Les trajectoires formantiques respectant les lois de la physique contribuent-elles à une meilleure perception de la parole ?

Daniel Pape[1,3], Pascal Perrier[1], Susanne Fuchs[2], Sonia Kandel[4&5]

[1]ICP/GIPSA-lab, CNRS, Grenoble INP, Grenoble, France
[2]Phonetik/ZAS, Berlin, Allemagne
[3] IEETA, Université d'Aveiro, Portugal
[4]LPNC, CNRS, Université Pierre-Mendès France, Grenoble, France
[5]Institut Universitaire de France
Pascal.Perrier@gipsa-lab.grenoble-inp.fr
http://www.gipsa-lab.inpg.fr/page_pro.php?vid=168

ABSTRACT

Physical properties of speech articulators contribute to shape articulatory and formant trajectories. This study aims at evaluating the role of this shaping in speech perception. We conducted perception tests of synthetic stimuli generated with speech production models accounting for different degrees of physical complexity. Our results do not provide any support to the hypothesis that the degree of physical realism in the models influences the perception of naturalness. However for degraded speech (silent center speech), significant differences are observed.

Keywords: Perception-Action Interaction; Perception and Physics, Speech Goals

1. INTRODUCTION

La question du rôle des trajectoires formantiques dans la perception de la parole est ancienne et elle a été au cœur de nombreux débats. Deux théories s'opposent. Pour la première, ces trajectoires ne contiennent pas l'information pertinente phonétiquement. Cette théorie est en particulier défendue par ceux qui prônent l'existence d'une cible perceptive associée à chaque son élémentaire et le caractérisant ([6, 11] pour les voyelles, et [14] pour les consonnes). Dans ce cadre théorique, la trajectoire serait cependant exploitée perceptivement dans les cas où la cible perceptive n'est pas atteinte. Elle serait donc, non pas l'objet de la perception, mais le vecteur d'une information à partir de laquelle la cible perceptive pourrait être retrouvée [3, 4]. Pour la seconde théorie, les caractéristiques temporelles intrinsèques de la transition porteraient en elles-mêmes, et indépendamment de toute cible, l'information pertinente, celle du geste nécessaire à la production du son élémentaire. Ainsi, c'est la trajectoire formantique dans tous ses détails qui serait l'objet de la perception ([12, 13])

Dans ce contexte, les travaux de Cai et al [1] sont extrêmement intéressants. Ces auteurs ont étudié les mécanismes d'adaptation de locuteurs chinois (Mandarin) produisant la triphtongue /iau/ quand leur feedback auditif était perturbé en temps réel. Cette perturbation modifie la trajectoire formantique, tout en préservant les patrons formantiques atteints pour chacune des voyelles. Statistiquement, les sujets ont eu tendance à réagir à la perturbation en modifiant leurs stratégies articulatoires de façon à corriger l'effet de la perturbation sur la trajectoire formantique perçue. Ceci laisse penser que la trajectoire formantique entre les cibles pourrait constituer l'objectif majeur de la tâche motrice. Et dans une perspective d'interaction locuteur-auditeur, on peut conclure que la trajectoire dans sa totalité serait porteuse de l'information que le locuteur veut transmettre à l'auditeur. Ces travaux peuvent donc être interprétés comme un soutien aux hypothèses défendues par Strange et collègues [12, 13].

Cependant une autre explication est envisageable. Elle est suggérée par les travaux de Viviani et collègues (par exemple [18]) sur la perception visuelle et l'identification des mouvements de la main. Ces auteurs ont d'une part observé que, dans ces mouvements, la vitesse varie comme la racine cubique de l'inverse de la courbure de la trajectoire. Ils ont d'autre part constaté expérimentalement que cette loi était exploitée par le système de perception visuelle pour catégoriser les trajectoires. Ainsi, par exemple, un point lumineux parcourant sur un écran un cercle à vitesse constante (conformément à la courbure) est perçu comme un mouvement circulaire, alors que si ce point décrit ce même cercle à une vitesse variable, rapide-lente-rapide-lente, c'est un mouvement elliptique qui est perçu. Viviani et al. en ont conclu que la perception visuelle du mouvement chez l'homme est influencée par les connaissances que l'homme a des propriétés intrinsèques de ses mouvements. Il est alors possible d'interpréter les résultats de Caï et al., non pas comme la preuve du caractère central de la trajectoire dans la perception de la parole, mais comme la conséquence d'une perturbation non écologique qui ne respecterait par les règles physiques et les règles de contrôle moteur régissant les mouvements de la parole. Les stratégies de compensation ne viseraient alors pas à reproduire une trajectoire pertinente phonétiquement, mais à donner au mouvement ses propriétés naturelles.

Cet article présente la première étape d'une démarche d'évaluation du rôle potentiel, dans la perception de la parole, de l'impact des propriétés physiques des articulateurs de la parole sur les trajectoires formantiques. Nous y décrivons et analysons des tests

perceptifs de stimuli synthétiques générés avec des modèles intégrant plus ou moins la complexité physique de ces articulateurs. Après la présentation de la méthodologie, nous exposerons les résultats et nous conclurons dans le contexte du débat sur le rôle des trajectoires dans la perception de la parole.

2. METHODE

2.1. Sujets

Vingt-trois sujets (16 hommes, 7 femmes, âgé(e)s entre 25 et 50 ans) ont participé à l'expérience. Aucun n'était au courant des méthodes de synthèse utilisées. Tous sont de langue maternelle française. Aucun d'entre eux n'a fait part de problème d'élocution ou auditif.

2.2. Matériel et procédure

Génération des stimuli synthétiques

Des stimuli Voyelle1-Voyelle2-Voyelle1 ($V_1V_2V_1$) et Voyelle1-/g/-Voyelle1 (V_1CV_1) ont été synthétisés. Les voyelles ont été choisies parmi /i/, /e/, /ɛ/, /a/, ou /ɔ/ Ces stimuli ont été obtenus par synthèse articulatoire, à partir de formes sagittales du conduit vocal. Dans tous les cas, la synthèse acoustique a impliqué la génération de la fonction d'aire à partir de la forme sagittale [8], puis celle du signal acoustique par un modèle de type Kelly-Lochbaum (développé par B. Story [15,16]) excité par un modèle de la source vocale [17]. Dans tous les cas, la fréquence fondamentale a été maintenue constante et égale 110Hz. Les modèles de synthèse diffèrent par la façon dont les formes sagittales sont générées dans le temps.

Pour une première classe de stimuli, que nous appellerons ***Mod1,*** les formes sagittales ont été obtenues avec un modèle biomécanique bidimensionnel du conduit vocal [9]. Le modèle de contrôle est de type *cible* : on spécifie les commandes motrices de chaque son élémentaire et les mouvements entre ces sons sont la conséquence d'une évolution temporelle linéaire des commandes entre leurs valeurs cibles [7], selon un timing bien défini (tenue des cibles : 150ms ; transition entre cibles : 120ms). Différentes évaluations de ce modèle ont attesté de sa capacité à rendre compte de manière réaliste des caractéristiques cinématiques importantes des mouvements de la parole, amplitude du mouvement, valeur du pic de vitesse, forme du profil de vitesse [7], formes des trajectoires dans le plan sagittal [9], et relation vitesse-courbure [10].

Pour la synthèse des autres classes de stimuli, les formes de la langue effectivement atteintes pour chaque son élémentaire dans les stimuli de la classe ***Mod1*** ont été extraites, et ont servi de formes cibles. Pour chacune d'elles les instants auxquels elles sont atteintes et les durées de leurs tenues ont été mesurés. De nombreuses données expérimentales ont montré que pour /g/ la langue se déplace vers l'avant pendant la tenue consonantique tout en restant en contact avec le palais. C'est pourquoi deux formes cibles ont été extraites pour ce son, l'une au début de la phase de contact avec le palais (g_f) et l'autre à l'instant du relâchement consonantique (g_o). Ces formes cibles sont définies par la position de 17 points dans le plan sagittal. Tous les stimuli atteignent et maintiennent les formes cibles selon le timing mesuré dans les stimuli de la classe ***Mod1***, et le passage d'une forme cible à une autre se fait par le déplacement des 17 points selon des trajectoires rectilignes. Les classes de stimuli se différencient par le décours temporel des déplacements de ces points. Dans la classe ***Mod2,*** le déplacement se fait à vitesse constante. Dans la classe ***Mod3***, la vitesse est un arc de sinusoïde conforme aux caractéristiques d'un système du second ordre. Différentes données expérimentales ont en effet montré que ce type de profil de vitesse était couramment observé dans les mouvements de la parole. Enfin pour V_1CV_1 une quatrième classe de stimuli a été générée, ***Mod4,*** dans laquelle seule la forme g_o a été retenue pour /g/. Cette forme est maintenue pendant toute la durée de la tenue consonantique et les transitions entre formes-cibles ont le schéma temporel de la classe ***Mod2.*** Les stimuli ont donc été générés à partir de 3 modèles intégrant différentes complexités dans la représentation physique des articulateurs : ***Mod1*** correspond à la description la plus réaliste, suivi par ***Mod3*** puis par ***Mod2*** (et ***Mod 4*** pour V_1CV_1).

Pour ce travail nous avions deux objectifs majeurs : évaluer dans quelle mesure le degré de réalisme des modèles influence la qualité perçue de la synthèse ; étudier si dans des conditions de parole dégradée la perception est influencée par le réalisme des modèles. La parole dégradée a été générée en transformant les stimuli $V_1V_2V_1$ selon le paradigme des *centres silencieux* [13]. Le signal de parole y est dégradé du fait du remplacement de la partie stable de V_2 par du silence. Cette transformation a été réalisée manuellement à l'aide du logiciel PRAAT, en remplaçant par des 0 la portion de signal déterminée par les passages par zéro situés 10ms avant et 10 ms après la zone de stabilité du formant F2.

Tests perceptifs

Tous les tests ont été réalisés en chambre sourde au GIPSA-lab, à l'aide du logiciel gratuit Alvin [2]. Les sujets ont d'abord passé le test sur les stimuli à centre silencieux, puis les tests d'évaluation de la qualité de la synthèse, d'abord pour $V_1V_2V_1$, puis pour V_1CV_1. De courts tests d'entrainement ont été effectués en début de session.

Lors des tests sur les stimuli à centre silencieux, la tâche de l'auditeur a consisté à identifier V_2. Toutes les combinaisons $V_1V_2V_1$, y compris celles du type $V_1V_1V_1$, ont été évaluées. Ainsi 2 répétitions de 75 stimuli (3 modèles x $5V_1$ x $5V_2$) ont été présentées aux sujets. L'instruction était d'« *identifier la voyelle2 manquante aussi vite et aussi précisément que*

possible ». Les sujets répondaient en appuyant sur l'un des 5 boutons affichés à l'écran. Sur chaque bouton était représenté le symbole phonétique de la voyelle V_2, associé à un mot monosyllabique dont la prononciation contient cette voyelle. Ces boutons étaient situés sur un cercle centré sur un 6ème bouton sur lequel il fallait appuyer pour continuer l'expérience. Ainsi, à chaque nouvelle écoute, la souris était positionnée à égale distance des 5 boutons réponses. Ceci a permis une mesure fiable des temps de latence, en évitant une variabilité liée à la celle de la distance à parcourir avec la souris. La réécoute des stimuli n'était pas possible.

Pour évaluer les liens entre le réalisme du modèle et la qualité des stimuli synthétiques, des paires de stimuli générés par deux modèles ont été présentés. Les sujets devaient effectuer une tâche de discrimination et choisir « *aussi rapidement et précisément que possible lequel de ces deux stimuli est le plus naturel* ». Pour cela, ils devaient appuyer avec l'index gauche sur la touche 1 du clavier (stimulus 1) ou avec l'index droit sur la touche 2 (stimulus 2) du clavier numérique, les mains restant immobiles. Les tests ont été élaborés selon la procédure à choix forcé 2I-2AFC, où les deux stimuli sont présentés séquentiellement, séparés par une pause. Chaque stimulus dure 650ms et la pause est de 500ms. Les différentes classes de stimuli ont été combinées au sein des paires de manière aléatoire pour chaque séquence (exemple « ***Mod1***-Pause-***Mod2*** », « ***Mod2***-Pause-***Mod3*** », ou « ***Mod1***-Pause-***Mod3*** »). Pour chaque combinaison de 2 modèles, la moitié des stimuli, toutes séquences confondues, a été présentée dans un ordre, et la moitié dans l'ordre inverse. Cette procédure a été choisie pour garantir au mieux un traitement équivalent de tous les stimuli et rendre ainsi pertinente la mesure du temps de latence. Certaines paires contenaient aussi des stimuli identiques afin de tester la fiabilité des sujets (Test aveugle). L'évaluation perceptive des stimuli $V_1V_2V_1$ et V_1CV_1 a été faite en deux tests séparés. Pour $V_1V_2V_1$, 2 répétitions de 80 paires de stimuli [(3 couples de modèles + test aveugle) x 5V1 x 4V_2] ont été présentées. Pour V_1CV_1, /i/ n'a pas été prise en compte et 4 répétitions de 28 paires de stimuli [(6 couples de modèles + test aveugle) x 4V1], ont été présentées. La réécoute n'était pas possible.

2.3. Analyse statistique

L'objectif de l'analyse statistique des données est de voir si, tous sujets et tous stimuli confondus, il existe des différences entre les classes de stimuli. Nous avons utilisé pour cela le modèle linéaire généralisé avec effets mixtes disponible dans le logiciel R (2008). Nous avons choisi cette analyse plutôt qu'une ANOVA classique, car elle permet de traiter des classes qui n'ont pas le même nombre de données. Elle offre aussi la possibilité d'éliminer dans le traitement statistique la contribution des stimuli ($V_1V_2V_1$ et V_1CV_1) et des sujets dans la variance globale, en les considérant comme des facteurs aléatoires de variabilité. La classe des stimuli (***Mod1, Mod2, Mod3, Mod4)*** a été choisie comme facteur fixe.

3. RÉSULTATS

Réalisme du modèle et qualité de la synthèse

Les réponses aux paires de stimuli identiques ont été utilisées pour vérifier les performances de nos sujets. Théoriquement, pour de tels stimuli, les choix devraient se répartir équitablement entre les deux réponses. C'est pourquoi nous avons éliminé de l'analyse, pour l'ensemble des tests, quatre sujets qui présentaient une répartition des réponses plus déséquilibrée que 70%-30%. Pour les 19 sujets restants, que ce soit pour les séquences $V_1V_2V_1$ ou les séquences V_1CV_1, les réponses aux tests de qualité de la synthèse n'ont pas permis de mettre en évidence une différence significative entre les classes de stimuli. Ces tests ne permettent donc pas de démontrer qu'il existe une influence du réalisme de la modélisation physique sur le caractère plus ou moins naturel de la synthèse.

Perception des stimuli à centres silencieux.

Si on considère toutes les réponses correctes données pour les stimuli à centre silencieux par les 19 sujets sélectionnés, aucune différence entre les modèles ne peut être montrée. Cependant diverses études [5] ont montré que les temps de latence (TL) pouvaient être très informatifs des mécanismes perceptifs impliqués, notamment les TL courts. En effet, de TL longs peuvent être liés à l'implication de traitements cognitifs de haut niveau, dépassant le stade de la perception auditive proprement dite.

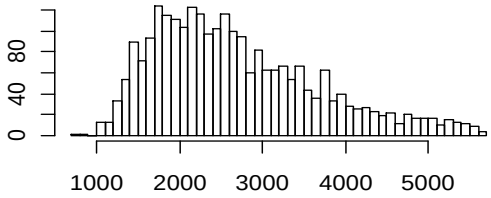

Figure 1 : Distribution des temps de latence (ms)

Pour notre test, la variabilité est grande, avec une moyenne, tous sujets et tous stimuli confondus, de 2885ms et un écart-type de 1375ms. La figure 1 donne la distribution des TL dans l'intervalle [-2σ +2σ].

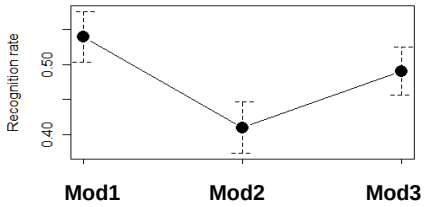

Figure 2 : Moyennes et erreurs-types (Int. Conf. :0.95) (nombre sujets : 19 ; nombre données =592)

Sur cette base, seules les réponses avec des TL dans l'intervalle [1000ms 2000ms] ont été prises en compte. Les moyennes et erreurs-types du pourcentage de bonnes réponses calculées par classe de stimuli sont

présentées fig. 2. Ce pourcentage est significativement meilleur (pMCMC=0.016, t=-2.45) pour les stimuli ***Mod1*** que ***Mod2***. Une tendance suggère un meilleur taux d'identification des stimuli ***Mod3*** que ***Mod2***, mais elle n'est pas significative (pMCMC=0.0874, t=1.678). Les différences entre les stimuli ***Mod1*** et ***Mod3*** ne sont pas significatives (pMCMC=0.4098, t=-0839).

4. CONCLUSIONS

Les tests évaluant le caractère naturel de la synthèse n'ont pas permis de faire apparaître un rôle quelconque du réalisme physique de la modélisation. Ce résultat ne va pas dans le sens de notre explication aux résultats expérimentaux de Cai et al. [1]. Cependant, il est possible que leurs stimuli perturbés soit encore moins écologiques que ceux qui ont été générés par le moins physique de nos modèles (***Mod2***). Nos résultats ne permettent pas non plus de dire que les auditeurs exploitent une connaissance du comportement physique des articulateurs pour évaluer le caractère naturel des gestes, contrairement à ce qu'ont montré Viviani et collègues pour la perception visuelle.

Les tests sur les stimuli à centre silencieux donnent des résultats significativement moins bons si les stimuli ont été générés par le modèle purement cinématique (***Mod2***) que s'ils l'ont été avec le modèle le plus réaliste (***Mod1***). Ainsi le réalisme physique semble aider à retrouver dans les transitions l'information phonétique manquante. Une information purement directionnelle (***Mod2***) sur la variation formantique est moins efficace qu'une description dynamique plus complexe. Ces résultats confirment le rôle des trajectoires dans la perception de la parole dégradée. Cependant pour les stimuli de la classe ***Mod1***, puisque le modèle de contrôle est de type « cible », les trajectoires sont intrinsèquement liées à la cible du mouvement. Nos résultats suggèrent donc que les auditeurs pourraient extraire des trajectoires l'information sur la physique des articulateurs afin de retrouver la cible manquante. Cela va dans le sens des hypothèses formulées par Loevenbruck & Perrier [3].

REMERCIEMENTS

A Brad Story (Univ. of Arizona) pour son synthétiseur acoustique. Ce travail est soutenu par l'Université Franco-Allemande (Sarrebruck) (Projet PILIOS) et le financement SFRH/BPD/48002/2008 du FCT Portugal.

BIBLIOGRAPHIE

[1] Cai, S., Boucek, M., Ghosh, S.S., Guenther, F. H. & Perkell, J.S. (2008). A system for online dynamic perturbation of formant trajectories and results from perturbations of the Mandarin triphthong /iau/. *Proc. of ISSP-2008.*, (pp. 65-68), Strasbourg, France.

[2] Hillenbrand J. & Gayvert R. (2005) Open Source Software for Experiment Design and Control, *J.S.L.H.R., 48,* 45-60.

[3] Lœvenbruck H. & Perrier P. (1996). How could undershot vowel targets be recovered? A dynamical approach based on the Equilibrium Point Hypothesis for the control of speech movements. *Proc. of ISSP-1996* (pp. 117-120), Autrans, France.

[4] Lindblom, B. & Studdert-Kennedy M. (1967) On the role of formant transitions in vowel recognition. *J Acoust Soc Am, 42*, 830-843.

[5] Miller, J.L. & Dexter, E.R. (1988). Effects of speaking rate and lexical status on phonetic perception. *J Exp Psychol Hum Percept Perform, 14*, 369–378.

[6] Nearey, T. (1977). *Phonetic feature systems for vowels*,Doctoral dissertation, University of Connecticut, Storrs, CT.

[7] Payan, Y. & Perrier, P. (1997). Synthesis of V-V sequences with a 2D biomechanical tongue model controlled by the equilibrium point hypothesis. *Speech Comm., 22 (2/3)*, 185-205.

[8] Perrier P., Boë L.J. & Sock R. (1992). Vocal Tract Area Function Estimation From Midsagittal Dimensions With CT Scans and a Vocal Tract Cast: Modeling the Transition With Two Sets of Coefficients. *J.S.H.R.,35*, 53-67

[9] Perrier, P., Payan, Y., Zandipour, M. & Perkell, J. (2003). Influences of tongue biomechanics on speech movements during the production of velar stop consonants: a modeling study. *J Acoust Soc Am, 114 (3)*, 1582-1599.

[10] Perrier, P. & Fuchs, S. (2008). Speed-curvature relations in speech production challenge the 1/3 power law. *J Neurophysiol, 100 (3), 1171-1183.*

[11] Stevens, K. N. (1972) The quantal nature of speech: Evidence from articulatory-acoustic data, in E. E. David, Jr. & P. B. Denes, (Eds.), *Human Communication: A Unified View* (pp. 51–66).

[12] Strange, W., Edman, T.R., & Jenkins, J.J. (1979). Acoustic and phonological factors in vowel identification. *J Exp Psychol Hum Percept Perform, 5(4),* 643-656

[13] Strange, W., Jenkins, J.J. & Johnson, T.L. (1983). Dynamic specification of coarticulated wovels, *J Acoust Soc Am, 74(3)*, 695-705.

[14] Stevens, K. N. & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *J Acoust Soc Am, 64 (5)*, 1358-1368

[15] Story, B. H. (2004). Vowel acoustics for speaking and singing, *Acta Acustica 90(4)*, 629--640.

[16] Story, B.H. (2005). A parametric model of the vocal tract area function for vowel and consonant simulation, *J. Acoust. Soc. Am., 117(5)*, 3231-3254.

[17] Titze, I.R. (1984). Parameterization of the glottal area, glottal flow, and vocal fold contact area, *J. Acoust. Soc. Am., 75*, 570-580.

[18] Viviani, P., & Stucchi, N. (1992). Biological movements look uniform: evidence of motor-perceptual interactions *J Exp Psychol Hum Percept Perform*, 18 (3), 603-623..