Apprentissage de contrastes non-natifs : Limites des entraînements statistiques

Gregory Collet[1,2,3], *Jacqueline Leybaert*[3], *Willy Serniclaes*[2,4] *et Cécile Colin*[2]
(1) FRS-FNRS, 5, rue d'Egmont, B-1000 Bruxelles, Belgique
(2) ULB, UNESCOG, CP191 Bruxelles, Belgique
(3) ULB, LCLD, CP191 Bruxelles, Belgique
(4) CNRS, UMR 8158, Paris, France
gcollet@ulb.ac.be, leybaert@ulb.ac.be, wsernic@ulb.ac.be, ccolin@ulb.ac.be

RESUME

Récemment, des études ont montré que l'information statistique contenue dans le signal de parole permettait l'acquisition des catégories phonologiques. Ainsi, l'exposition à des distributions bimodales engendrait une augmentation de la discrimination des phonèmes alors que l'exposition à des distributions unimodales ne modifiait pas les capacités perceptives. Le but de cette étude était de déterminer dans quelle mesure différentes distributions (bimodale, unimodale, uniforme) pouvaient avoir un impact sur l'augmentation des capacités de discrimination de stimuli linguistiques allophoniques séparés par de fines différences acoustiques (i.e. 20 ou 30 ms de Délai d'Établissement du Voisement – DEV). Les résultats indiquent que malgré l'augmentation des performances de discrimination pour certains contrastes, l'extraction de régularités statistiques reste difficile pour de fines différences acoustiques.

ABSTRACT

Learning non-native contrasts: Limits of the statistical training

Recently, studies showed that distributional information contained in the speech signal can contribute to the acquisition of phoneme categories. Indeed, exposition to bimodal distributions leads to an improvement of phoneme discrimination compared to unimodal distributions which did not improve speech perception. The aim of the present study was to determine to what extent distributional information (bimodal, unimodal, uniform) can induce change in allophonic fine-grained speech perception (i.e. 20 or 30 ms of Voice Onset Time – VOT). Results showed that despite discrimination improvement across some voicing contrasts, extraction of statistical regularities remains difficult for small acoustic differences.

MOTS-CLES : Entraînements statistiques, Perception de la parole, Voisement.
KEYWORDS : Statistical training, Fine-grained speech perception, Voicing.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 489–496,
Grenoble, 4 au 8 juin 2012. ©2012 ATALA & AFCP

1 Introduction

De nombreuses études ont montré que l'extraction de l'information statistique contribue à la construction des capacités de perception de la parole dès le plus jeune âge (Guenther & Gjaja, 1996 ; Jusczyk, Luce, & Charles-Luce, 1994 ; Peperkamp, Pettinato, & Dupoux, 2003 ; Saffran, Aslin, & Newport, 1996).

Parmi ces travaux, Maye et Gerken (2000 ; 2001) se sont particulièrement intéressés à la capacité des adultes à extraire de l'information statistique dans leur environnement afin de former de nouvelles catégories phonologiques. Des adultes anglophones étaient exposés pendant quelques minutes à des syllabes issues d'un continuum de Délai d'Etablissement du voisement (DEV : délai entre la fin de l'occlusion de la consonne et le début de la vibration des cordes vocales ; ou VOT, Voice Onset Time). Les huit stimuli utilisés (dont les valeurs extrêmes étaient séparées par 120 ms de DEV) étaient caractérisés par différentes valeurs de DEV dont la fréquence d'occurrence variait. Deux groupes de participants étaient constitués sur la base des différents types de distributions proposées : 1) distribution bimodale : deux stimuli situés aux extrémités du continuum présentaient une plus haute fréquence d'occurrence que les autres (formation de deux groupes de stimuli de part et d'autre du continuum) ; et 2) distribution unimodale : deux stimuli centrés au milieu du continuum présentaient une plus haute fréquence d'occurrence (formation d'un groupe de stimuli au milieu du continuum). Les résultats ont montré que les participants parvenaient à créer de nouvelles catégories en se basant sur l'information contenue dans les blocs de stimulation, montrant ainsi une amélioration de la discrimination après exposition à des stimuli formant deux groupes (stimulation bimodale).

Le but de notre recherche était de tester les limites de l'extraction de l'information statistique, chez des adultes francophones, en utilisant des stimuli allophoniques séparés par de beaucoup plus fines différences acoustiques que chez Maye et Gerken (2000 ; 2001).

Pour ce faire, nous avons décidé de comparer des contrastes séparés par 20 et 30 ms de DEV. De plus, étant donné que la frontière perceptive de DEV en Français se situe aux alentours de 0 ms de DEV (Serniclaes, 2011), opposant la perception du DEV positif et négatif, nous avons décidé d'effectuer nos recherches sur les deux côtés de ce continuum de voisement. Un total de quatre contrastes a donc été utilisé au travers de cette étude (i.e. +15/+45, +20/+40 ; -15/-45 et -20/-40). Sur la base de la perception de locuteurs francophones, les deux contrastes provenant du côté positif (i.e. +15/+45, +20/+40) sont constitués d'allophones de la catégorie /te/ et les deux contrastes du côté négatif (-15/-45 et -20/-40) sont constitués d'allophones de la catégorie /de/.

Enfin, en plus des deux conditions d'exposition proposées par Maye et Gerken (2000 ; 2001), nous avons décidé d'en ajouter une troisième pour chacun de nos quatre contrastes. Il s'agit de la condition Uniforme dans laquelle la même fréquence d'occurrence était appliquée pour l'ensemble des stimuli. Cette condition permettait de contrôler que l'augmentation de la discrimination en post-test était bien due à l'extraction de régularités statistiques et non pas à la simple exposition aux stimuli. De plus, lors d'expériences pilotes, nous nous sommes assurés que le test-retest n'engendrait pas d'augmentation des performances en discrimination chez les participants.

2 Méthode

2.1 Participants

Pour chacun des quatre contrastes étudiés, trois groupes de 10 participants francophones ont été constitués. La moyenne d'âge générale de ces 12 groupes était de 20.3 ans (ET = 2.1) et ce facteur ne constituait pas une source de différence entre les groupes ($F(11,119) = 1.6$; $p = .10$).

Pour chaque contraste, chacun des groupes a été soumis, lors de la phase de stimulation, à un des trois types de distribution des stimuli présentés à la Figure 1. Pour chaque distribution, la fréquence d'occurrence des stimuli présentés était différente : 1) distribution bimodale : deux stimuli situés aux extrémités du continuum présentaient une plus haute fréquence d'occurrence que les autres (formation de deux groupes) ; 2) distribution unimodale : deux stimuli centrés au milieu du continuum présentaient une plus haute fréquence d'occurrence (formation d'un groupe) ; et 3) distribution uniforme : la fréquence d'occurrence était identique pour tous les stimuli du continuum (pas de formation de groupe).

2.2 Stimuli

Les stimuli utilisés provenaient d'un continuum de voisement généré par un synthétiseur de parole (Carré, 2004). Les fréquences des transitions initiales de F1, F2 et F3 étaient respectivement de 200, 2200 et 3100 Hz et les parties stables des trois formants étaient respectivement de 500, 1500 et 2500 Hz. La valeur de F0 était quant à elle de 120 Hz. La durée des transitions était de 24 ms et la durée totale des stimuli était de 200 ms.

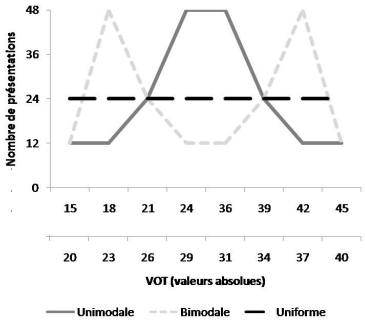

FIGURE 1 – Nombre de présentations pour les différents stimuli lors de la phase de stimulation. L'axe des abscisses présente les huit valeurs de VOT (en valeurs absolues) impliquées lors de la phase de stimulation pour les contrastes +15/+45 et -15/-45 (ligne supérieure) et pour les contrastes +20/+40 et -20/-40 (ligne inférieure). Les courbes représentent la stimulation Unimodale (gris foncé), la stimulation Bimodale (gris clair pointillé) et la stimulation Uniforme (noir pointillé).

Pour chacun des quatre contrastes, un ensemble spécifique de huit stimuli différents était utilisé (Figure 1). Dans le cas des deux contrastes séparés par 30 ms de DEV (i.e.

+15/+45 et -15/-45), les mêmes valeurs absolues de DEV positif ou négatif étaient utilisées (i.e. 15, 18, 21, 24, 36, 39, 42 et 45). De même, pour les deux contrastes séparés par 20 ms de DEV (i.e. +20/+40 et -20/-40), les mêmes valeurs absolues de DEV positif ou négatif étaient utilisées (i.e. 20, 23, 26, 29, 31, 34, 37 et 40).

2.3 Procédure

L'expérience était constituée d'un pré-test et d'un post-test séparés par deux blocs de stimulation.

Lors des phases de pré- et post-test, les participants réalisaient une tâche de discrimination. Un bloc de 20 paires de stimuli était présenté. Chaque bloc était constitué des cinq répétitions de chaque combinaison possible entre les deux stimuli du contraste. Par exemple, pour le contraste -15/-45 : les combinaisons étaient -15/-15, -45/-45, -45/-15 et -15/-45. Pour les paires -15/-15 et -45/-45, la réponse « *même* » était attendue alors que pour les paires -45/-15 et -15/-45, la réponse « *différente* » était attendue. L'intervalle inter-stimuli (IIS) à l'intérieur de la paire était de 500 ms et les participants disposaient de 2500 ms entre chaque paire afin de donner leur réponse. Les paires de stimuli étaient présentées dans un ordre aléatoire différent pour chaque phase et pour chaque sujet.

Lors des phases de stimulation, les participants réalisaient une tâche de détection de stimuli. Il leur était demandé d'appuyer le plus rapidement possible sur une touche cible du clavier de l'ordinateur dès qu'ils entendaient un stimulus. Cette tâche requérant une attention importante, une pause d'une minute était proposée au milieu de l'épreuve. De plus, afin d'éviter la mise en place de potentiels mécanismes d'anticipation, l'intervalle après chaque réponse du participant variait entre 1000 et 2500 ms. L'ordre de présentation des stimuli était aléatoire et différent pour chaque phase de stimulation et pour chaque participant.

Quel que soit le contraste ou le type de stimulation, l'expérience comportait un total de 192 stimuli réparti sur deux blocs (se référer à la Figure 1 pour le nombre de présentations associées à chaque stimulus lors de la stimulation).

2.4 Traitement des données

Chaque contraste a été analysé séparément au moyen d'une ANOVA à mesures répétées avec le facteur Session comme facteur intra-sujet (deux niveaux : pré- et post-test), le facteur Groupe comme facteur inter-sujet (trois niveaux : bimodale, unimodale et uniforme) et le score de discrimination comme variable dépendante.

3 Résultats

3.1 Contraste +15/+45 ms DEV

Le facteur Session était significatif ($F(1,27) = 25.4$; $p < .0005$; $\eta^2 = .48$), les participants obtenant en moyenne 13% de discriminations correctes de plus lors du post-test (voir Figure 2). Le facteur groupe n'était pas significatif ($F < 1$) et n'interagissait pas avec le facteur Session ($F < 1$).

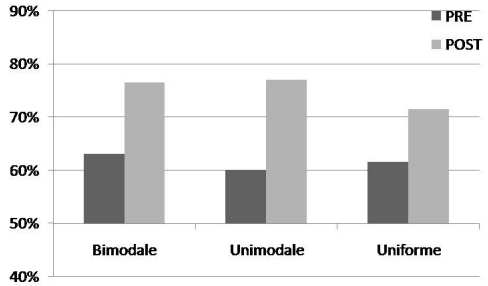

FIGURE 2 – Pourcentages de discriminations correctes pour le contraste +15/+45 lors du pré-test (gris foncé) et lors du post-test (gris clair) pour les trois groupes entraînés.

3.2 Contraste -15/-45 ms DEV

A nouveau, le facteur Session était significatif ($F(1,27) = 7.5$; $p < .05$; $\eta^2 = .22$), les participants obtenant en moyenne 9% de discriminations correctes de plus lors du post-test (voir Figure 3). Le facteur groupe n'était pas significatif ($F < 1$) et n'interagissait pas avec le facteur Session ($F < 1$).

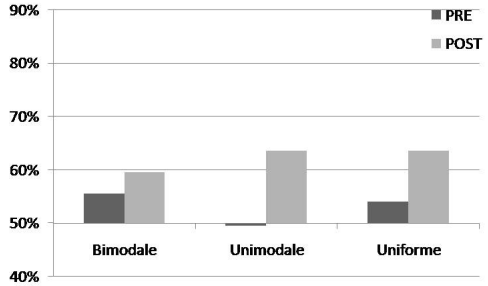

FIGURE 3 – Pourcentages de discriminations correctes pour le contraste -15/-45 lors du pré-test (gris foncé) et lors du post-test (gris clair) pour les trois groupes entraînés.

3.3 Contraste +20/+40 ms DEV

Le facteur Session était ici encore significatif ($F(1,27) = 13.9$; $p < .005$; $\eta^2 = .34$), les participants obtenant en moyenne 10% de discriminations correctes de plus lors du post-test (voir figure 4). Le facteur groupe n'était pas significatif ($F < 1$) et n'interagissait pas avec le facteur Session ($F < 1$).

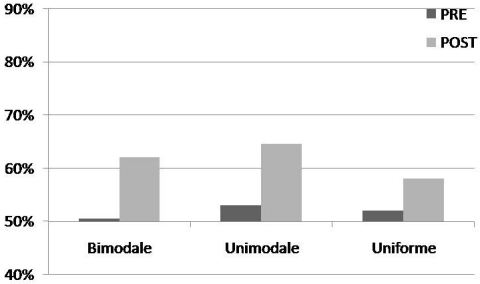

FIGURE 4 – Pourcentages de discriminations correctes pour le contraste +20/+40 lors du pré-test (gris foncé) et lors du post-test (gris clair) pour les trois groupes entraînés.

3.4 Contraste -20/-40 ms DEV

Comme l'illustre la Figure 5, aucun effet n'était ici significatif (facteur Session : $F(1,27) = 2.6$; $p = .012$; $\eta^2 = .09$; facteur Groupe : $F < 1$). Il n'y avait pas non plus d'interaction entre les deux facteurs ($F < 1$).

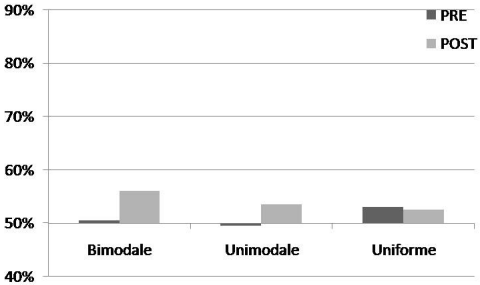

FIGURE 5 – Pourcentages de discriminations correctes pour le contraste -20/-40 lors du pré-test (gris foncé) et lors du post-test (gris clair) pour les trois groupes entraînés.

Au vu de ces résultats, nous avons voulu déterminer si certains contrastes étaient plus difficiles à percevoir en pré-test. Les résultats montrent une différence significative entre les quatre contrastes au pré-test ($F(3,116) = 7.4$; $p < .0005$). Les contrastes a posteriori (correction de Bonferroni) indiquent que +15/+45 diffère significativement des trois autres (+20/+40 : $p < .0005$; -15/-45 : $p = .001$; -20/-40 : $p < .0005$). Aucune autre différence entre contrastes n'a été observée.

4 Discussion

Les résultats de cette étude montrent à la fois les limites des entraînements statistiques mais aussi les limites que peut avoir le système perceptif dans l'extraction de l'information en vue d'améliorer la discrimination des contrastes allophoniques séparés

par de fines différences acoustiques malgré ses capacités à discriminer des différences aussi fines que 3 ms (Samuel, 1977).

L'analyse des trois premiers contrastes montre une augmentation systématique et significative des performances de discrimination après la phase de stimulation et ce quel que soit le type de stimulation proposé. Contrairement aux conclusions tirées par Maye et Gerken (2000 ; 2001), mettant en évidence l'amélioration des performances de discrimination et la construction des catégories phonologiques suite à l'exposition à une distribution bimodale, nos résultats indiquent que l'information statistique n'est pas strictement nécessaire afin d'améliorer la perception de stimuli allophoniques séparés par de fines différences acoustiques. En effet, dans le cadre de cette étude, les participants ne semblent pas tirer profit exclusivement de la distribution bimodale, les performances de discrimination étant également meilleures après exposition aux distributions unimodales et uniformes.

Par contre, pour le dernier contraste (i.e. -20/-40), les performances des participants ne s'améliorent dans aucune des conditions. Cette différence pourrait s'expliquer par la conjonction de deux arguments, l'un acoustique, l'autre psychoacoustique.

Selon l'argument acoustique, il se pourrait que l'amélioration de la perception de contrastes allophoniques soit plus difficile lorsque la différence acoustique entre les deux stimuli est réduite. Cette hypothèse permettrait d'expliquer l'augmentation des performances dans le cas du contraste -15/-45 (30 ms de différence acoustique) alors que le contraste -20/-40 (20 ms de différence acoustique) reste difficilement discriminable même après stimulation. Cependant, cet argument ne permet pas à lui seul d'expliquer l'absence d'augmentation des performances pour ce dernier contraste, son équivalent du côté positif du continuum (i.e. $+20/+40$) présentant quant à lui une augmentation des performances.

L'argument psychoacoustique contribue à mieux comprendre cette différence. En effet, il faut tenir compte de l'effet de masquage que peuvent exercer les hautes fréquences sur les basses fréquences (Aslin, Pisoni, & Hennessy, 1981 ; Burnham, Earnshaw, & Clark, 1991 ; Pisoni, 1977). Dans le cas du DEV négatif, une composante de basse fréquence (prévoisement) précède une composante de haute fréquence (le relâchement de l'occlusion). A l'inverse, lorsque les valeurs de DEV sont positives, le relâchement de l'occlusion (haute fréquence) précède le voisement (basse fréquence). Ces résultats comportementaux ainsi que d'autres données neurophysiologiques (Hoonhorst, Serniclaes, Collet, Colin, Markessis et al., 2009) soutiennent l'idée que cette asymétrie perceptive est basée sur des mécanismes auditifs généraux. Ceci permettrait d'expliquer pourquoi les performances de discrimination autour du contraste -20/-40 n'augmentent pas alors que celles liées au contraste $+20/+40$ ms DEV augmentent.

Notons que la conjonction de ces deux arguments pourrait également expliquer pourquoi, lors du pré-test, les participants sont meilleurs pour le contraste $+15/+45$ que pour les trois autres.

5 Conclusion

Les résultats de cette expérience permettent de mettre en évidence les limites de l'extraction de l'information statistique par notre système perceptif. En effet, chez des

locuteurs francophones, dès 30 ms de DEV de différence acoustique, le système perceptif ne semble plus tirer avantage de cette information. La simple présentation répétée de stimuli semble suffisante pour améliorer les performances de discrimination. De plus, pour de faibles différences acoustiques (20 ms de DEV), l'extraction d'information peut s'avérer impossible (en particulier pour les valeurs de DEV négatives) et conduire à l'absence d'amélioration des performances comportementales.

Références

Aslin, R. N., Pisoni, D. B., Hennessy, B. L., & Perey, A. V. (1981). Discrimination of voice onset time by human infants: New findings and implications for the effect of early experience. *Child Development, 52,* 1135-1145.

Carré, R. (2004). Program SyntFormVoy. Laboratoire Dynamique du Language, CNRS, Lyon, France.

Burnham, D. K., Earnshaw, L. J., & Clark, J. E. (1991). Development of categorical identification of native and non-native bilabial stops: Infants, children and adults. *Journal of Child Language, 18,* 231–260.

Guenther, F. H., & Gjaja M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America, 100(2),* 1111-1120.

Hoonhorst, I., Serniclaes, W., Collet, G., Colin, C., Markessis, E., Radeau-Loicq, M., & Deltenre, P. (2009). N1b and Na subcomponents of the N100 long latency auditory evoked-potential: neurophysiological correlates of voicing in French-speaking subjects. *Clinical Neurophysiology, 120,* 897-903.

Jusczyk, P. W., Luce,P., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language, 33,* 630-645.

Maye, J., & Gerken, L. (2000). Learning phonemes without minimal pairs. In S. C. Howell, S. A. Fish & T. Keith-Lucas (Eds.), *Proceedings of the 24th Boston University Conference on Language Development* (pp. 522-533). Somerville, MA: Cascadilla Press.

Maye, J., & Gerken, L. (2001). Learning phonemes: how far can the input take us? In A.H-J. Do, L. Domínguez, & A. Johansen (Eds.), *Proceedings of the 25th Annual Boston University Conference on Language Development* (pp. 480-490). Somerville, MA: Cascadilla Press.

Peperkamp S., Pettinato, F. & Dupoux, E. (2003). Allophonic variation and the acquisition of phoneme categories. In B. Beachley, A. Brown, & F. Conlin (Eds.), Proceedings of the 27th Annual Boston University Conference on Language Development, vol. II (pp. 650-661). Somerville, MA: Cascadilla Press.

Pisoni, D. B. (1977). Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops. *Journal of the Acoustical Society of America*, *61,* 1352-1361.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274,* 1926–1928.

Samuel, A. G. (1977). The effect of discrimination training on speech perception: noncategorical perception. *Perception & Psychophysics, 22*, 321–330.

Serniclaes, W. (2011). Features are phonological transforms of natural boundaries. In G. N. Clements and R. Ridouane (Eds.), *Cognitive, physical and developmental bases of distinctive speech categories* (pp. 237-257). London, UK: John Benjamins.