers from the Parasession on Language and Behavior*, edited by R. A. Masek, R. A. Hendrick and M. F. Miller (Chicago Ling. Soc., Chicago) pp. 178-203.

Peterson, G. E. and Barney, H. L. (**1952**). "Control methods used in the study of the vowels," J. Acoust. Soc. Am. **24**, 175-184.

Strange, W., Verbrugge, R. R., Shankweiller, D. P. and Edman, T. R. (**1976**). "Consonant environment specifies vowel identity," J. Acoust. Soc. Am. **60**, 213-224.

Sussman, H. M., Fruchter, D., Hilbert, J. and Sirosh, J. (**1998**). "Linear correlates in the speech signal: The orderly output constraint," Behavioral and Brain Sciences **21**, 241-299.

Sussman, H. M., McCaffrey, H. A. and Matthews, S. A. (**1991**). "An investigation of locus equations as a source of relational invariance for stop place categorization," J. Acoust. Soc. Am. **90**, 1309-1325.