UNIVERSITÉ DU QUÉBEC À MONTRÉAL

LA POURSUITE AUDITIVE DU MOUVEMENT ACOUSTIQUE VERS L'ACQUISITION DES CATÉGORIES PHONÉTIQUES

THÈSE
PRÉSENTÉE
COMME EXIGENCE PARTIELLE
DU DOCTORAT EN PSYCHOLOGIE

PAR
BRUNO GAUTHIER

JUIN 2009

REMERCIEMENTS

Ceci est un travail d'équipe, merci à tous et à la complémentarité de chacun. Merci à Rushen Shi, professeure à l'Université du Québec à Montréal (UQàM), ma directrice et instance linguistique, pour voir grand et précis à la fois, pour la rigueur et la flexibilité, la confiance et le support. Merci au professeur Robert Proulx (UQàM), mon co-directeur et influence de la pensée et de sa mécanique, pour la connexion mathématique, la zone proximale didactique et l'autonomie.

Merci à la professeure Lucie Ménard (UQàM), membre du Jury, pour le savoir phonétique et l'encouragement. Merci au professeur Yi Xu (*University College London*), phonéticien et collaborateur, pour les données de production, à l'oral comme à l'écrit. Merci au professeur Peter Scherzer (UQàM), membre du Jury et assise neuropsychologique, en théorie et en pratique, pour le respect et la sollicitude. Merci à la professeure Linda Polka (Université McGill), membre du jury et inspiration initiale pour la parole chez l'enfant. Merci à mes collègues du tout début, Sylvain Chartier pour la générosité, Christian Thériault pour la passion du domaine et de la vie, Suzanne Mackay pour le soutien, le calme et la douceur. Merci à Marie-Ève Bouchard pour l'accompagnement de parcours et la vision future.

Merci à mes amis à qui je pourrai maintenant me consacrer, Judith le bon sens et l'émotion, Pierre et Frank les fauteurs de troubles et amis de toujours, Annie la joie de vivre, André la simplicité, et André l'intelligencia. Merci à ceux à qui je dois tout, à ma mère Nicole et mes frères Jean-Pierre et Nicolas, à mon regretté père qui continue de m'épauler, et à Michel, mon bonheur constant et mon emprise sur l'avenir.

Pour le soutien financier, merci au Laboratoire d'Études en Intelligence Naturelle et Artificielle, au Groupe de Recherche sur le Langage et à la Faculté des Sciences Humaines de l'UQàM; au Fonds pour la formation de chercheurs et l'aide à la recherche (FCAR) pour la bourse de maîtrise et de doctorat et l'appui à la multidisciplinarité; au Centre de Recherche sur le Langage, l'Esprit et le Cerveau de l'Université McGill; à l'équipe du *Boston University Conference on Language Development*. Enfin, un merci particulier au Consulat d'Angleterre et à Emmanuel Dupoux et Anne Christophe du Laboratoire de Science Cognitive et Psycholinguistique de l'École Normale Supérieure de Paris, pour une expérience inoubliable.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

RÉSUMÉ

Cette thèse explore le développement de la parole chez l'enfant durant la première année de vie. Elle vise précisément à caractériser le mécanisme à la base de l'acquisition des catégories phonétiques. Les nombreuses recherches dans le domaine offrent un portrait compréhensible de la trajectoire développementale de la perception de la parole. Les mécanismes de ce développement demeurent toutefois mal compris, particulièrement en ce qui a trait aux stratégies de l'enfant pour faire face au problème de la variabilité. En s'inspirant des théories de l'invariance chez l'adulte et sur la base de données empiriques chez l'enfant, cette thèse présente trois études visant à soutenir l'hypothèse selon laquelle la poursuite auditive du mouvement acoustique sous-tend l'acquisition des catégories phonétiques.

Tout d'abord, deux études de modélisation simulent l'acquisition d'un type particulier de catégories phonétiques, les tons lexicaux, par le biais de réseaux neuronaux artificiels de type non supervisé. Ces simulations évaluent l'impact de diverses sources de variabilité et l'efficacité du mouvement acoustique sur la catégorisation des tons en chinois mandarin. Les résultats montrent que malgré un degré modéré de variabilité, les patrons de fréquence fondamentale présentent des régularités permettant de distinguer les quatre tons mandarins, sans information préalable quant au nombre de catégories à découvrir. Ceci suggère que le signal acoustique continu peut suffire à l'acquisition des tons lexicaux, sans besoin de faire appel à un ensemble de propriétés phonétiques abstraites. En présence de multiples sources de variabilité cependant, l'information spectrale du signal de surface n'entretient qu'une faible relation avec les sons de la parole recherchés. À l'opposé, l'information dynamique (les profils de vélocité de la fréquence fondamentale) permet d'atteindre un niveau de performance comparable à celui de l'adulte pour l'identification des tons. De plus, les quatre profils de vélocité découverts par le réseau neuronal correspondent aux quatre tons mandarins et permettent de caractériser les gestes invariants impliqués dans la production tonale.

Afin de vérifier si l'enfant peut faire usage de cette stratégie dynamique pour normaliser le signal de la parole, une étude comportementale examine ensuite la capacité d'enfants préverbaux à percevoir des variations acoustiques reflétant une contrainte articulatoire. Une procédure de regard préférentiel est utilisée afin de vérifier si des enfants de 4 et 8 mois peuvent distinguer entre eux des patrons d'intonation possibles et impossibles sur le plan articulatoire et produits par un locuteur inconnu. Les résultats montrent que les enfants des deux groupes d'âge écoutent plus longuement les stimuli possibles, indiquant qu'ils peuvent détecter des variations subtiles de vélocité de la fréquence fondamentale et préfèrent les variations qui respectent la contrainte articulatoire. Ces résultats suggèrent qu'en bas âge déjà, les enfants peuvent calculer la première dérivée d'informations spectrales continues et réduire la variabilité interlocuteur à partir de la dynamique du signal acoustique.

Le modèle proposé par les études de simulations permet d'établir l'efficacité de l'information dynamique dans le développement phonétique. L'étude comportementale permet pour sa part de vérifier la sensibilité à cette information chez l'enfant bas âge. Ces résultats suggèrent que l'invariance se situe à la fois au niveau acoustique/auditif, moteur et développemental, et que la poursuite auditive du mouvement acoustique reflétant les gestes articulatoires représente une stratégie efficace pour l'acquisition des catégories phonétiques.

Mots clés : Catégories phonétiques, apprentissage distributionnel, perception de la parole, production de la parole, acquisition du langage, réseaux neuronaux artificiels non-supervisé.

INTRODUCTON

L'étude de la perception de la parole chez l'enfant a débuté dans les années 70, avec objectifs principaux de déceler les capacités perceptives initiales, d'objectiver le développement de ces capacités et de découvrir l'âge à partir duquel l'enfant traite la parole comme l'adulte le fait. Les enfants abordent initialement la parole en termes de patrons acoustiques globaux et présentent certains biais auditifs qui rappellent à plusieurs égards les propriétés de la perception de la parole chez l'adulte. Les petits se distinguent toutefois des grands quant à leur capacité initiale à distinguer la plupart des sons des langues du monde. Durant la première année de vie cependant, cette capacité devient spécifique au système phonétique de leur langue maternelle. Alors que la recherche actuelle continue de tracer la trajectoire du développement phonétique, plusieurs se penchent maintenant sur la façon dont la perception de la parole évolue. Cette thèse vise à caractériser les mécanismes à la base du développement de la perception de la parole chez l'enfant en bas âge.

Les mécanismes de la perception de la parole suscitent un intérêt depuis bon moment déjà, par exemple afin de déterminer la part de l'inné et de l'acquis dans l'acquisition du langage, ou de préciser la nature générale ou spécialisée du traitement de la parole. Plusieurs chercheurs tentent d'élucider cette question à l'aide d'une approche comparative, en examinant si seulement les humains traitent le signal de la parole comme ils le font, et si seulement le signal de la parole est traité par l'humain comme il l'est. Par exemple, la perception catégorielle des sons de la parole, l'une des premières propriétés de la perception adulte observée chez l'enfant, reflétait à l'origine l'unicité de la capacité innée du langage chez l'humain. Des études subséquentes ont toutefois montré que le chinchilla, le singe macaque et la caille japonaise répondent de façon catégorielle à la parole, et l'humain de façon catégorielle aux sons musicaux. Ces études renforcent certes le besoin de déterminer si les similarités de surface observées chez l'humain et l'animal partagent le même algorithme (Weiss & Newport, 2006), spécialement en ce qui a trait aux mécanismes de l'acquisition phonétique chez l'enfant. Néanmoins, la différence évidente mais non triviale entre les systèmes de production vocale chez l'humain et l'animal suggère que l'unicité du développement phonétique relève de la relation entre la production et la perception de la parole.

Cette relation pourrait être à la base d'une habileté nécessaire chez l'enfant pour acquérir le système de sons de sa langue maternelle, la constance perceptive. Le signal acoustique que reçoit l'enfant fluctue en fonctions de multiples sources de variabilité. Si l'auditeur adulte ne présente aucune difficulté à faire face à la variabilité du signal de la parole, la façon dont le système perceptif peut résoudre ce problème demeure un défi de taille pour les théories de la perception de la parole et de l'acquisition du langage, et une solution potentielle pour la conception de technologies efficaces du traitement automatique de la parole et du langage. Chez l'adulte, deux approches s'opposent quant au lieu de l'invariance phonétique, situant celui-ci au niveau acoustique ou au niveau moteur. Une troisième approche suggère que les standards phonétiques relèvent à la fois du stimulus proximal et de l'objet distal, le signal acoustique de surface reflétant directement la dynamique articulatoire sous-jacente. Le développement de la parole peut ainsi être conçu comme un système dynamique où interagissent les systèmes de perception et de production. Dans cette thèse, trois études visent à tester l'hypothèse selon laquelle l'information dynamique du signal acoustique reflétant les gestes articulatoires du locuteur permet l'émergence de la constance perceptive des sons de la parole chez l'enfant.

Cette thèse est présentée sous forme de thèse par articles[1]. Tout d'abord, deux études de modélisation simulent la perception, la normalisation et l'acquisition des catégories phonétiques à partir d'un algorithme simple et plausible sur les plans biologique et psychophysique. Plus précisément, l'approche connexionniste est utilisée afin de simuler l'acquisition des tons lexicaux, un type particulier de catégories phonétiques, à partir d'une dimension particulière du signal de la parole, la fréquence fondamentale. L'objectif principal vise à proposer un algorithme qui permette de catégoriser les sons de la parole malgré diverses sources importantes de variabilité. Une étude de perception vise ensuite à vérifier si,

---

[1] Publication 1: Gauthier, B., Shi, R., & Xu, Y. (2007a). Learning phonetic categories by tracking movements. *Cognition, 103*(1), 80-106.

Publication 2 : Gauthier, B., Shi, R., & Xu, Y. (2007b). Simulating the acquisition of lexical tones from continuous dynamic input. *Journal of the Acoustical Society of America, 121*(5), EL190-EL195.

Publication 3 : Gauthier, B. & Shi, R. (soumis). The perception of acoustic movement in preverbal infants. *Journal of the Acoustical Society of America.*

et à quel âge, l'enfant peut recourir aux mécanismes proposés dans les études de modélisation en vue d'apprendre les catégories phonétiques de sa langue maternelle. À l'aide d'une procédure de regard préférentiel, cette expérience explore le développement chez l'enfant de la poursuite auditive du mouvement acoustique, ou sa capacité à détecter l'information dynamique du signal acoustique, et vise à vérifier si l'enfant est en mesure d'utiliser cette stratégie pour normaliser la variabilité du signal de la parole.

Cette thèse est divisée en cinq chapitres. Le Chapitre 1 présente une recension des écrits du domaine de la perception de la parole chez l'enfant, exposant les aspects empiriques et théoriques du développement de la perception phonétique. La première section y décrit les capacités perceptives initiales, le développement de ces capacités, l'unité de perception et les capacités de normalisation de la parole durant la première année de vie. Cette section présente également un cadre théorique général unissant les connaissances actuelles sur l'acquisition phonétique. La Section 2 présente ensuite les principaux modèles de la perception de la parole chez l'enfant et les théories de la perception de la parole et de la normalisation chez l'adulte. La troisième section aborde la linguistique et la phonétique des tons lexicaux en général et du système de tons chinois mandarin en particulier, suivi d'une recension des études portant sur l'acquisition tonale chez l'enfant et sur la façon dont le problème de la variabilité s'applique à ce type de catégorie phonétique. Les buts et les hypothèses spécifiques de cette thèse complètent le premier chapitre. Les trois chapitres qui suivent présentent les trois études précédemment décrites. Enfin, le Chapitre 5 présente une discussion des résultats obtenus dans chaque étude et propose un modèle permettant de faire le pont entre la perception de la parole, la production de la parole et l'acquisition du langage.

1 CONTEXTE THÉORIQUE

1.1 La perception de la parole durant la première année de vie

1.1.1 Habiletés perceptives initiales

Inaugurant le domaine de la perception de la parole chez l'enfant avec une étude désormais classique, Eimas et collaborateurs (Eimas, Siqueland, Jusczyk, & Vigorito, 1971) ont exploré le moment où l'enfant peut traiter la parole à la façon d'un adulte. Les chercheurs ont examiné la capacité d'enfants de 1 et 4 mois à distinguer entre des sons synthétiques de la parole variant le long de la dimension acoustique du moment du début du voisement (*voice-onset-time*, VOT). Dans les consonnes, le VOT correspond à l'intervalle temporel entre le dégagement articulatoire et le début de la vibration des cordes vocales. L'adulte utiliserait cet indice acoustique afin de distinguer les consonnes voisées et non-voisées en anglais (Liberman, Delattre, & Cooper, 1958). Eimas et collaborateurs ont employé le paradigme de succion à haute amplitude (*high-amplitude-sucking*, HAS) (Siqueland & Delucua, 1969). Cette procédure consiste d'abord à établir le taux de base de succion non nutritive chez l'enfant, et à lui présenter ensuite des stimuli auditifs répétitifs conditionnels à la succion d'une tétine. Une augmentation de la succion relativement au taux de base indique le début de la phase d'habituation. Lorsque la réponse de l'enfant diminue sous un seuil préétabli, indiquant l'habituation aux stimuli, les enfants du groupe contrôle continuent d'entendre les stimuli d'habituation, alors que le groupe expérimental fait face à de nouveaux stimuli. Une augmentation de la réponse du groupe expérimental relativement au groupe contrôle suggère que l'enfant peut discriminer les stimuli pré- et post-habituation.

Un second but de l'étude de Eimas et collaborateurs (1971) était d'examiner si les enfants en bas âge perçoivent les consonnes de façon catégorielle, comme les adultes le font. La perception du VOT est catégorielle du fait que l'adulte peut distinguer deux stimuli chevauchant une frontière le long du continuum de VOT et les identifier comme appartenant à différentes catégories, alors que des stimuli de même différence acoustique mais qui n'empiètent pas la frontière sont perçus comme le même son de la parole (Liberman et al., 1958). En anglais, la frontière phonémique des consonnes voisées et non-voisées chez l'adulte se situe à environ 30 millisecondes après le dégagement articulatoire. Deux groupes

d'enfants ont donc été constitués pour la condition expérimentale. Les stimuli pré- et post-habituation différaient d'un VOT de 20 millisecondes dans les deux conditions. Pour le groupe *intra-catégorie*, les stimuli pré- et post-habituation [ba] ou [pa] appartenaient à la même catégorie, avec des paires de VOT de -20 et 0 ou de 60 et 80 millisecondes. Le groupe *inter-catégorie* faisait face à des valeurs de VOT inter-phonémiques, c.-à-d. 20 et 40 millisecondes. Les résultats indiquent que le groupe contrôle n'a pas récupéré de l'habituation, et que le groupe inter-catégorie a récupéré de façon significative au changement de stimulus relativement au groupe intra-catégorie. Ces résultats suggèrent que les enfants en bas âge peuvent distinguer des différences subtiles de VOT, qu'ils le font de façon catégorielle, et que leur frontière de VOT ressemble à celle des adultes.

Eimas (1975) a ensuite montré que les enfants en bas âge distinguent également des différences de VOT de façon catégorielle pour les consonnes occlusives dentales [d] et [t], et que la perception d'enfants de 2 mois pour des occlusives qui diffèrent en termes de lieu d'articulation ([b] - [g]), caractéristique articulatoire reflétée par la transition du second formant (F2)[2], est également catégorielle (Eimas, 1974). Une autre étude a montré que les enfants de 2 mois peuvent discriminer les consonnes fricatives dont le lieu d'articulation diffère ([f] - [θ] et [v] - [ð]), et que cette discrimination est catégorielle (Levitt, Jusczyk, Murray, & Carden, 1988).

À la même période, Trehub (1973) s'est penchée sur la sensibilité précoce aux contrastes vocaliques en examinant à l'aide de la procédure HAS la capacité d'enfants âgés de 4 à 17 semaines à distinguer les contrastes [a] - [i] et [i] - [u]. Les stimuli étaient des exemplaires naturels produits de façon isolée ou incorporés dans des syllabes consonne-voyelle (CV) avec la consonne occlusive [p] ou [t]. Les résultats indiquent que les enfants peuvent distinguer chaque contraste, présenté en isolation ou dans divers contextes consonantiques. Dans une étude subséquente et utilisant la même procédure, Trehub (1976) a montré que les enfants de 5 à 17 semaines peuvent également distinguer entre les voyelles contrastant selon le mode d'articulation (oral versus nasal : [a] - [ã]). Ces études, cependant, n'ont pas examiné si la

[2] Les formants correspondent à des régions de haute énergie dans le signal de la parole et sont déterminés par les fréquences de résonances de l'appareil vocal.

perception des contrastes vocaliques est catégorielle chez l'enfant, c.-à-d. s'ils peuvent aussi bien distinguer des différences vocaliques inter- qu'intra-catégorielles.

L'adulte percevrait pour sa part les voyelles de façon continue lorsqu'il a accès à l'information présente en mémoire auditive à court terme (Pisoni, 1973). Afin de vérifier si les enfants perçoivent également les voyelles de façon continue, Swoboda, Morse and Leavitt (1976) ont évalué la capacité d'enfants âgés de 8 semaines à distinguer le contraste [i] - [I]. Les stimuli (empruntés de Pisoni, 1973) comprenaient quatre voyelles synthétiques dont les trois premiers formants variaient de [i] à [I] en échelons logarithmiques égaux. Les stimuli ont d'abords été présentés à six adultes, qui ont tous identifié deux stimuli comme [i] et les deux autres comme [I]. Les chercheurs ont ensuite examiné la perception de diverses paires de stimuli chez les enfants à l'aide la procédure HAS. Un groupe a reçu un changement inter-catégorie, l'autre groupe un changement intra-catégorie, et le groupe contrôle n'a reçu aucun changement. Les résultats montrent une augmentation significative de la réponse suite au changement de stimuli pour les deux groupes expérimentaux relativement au groupe contrôle, mais aucune différence entre la discrimination de paires intra- et inter-catégories, suggérant que le système perceptif précoce traite les voyelles de façon continue, comme l'adulte.

Ces études suggèrent que les enfants en bas âge possèdent des capacités de discrimination auditives/perceptives leur permettant de distinguer plusieurs sons de la parole, et que cette capacité dispose des propriétés non-linéaire et continue observées dans la perception des consonnes et des voyelles chez l'adulte. Ces résultats, et particulièrement ceux ayant trait à la perception catégorielle, étaient à l'origine interprétés comme reflétant l'existence de frontières phonémiques innées et de contraintes imposées par un module spécialisé du traitement de la parole (Eimas et al., 1971). Cependant, des études subséquentes ont fournit trois types d'évidence défiant cette interprétation. D'abord, la discrimination catégorielle a été observée pour des sons autres que ceux de la parole chez l'adulte (Pisoni, 1977) et l'enfant (Jusczyk, 1980; Jusczyk, Rosner, Reed, & Kennedy, 1989). Ensuite, des similarités de réponse aux sons de la parole chez l'humain et l'animal ont été observées. Par exemple, des chinchillas ont été entraînés avec des exemplaires synthétiques de [ba] - [pa] provenant des extrêmes d'un continuum de VOT (10 et 80 millisecondes) (Kuhl & Miller, 1975). Les fonctions de discrimination résultantes étaient presque identiques à celles

d'auditeurs anglophones adultes en termes d'inclinaison et de valeur absolue de localisation des frontières. Des résultats similaires ont été obtenus avec des macaques pour la perception du contraste de voisement (Kuhl & Padden, 1982), ainsi que pour la perception du contraste de lieu d'articulation, également chez le macaque, avec des fonctions de discrimination pour [b] - [d] et [d] - [g] très similaires à celles d'auditeurs adultes (Kuhl & Padden, 1983). Ces résultats suggèrent la présence de frontières perceptives innées soutenues par des mécanismes auditifs généraux plutôt que par une connaissance phonémique préétablie. À l'appui de cette hypothèse, des études inter-linguistiques explorant d'autres contrastes de la parole dans diverses langues indiquent que les frontières catégorielles des enfants ne correspondent pas toujours à celles de leur langue maternelle (Lasky, Syrdal-Lasky, & Klein, 1975; Streeter, 1976), suggérant que l'expérience façonne la perception auditive selon la structure de la langue ambiante.

1.1.2 Développement de la perception de la parole

En conjonction avec la découverte du rôle de l'expérience dans le développement phonologique initial, un champ d'étude important émergea dans le domaine du développement de la perception de la parole dans les années 80. S'éloignant de la question de la perception catégorielle, de nouvelles études visaient alors à spécifier l'âge auquel la sensibilité générale pour les sons paroliers devient spécifique à la langue maternelle. Werker et Tees (1984) ont montré que la capacité de distinguer des consonnes non-natives diminue vers la fin de la première année de vie. En fait, Eilers, Wilson et Moore (1977) se sont d'abord penchés sur la question et ont observé des changements perceptifs chez l'enfant durant cette période. Cependant, le recours à différentes procédures pour examiner différents groupes d'âge, à savoir HAS à 3 mois et la *Conditioned Headturn Procedure* (CHP) à 6 mois, peut expliquer la réponse différente de ces deux groupes d'âge. Néanmoins, l'introduction de la CHP représente une avancée méthodologique importante dans l'étude de la perception de la parole chez l'enfant. Cette procédure implique la présentation d'un son répétitif et le renforcement visuel du mouvement de la tête vers un jouet animé lorsque le son change. À l'aide de cette procédure, Werker et Tees (1984) ont constaté chez des enfants de 6 à 8 mois apprenant l'anglais la capacité de distinguer des contrastes natifs ([b] - [d]) ainsi que des contrastes non-natifs (Hindi [t] - [ʈ] et Nthlakampx [k̓] - [q̓]), observations compatibles avec

les habiletés générales de discrimination déjà documentées. Cependant, les enfants de 10 à 12 mois ne distinguaient plus que les contrastes natifs, suggérant que l'absence d'exposition aux contrastes non-natifs résulte en une perte de sensibilité pour ceux-ci, phénomène connu sous le nom de réorganisation perceptive.

Une réorganisation perceptive semble également s'opérer sur les voyelles. Polka et Werker (1994) ont examiné à l'aide d'une variante de CHP la capacité d'enfants de 6-8 et 10-12 mois apprenant l'anglais à distinguer deux contrastes vocaliques allemands ([Y] - [U] et [y] - [u]). Les stimuli ont été produits par un locuteur masculin dans le contexte [dVt]. Les résultats montrent une meilleure performance chez les plus jeunes enfants, toutefois en deçà de celle rapportée précédemment pour des contrastes consonantiques non-natifs. À partir des mêmes stimuli, les chercheurs ont donc examiné des enfants de 4 et 6 mois avec la procédure d'habituation visuelle, qui consiste à présenter des exemplaires d'une catégorie sonore de façon contingente à un damier télévisé (Best, McRoberts, & Sithole, 1988). Lorsque le temps de regard diminue en-deçà d'un critère préétabli, indiquant l'habituation, une nouvelle catégorie sonore est présentée. Une augmentation du temps de regard au changement de catégorie indique la discrimination des deux sons paroliers. Les résultats montrent que seulement les plus jeunes enfants distinguent les deux contrastes allemands, suggérant une réorganisation perceptive des voyelles, mais à un plus jeune âge que pour les consonnes.

En se basant sur la théorie générale du développement de Gottlieb (1983), Aslin, Pisoni, et Jusczyk (1983) ont proposé quatre processus à la base du développement de la perception phonétique. Ces processus sont illustrés à la figure 1.1, où les lignes continues correspondent aux fonctions distinctives initiales et les lignes pointillées aux changements de ces fonctions.

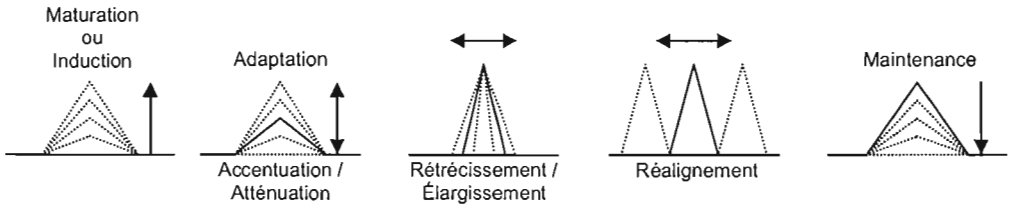


Figure 1.1 Processus à la base du développement de la perception de la parole durant la première année de vie (d'après: Aslin et al., 1983).

La *maturation* se rapporte aux transformations physio-anatomiques préprogrammées qui permettent éventuellement au système de la parole d'accomplir sa fonction. À l'opposé, l'*induction* repose entièrement sur l'expérience pour construire la structure perceptive des sons de la parole. La *maintenance* se rapporte aux capacités distinctives innées et universelles pour tous les sons paroliers, capacités ensuite perdues ou conservées en fonction de la demande extérieure. Enfin, l'*adaptation* se rapporte aux capacités distinctives générales qui sont ensuite modulées par l'expérience. Le processus d'adaptation se subdivise en ceux d'accentuation et atténuation, rétrécissement et élargissement, et réalignement (Aslin et al., 1983).

Afin de rendre compte du développement perceptif durant la première année de vie, Werker et Lalonde (1988) ont proposé le Modèle de Réorganisation Fonctionnelle, dont la version initiale reposait sur le processus de maintenance. De nouvelles évidences, par exemple la capacité d'adultes à recouvrir une sensibilité aux sons non-natifs suite à un entraînement adéquat (Flege, 1989; Logan, Lively, & Pisoni, 1991), ont toutefois mené les auteurs à adopter une interprétation différente. Ainsi, la capacité de distinguer les contrastes non-natifs semble s'atténuer avec l'expérience plutôt que de disparaître.

Dans le but spécifique d'examiner le rôle de l'induction dans le développement phonétique, Maye, Werker and Gerken (2002) ont manipulé les propriétés distributionnelles de stimuli expérimentaux afin d'en observer l'effet sur les habiletés de discrimination chez l'enfant. À l'aide d'une variation de la Procédure de Regard Préférentiel (Jusczyk & Aslin, 1995), les chercheurs ont exposé des enfants de 6 et 8 mois à différentes distributions de syllabes se situant sur le continuum phonétique des consonnes plosives alvéolaires non-aspirées [da] - [ta], toutes deux formant une seule catégorie de plosive voisée en anglais. Dans la condition *bimodale*, des enfants apprenant l'anglais ont été familiarisés avec une majorité d'exemplaires situés aux extrémités du continuum. La condition *monomodale* impliquait une majorité de stimuli occupant le centre du continuum. L'étape suivante consistait à examiner la capacité des enfants à distinguer deux exemplaires situés près des extrémités du continuum. Les résultats montrent une meilleure capacité à distinguer le contraste de VOT chez les enfants de la condition bimodale. Selon les auteurs de cette recherche, ceci indique que les enfants ont formé une ou deux catégories selon le type

d'exposition, soutenant l'hypothèse selon laquelle l'expérience modifie la perception en fonction des propriétés distributionnelles du signal de la parole. Ceci soulève la question de savoir si un tel mécanisme d'apprentissage est spécifique aux humains. Récemment, la capacité de former des frontières catégorielles selon les propriétés distributionnelles du signal de la parole a également été observée chez le rat (Pons, 2006), suggérant que l'apprentissage statistique ne relève pas spécifiquement du langage.

Plus récemment, Kuhl et al. (2006) ont exploré le rôle de l'accentuation dans le développement de la perception de la parole, prédisant une facilitation pour la perception de contrastes natifs parallèlement au déclin de sensibilité pour les contrastes non-natifs. À l'aide de la procédure CHP, les chercheurs ont examiné la capacité d'enfants de 6 à 8 et de 10 à 12 mois apprenant l'anglais ou le japonais à distinguer le contraste anglais américain [r] - [l], qui n'est pas phonémique en japonais. Alors que les plus jeunes ont réussi dans les deux langues, seul les plus vieux du groupe anglophone pouvait encore percevoir le contraste, répliquant le déclin de sensibilité pour les contrastes non-natifs. En outre cependant, les enfants anglophones de 10-12 mois ont montré une sensibilité supérieure à celle des plus jeunes au contraste natif, suggérant que la réorganisation perceptive implique également la spécification des contrastes de la langue maternelle.

En plus d'étudier la frontière des catégories phonétiques, la recherche s'intéresse également à leur structure interne. Grieser et Kuhl (1989) ont montré que chez l'adulte, deux exemplaires d'une voyelle situés près de son prototype se distinguent moins facilement que deux exemplaires situés en région éloignée du centre vocalique. Afin d'examiner si les enfants manifestent le même effet de perception « magnétique », les chercheurs ont exposé deux groupes d'enfants de 6 mois, l'un avec le prototype d'une voyelle (précédemment jugé comme tel par des adultes), et l'autre avec un exemplaire atypique de la même voyelle. Les deux groupes ont ensuite été testés avec deux nouveaux exemplaires situés à équidistance entre le prototype et l'exemplaire atypique. Tel que prédit, les enfants entraînés avec la voyelle atypique pouvaient mieux distinguer les nouveaux exemplaires. La Théorie Magnétique de la Langue Maternelle a été proposée pour rendre compte de ces résultats (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992). Selon ce point de vue, les centres de catégories, les prototypes, attirent vers eux à la façon d'un aimant les exemplaires

périphériques. Par exemple, les enfants apprenant le suédois traitent en prototype le son suédois [y] et en non-prototype le son anglais américain [i], alors que les enfants anglophones montrent un patron de réponse inverse (Kuhl et al., 1992). La Théorie Magnétique stipule que les enfants naissent pourvus de frontières auditives naturelles leur permettant de distinguer les sons de la parole, et que l'expérience avec la langue maternelle apporte, outre un réajustement des frontières phonétiques, une réorganisation interne des catégories phonétiques. Comme l'effet magnétique ne s'observait que pour les sons de la langue maternelle chez l'adulte comme l'enfant, mais non chez l'animal (Kuhl, 1991), il a été proposé qu'il s'agit-là d'un mécanisme propre à l'acquisition du langage. Une étude subséquente observait toutefois le même effet chez l'étourneau sansonnet (Sturnus vulgaris) (Kluender, Lotto, Holt, & Bloedel, 1998), remettant en question la nature spécifique de l'effet magnétique.

Les études rapportées jusqu'ici indiquent que les enfants naissent pourvus de biais perceptifs généraux leur permettant de distinguer la plupart des sons des langues du monde. Une réorganisation perceptive impliquant divers processus s'installe ensuite durant la première année de vie, à la fin de laquelle s'observent une sensibilité accrue aux contrastes natifs et une diminution de la sensibilité envers les contrastes non-natifs. À ce processus contribue une sensibilité envers les propriétés distributionnelles du signal de la parole, mécanisme d'apprentissage général que partagent l'animal et l'enfant, mais qui n'entre en jeu dans l'acquisition des catégories phonétiques que chez ce dernier.

1.1.3 Unité de traitement de la parole

Une autre question soulevée dans les années 1980 concerne la façon dont les enfants en bas âge encodent les sons de la parole, c.-à-d. globalement sous forme de syllabes, ou selon de plus petites unités comme les segments ou les caractéristiques phonétiques. La syllabe est souvent considérée comme l'unité fonctionnelle primitive en production de la parole (Fujimura, 2000; Krakow, 1999; Xu & Liu, 2006). Acoustiquement, les unités syllabiques correspondent à des patrons spectro-temporels d'une durée d'environ 100 millisecondes. Selon la théorie du Cadre-Contenu (*Frame-Content Theory*, MacNeilage, 1998), le mouvement de base de la syllabe (ouverture-fermeture du mandibule) a évolué pour devenir un outil de communication en y incorporant le contenu segmental (mouvements des articulateurs). Les segments correspondent pour leur part à des fenêtres temporelles d'une

durée d'environ 50 millisecondes. Comme les syllabes, une majorité des segments peuvent facilement être identifiées à partir du spectrographe. Sur le plan articulatoire, les segments correspondent aux unités de base de prononciation, appelées phones, un cas particulier de catégories phonétiques incluant les voyelles et les consonnes, telles que répertoriées dans la charte de l'alphabet phonétique international (IPA) (les tons lexicaux forment un autre type de catégorie phonétique et seront décrits en détails à la Section 1.3). Un segment se définit à son tour par un ensemble de caractéristiques phonétiques, elles-mêmes représentant les divers articulateurs et mouvements de base impliqués dans la production de la parole.

Suite à leur étude classique, Bertoncini et Mehler (1981) ont proposé la syllabe comme unité de segmentation naturelle du signal de la parole chez l'enfant. À l'aide de la procédure HAS, ils ont vérifié la capacité d'enfants de 1 mois à distinguer entre des paires de syllabes (CVC) et entre des paires de non-syllabes (CCC). Un stimulus répétitif était présenté (par exemple, [tap] dans le groupe CVC et [tsp] dans le groupe CCC) jusqu'à l'atteinte d'un critère d'habituation. Les stimuli post-habituation étaient composés des syllabes pré-habituation inversées, de sorte que chaque condition impliquait le même degré de variation (par exemple, [pat] et [pst]). Les résultats montrent que le groupe CVC a mieux réussi la tâche que le groupe CCC.

Plus récemment, Bijeljac-Babic, Bertoncini et Mehler (1993) ont exploré la capacité d'enfants de 4 jours à distinguer des énoncés multisyllabiques à l'aide de la procédure HAS. En supposant que l'unité primitive est la syllabe, la prédiction voulait que les enfants détectent mieux dans un énoncé un changement du nombre de syllabes qu'un changement du nombre de segments phonétiques. Une première expérience a montré la capacité de l'enfant à distinguer entre des listes d'énoncés de deux versus trois syllabes CV. Pour éliminer la possibilité que la longueur entre en jeu, une seconde expérience a montré que la discrimination persistait malgré la réduction des différences de durée. Enfin, une troisième expérience a montré l'incapacité de l'enfant à distinguer entre une liste d'énoncés bisyllabiques contenant quatre versus six segments, suggérant que les enfants portent attention à la structure syllabique plutôt que segmentale des énoncés. Dans l'ensemble, ces résultats sont compatibles avec l'hypothèse selon laquelle la syllabe représente l'unité de traitement de la parole chez l'enfant en bas âge.

En ce qui a trait aux plus petites unités de traitement, Miller et Eimas (1979) ont employé la procédure HAS pour vérifier la capacité d'enfant de 2, 3 et 4 mois à distinguer entre des listes de deux syllabes qui partagent les mêmes caractéristiques articulatoires (par exemple, voisement et lieu d'articulation), mais qui diffèrent dans la combinaison de ces caractéristiques (par exemple, [ba, ta] versus [da, pa]). Malgré la présence des mêmes caractéristiques dans les deux paires de stimuli, les enfants pouvait distinguer la recombinaison de ces caractéristiques, démontrant selon les auteurs de cette recherche qu'ils sont sensibles à la relation entre les caractéristiques phonétiques. Dans une étude similaire, Hillenbrand (1983) a utilisé la procédure CHP afin de déterminer si les enfants de 6 mois peuvent percevoir les caractéristiques phonétiques, et plus précisément s'ils peuvent distinguer entre des ensembles de syllabes qui diffèrent selon une seule caractéristique (par exemple le mode d'articulation (oral-nasal) tel que [ba, da] versus [ma, na]) et des ensembles de syllabes qui diffèrent selon deux caractéristiques (par exemple le mode d'articulation (oral-nasal) et le lieu d'articulation (labial-labiodental-vélaire-uvulaire), tel que [ba, ŋa] versus [na, ga]). Les résultats montrent une meilleure performance chez le premier groupe, constituant pour l'auteur une évidence de la capacité des enfants de 6 mois à percevoir les corrélats acoustiques des caractéristiques consonantiques.

Plus récemment, Jusczyk, Goodman et Baumann (1999) ont utilisé la procédure CHP afin d'examiner la capacité d'enfants de 9 mois à distinguer entre deux types de listes de syllabes. Les listes expérimentales contenaient des syllabes CVC dont la consonne initiale partage le mode d'articulation (par exemple, plosives voisées [b, d, g], fricatives non voisées [f, ʃ, θ], ...), alors que les listes contrôles contiennent des syllabes dont la consonne initiale ne partage aucune caractéristique ([j, t, h, w, s, θ, m, ...]). Les résultats montrent des temps de regards plus longs pour les listes expérimentales que pour les listes contrôles, suggérant que les enfants portent attention à l'information relative aux caractéristiques phonétiques. Cependant, les listes expérimentales contiennent de quatre à six répétitions des mêmes deux ou trois consonnes, alors que les listes contrôles ne présentent la même consonne qu'une seule fois. Les résultats peuvent donc s'expliquer par un attrait envers les segments communs des listes expérimentales plutôt qu'envers les caractéristiques phonétiques communes aux segments initiaux.

Du point de vue de l'apprentissage, Maye et Weiss (2003) ont récemment exploré si la sensibilité des enfants aux propriétés distributionnelles du signal de la parole se base sur les caractéristiques phonétiques ou sur les segments. Lors d'une première expérience, les chercheurs ont constaté que les enfants de 8 mois apprenant l'anglais américain peuvent distinguer le contraste pré-voisé [da] et court voisé (*short lag*) [ta] une fois entraînés à partir d'exemplaires provenant d'une distribution bimodale sur le continuum du VOT, mais non d'une distribution monomodale. Dans une deuxième expérience, les chercheurs ont vérifié la capacité des enfants entraînés avec la même distribution bimodale à distinguer le contraste VOT, cette fois situé sur un lieu d'articulation différent ([ga - ka]). Les résultats montrent une augmentation du temps de regard chez ce groupe relativement à celui du groupe monomodal de la première expérience. Selon les auteurs de cette recherche, les enfants ont généralisé les plosives alvéolaire aux vélaires (et inversement), suggérant qu'ils traitent les caractéristiques phonétiques afin de catégoriser les sons de la parole.

En somme, les évidences suggèrent que les enfants en deçà de 4 mois traitent le signal de la parole en termes d'unités syllabiques, qu'ils commencent à traiter l'information sous-segmentale vers l'âge de 6 mois, et pourrait dès l'âge de 8 mois employer cette information en vue d'acquérir les catégories phonétiques de leur langue maternelle. Toutefois, une autre interprétation suggère que les enfants, plutôt que d'extraire les caractéristiques phonétiques du signal, remarquent simplement la similarité acoustique entre des portions de taille segmentale ou syllabique. Afin de résoudre ce débat, il pourrait suffire de questionner la nécessité d'un mécanisme d'extraction de caractéristiques pour l'acquisition phonétiques, ou de façon alternative de vérifier si des portions ininterrompues du signal acoustique peuvent suffire à la tâche. Une réponse positive à cette question indiquerait que les catégories phonétiques peuvent être dérivées du signal acoustique continu, et que l'extraction des caractéristiques phonétiques peut survenir ensuite. Une telle stratégie, cependant, fait face au problème de la variabilité.

1.1.4 Le problème de la variabilité

La parole est produite dans une variété de contextes, qui sur le plan acoustique entraînent pour une même information des fluctuations importantes, ou au contraire des patrons similaires pour différentes catégories phonétiques. Du point de vue psycholinguistique, le

problème de la variabilité se rapporte à la constance perceptive, et précisément au mécanisme à la base de la catégorisation des sons de la parole (Perkell & Klatt, 1986; Pisoni, 1997). Chez l'enfant, le problème de la variabilité de la parole a été soulevée voilà plus d'une demi siècle (Irwin, 1947), et il demeure l'un des grands défis pour les théories de l'acquisition du langage (Pierrehumbert, 2003).

Deux types de variabilité peuvent être distingués. La variabilité aléatoire (*unlawful*), imprévisible, et la variabilité systématique (*lawful*), c.-à-d. les fluctuations récurrentes, mais non manifestes du signal (Studdert-Kennedy, dans Perkell & Klatt, 1986). Ce second type de variabilité est considéré utile pour la perception de la parole (par ex., Elman & McClelland, 1986), bénéfique pour l'acquisition du langage (Singh, 2003) et nécessaire pour la formation de catégories (Needham, Dueker, & Lockhead, 2005). Ceci expliquerait pourquoi une majeure partie de la recherche porte sur la variabilité systématique, qui vise à en identifier les sources et les conséquences.

Une source majeure de variabilité systématique se rapporte aux variations volontaires et linguistiquement déterminées. L'identité vocalique, par exemple, expliquerait en majeure partie les fluctuations du signal dans la production des voyelles (Nearey, 1989). Une seconde source intrusive de variabilité, l'une des plus largement documentées dans la production des voyelles et objet de cette thèse, concerne la variabilité induite par le locuteur (par ex., Johnson & Mullennix, 1997; Nearey, 1989). La variabilité intra-locuteur se rapporte à la façon dont le locuteur s'adapte à la situation, par exemple en ajustant l'intensité (chuchotée ou criée), le débit (rapide ou lent) et le style de voix (grinçante, aspirée, etc.). Elle concerne aussi divers facteurs personnels (neutre ou émotif), interpersonnels (parole dirigée vers l'enfant ou l'adulte), sociaux (familier ou formel, spontané ou lu), et autres contraintes semblables. Les études proposées dans cette thèse portent spécifiquement sur la variabilité induite par certaines caractéristiques intrinsèques au locuteur, tel que l'âge et le sexe, et plus précisément la taille et la configuration de l'appareil vocal. Un autre facteur important qui contribue au problème de la variabilité et abordé dans cette thèse concerne la coarticulation, également désignée sous le nom de variations contextuelles (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). La variabilité contextuelle se rapporte au fait que dans le

discours continu, les gestes de la parole ne sont pas produits séquentiellement, mais empiètent les uns sur les autres en fonction des gestes précédents et à venir.

Ces différentes sources de variabilité ont pour principales conséquences la dispersion intra-catégorielle et le chevauchement inter-catégoriel des catégories phonétiques : des patrons acoustiques similaires pour des sons distincts, et différents patrons pour le même son. L'étude classique de Peterson et Barney (1952) illustre l'impact de la variabilité interlocuteur sur la réalisation des voyelles. Le corpus de données contient 1520 exemplaires des 10 voyelles de l'anglais américain produites par 76 locuteurs (enfants, femmes et hommes). Chaque exemplaire est représenté dans l'espace planaire vocalique par les valeurs du premier et second formant extraites de la région stable des voyelles. Les résultats révèlent des nuages de points qui se recouvrent pour certaines voyelles, mais aussi des exemplaires de certaines voyelles approchant le centre d'autres catégories. Une étude similaire a été reconduite récemment, incluant cette fois un système de 12 voyelles produits par 139 locuteurs anglo-américains (Hillenbrand, Getty, Clark, & Wheeler, 1995). Les résultats révèlent une détérioration de la séparation des voyelles à partir des patrons statiques, mais une meilleure précision à partir de l'information dynamique, c.-à-d. de changements spectraux et de durée.

Durant la même période, d'autres chercheurs se sont intéressés à l'impact de la variabilité contextuelle sur le signal de la parole (Liberman, Delattre, Cooper, & Gerstman, 1954). Les chercheurs, à la recherche d'indices acoustiques spécifiant les consonnes plosives, ont découvert que l'adulte traite différents patrons acoustiques comme la même consonne. Les consonnes anglaises [b, d, g] appartiennent à la classe des plosives voisées, telles que spécifiées par des valeurs de VOT similaires (Liberman, Delattre, & Cooper, 1955). Le lieu d'articulation distingue ces plosives, et au niveau acoustique la transition de F2 qui en résulte (Liberman et al., 1954). Les valeurs continues de la pente de la transition F2 permettent d'identifier des frontières catégorielles séparant les trois sons produits dans le même contexte vocalique. Cependant, la transition F2 peut varier pour la même consonne en fonction de la voyelle suivante. Par exemple, la transition F2 pour [d] est positive avant [i] et négative avant [u]. Ces exemples de variabilité contextuelle et interlocuteur suggèrent une relation complexe de plusieurs à plusieurs (*many-to-many*) entre les signaux de la parole et les catégories phonétiques. Néanmoins, les adultes ne manifestent aucun problème à identifier les sons de la

parole (Creelman, 1957; Hillenbrand et al., 1995; Peterson & Barney, 1952; Verbrugge, Strange, Shankweiler, & Edman, 1976), soulevant la question de l'invariance perceptive.

Les premières tentatives visant à explorer la constance perceptive chez l'enfant en bas âge datent du milieu des années 70. L'une de ces études proposait de vérifier à l'aide de la procédure CHP si des enfants âgés de 14 à 18 semaines peuvent percevoir la similarité entre la même consonne produite dans différents contextes vocaliques (Fodor, Garrett, & Brill, 1975). Les enfants ont été exposés à la présentation aléatoire des syllabes [pi], [ka] et [pu]. Un stimulus visuel accompagnait les syllabes [pi] et [pu] dans la condition *similaire* et [pi] et [ka] dans la condition *différent*. Les résultats indiquent une plus grande réponse du premier groupe, ce qui reflète selon les auteurs de cette recherche la capacité des enfants à percevoir une même consonne malgré la variabilité induite par différents contextes vocaliques.

Au cours de la même période, Kuhl et Miller (1975) ont utilisé la procédure HAS afin d'examiner chez des enfants de 4 à 16 semaines la capacité de distinguer les contrastes vocaliques malgré une variabilité acoustique non systématique. Les chercheurs ont d'abord vérifié la capacité des enfants à distinguer les deux voyelles [a] et [i] ainsi que deux patrons distincts de fréquence fondamentale (F0) (stable versus montant-descendant). Comme prédit et en concordance avec d'autres observations (Trehub, 1973), les enfants pouvaient distinguer les deux voyelles, mais également les deux patrons de F0. Dans deux conditions supplémentaires, les chercheurs ont ensuite examiné si les enfants pouvaient toujours détecter un changement de dimension acoustique cible (identité vocalique ou patron de F0) malgré la variabilité non pertinente induite par des fluctuations de l'autre dimension. Dans la première condition, les enfants étaient habitués à une voyelle alors que le patron de F0 variait de façon aléatoire durant les périodes pré- et post- changement. Dans l'autre condition, le patron de F0 demeuraient constant alors que l'identité de la voyelle variait. Les enfants ont distingué la voyelle malgré la variabilité de F0, mais ne pouvait plus distinguer les patrons de F0 lorsque l'identité vocalique variait, possiblement en raison de la saillance des voyelles dans certains contextes (Carrell & Smith, 1979). Néanmoins, ces résultats suggèrent une capacité précoce à faire face à un certain degré de variabilité contextuelle impliquant diverses dimensions acoustiques.

Dans une étude similaire à celle de Fodor et al. (1975) mentionné précédemment, Jusczyk et Derrah (1987) ont vérifié si les enfants de 2 mois considèrent la même consonne produite dans différents contextes vocaliques comme une seule catégorie. À l'aide de la procédure HAS, les chercheurs ont habitué des enfants à des listes de syllabes CV partageant la même consonne plosive (par exemple, [bi, bo, bə, ba]). Suite à l'habituation, les enfants du groupe contrôle ont continué d'entendre la même liste, alors que deux groupes expérimentaux recevaient différents stimuli. Pour les deux groupes, une nouvelle syllabe était ajoutée à la liste, contenant respectivement la même consonne mais une voyelle différente (par exemple, [bu]) (groupe *similaire-différent*) et deux segments différents (par exemple, [du]) (groupe *différent-différent*). Si les enfants traitent [b] produit dans divers contextes comme une seule catégorie, le groupe similaire-différent devrait moins récupérer de l'habituation que le groupe différent-différent. Les résultats montrent que les deux groupes expérimentaux ont récupéré de façon significative relativement au groupe contrôle, mais aucune différence n'a été observée entre ces deux groupes. Ces résultats ont été répliqués par la suite (Bertoncini, Bijeljac-Babic, Jusczyk, Kennedy, & Mehler, 1988) et interprétés comme une évidence du rôle de la syllabe en tant qu'unité de traitement de la parole chez l'enfant en bas âge, tel que décrit précédemment (1.1.3).

Dans une série d'études, Kuhl (1976a; 1977; 1979) a examiné le problème de la variabilité à l'aide de la procédure CHP impliquant une tâche de transfert d'apprentissage chez les enfants de 5.5 à 6.5 ans. Les enfants, d'abord entraînés à discriminer un contraste vocalique, étaient ensuite progressivement exposés à de nouveaux exemplaires contenant un degré croissant de variabilité. L'entraînement impliquait un exemplaire de deux voyelles synthétiques, dont la fréquence fondamentale et les fréquences formantiques correspondaient à ceux d'un locuteur masculin adulte et dont le contour de F0 était maintenu constant. Les enfants étaient ensuite testés avec des exemplaires synthétiques dont la voix correspondait à celle de femmes et d'enfants et dont le contour de F0 variait (par exemple montant et montant-descendant). Les résultats montrent que les enfants pouvaient aussi bien distinguer le contraste [a] - [i] (Kuhl, 1976a, 1979) que le contraste plus difficile [a] - [ɔ] (Kuhl, 1977) malgré les variations de contours de F0 et de fréquences fondamentales et formantiques.

Cette capacité de généraliser la distinction à de nouveaux exemplaires suggère une constance perceptive pour les voyelles dès l'âge de six mois.

Dans une autre étude, Marean, Werner, et Kuhl (1992) ont examiné cette capacité chez des enfants de 2 mois à l'aide d'une variante de la procédure CHP. Les enfants étaient entraînés à tourner la tête lorsqu'ils percevaient un changement d'identité vocalique ([a] et [i]) et à ne pas répondre lorsqu'il s'agissait de la même catégorie. Les résultats montrent que les enfants ont réussi à catégoriser les stimuli malgré les variations spectrales liées aux changements de locuteur et de fréquence fondamentale.

La capacité d'enfants de 2 mois à percevoir des contrastes consonantiques malgré la variabilité induite par le locuteur a été étudiée dans une série d'expériences utilisant la procédure HAS (Jusczyk, Pisoni, & Mullennix, 1992). Les enfants du groupe *locuteur-simple* étaient exposés à de multiples exemplaires du mot [bʌg] ou [dʌg] produits par un seul locuteur, alors que les enfants du groupe *locuteurs-multiples* étaient exposés à de multiples exemplaires produits par plusieurs locuteurs (6 hommes et 6 femmes). Dans chaque groupe, la condition *changement phonétique* présentait des exemplaires de l'autre syllabe en période post-habituation, alors que la condition contrôle impliquait des exemplaires de la même syllabe. Les résultats montrent que les groupes locuteur-simple et locuteurs-multiples de la condition changement-phonétique ont tous deux récupéré de façon significative de l'habituation, suggérant que les enfants peuvent normaliser la variabilité intra- et interlocuteur. Cependant, lorsqu'un délai de deux minutes était introduit entre les périodes de pré- et post-habituation, seuls les enfants entraînés avec un seul exemplaire produit par un seul locuteur ont répondu au changement phonétique. Aucune différence significative n'a été observée entre les groupes contrôle, locuteur-simple et locuteurs-multiples de la condition changement-phonétique lorsque les enfants étaient entraînés avec de multiples exemplaires. Selon les auteurs de cette recherche, ces résultats indiquent que la variabilité intra- et interlocuteur entrave l'encodage phonétique en mémoire à long terme, suggérant que la capacité de normaliser la variabilité induite par le locuteur ne serait pas complètement acquise à l'âge de deux mois.

Les nouveau-nés et les enfants en bas âge semblent donc posséder une capacité au moins rudimentaire de faire face à diverses sources de variabilité. Cependant, le degré de variabilité des stimuli utilisés dans ces études ne se compare pas à celui que les enfants rencontrent au quotidien. En outre, certaines procédures utilisées, notamment le transfert d'apprentissage, implique une augmentation progressive de la variabilité à travers les essais, pouvant informer l'enfant de manière explicite quant à l'information à laquelle il doit porter attention. Cette procédure correspond ainsi à un type d'apprentissage supervisé qui est peu compatible avec l'absence de rétroaction durant l'acquisition du langage. Bien que les preuves directes de constance perceptive chez l'enfant en bas âge demeurent équivoques, d'autres types d'évidences, indirectes cette fois, peuvent jeter une certaine lumière sur la question.

Les enfants acquièrent rapidement les règles phonotactiques de leur langue maternelle, c.-à-d. les règles de combinaison et de positionnement des sons de la parole dans une langue. Jusczyk, Luce et Charles-Luce (1994) ont montré que les enfants de 9 mois portent une plus grande attention à des non-mots contenant des séquences de sons fréquentes que non fréquentes dans leur langue maternelle. D'autres études révèlent qu'entre l'âge de 6 et 9 mois, les enfants peuvent distinguer entre des séquences de sons légales et illégales dans leur langue maternelle (Friederici & Wessels, 1993; Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993). Pour savoir qu'une séquence de sons est fréquente dans sa langue, et surtout qu'un son particulier peut en précéder un autre mais non le suivre, l'enfant doit reconnaître l'identité phonétique des sons, et donc pouvoir normaliser la variabilité contextuelle.

La capacité de l'enfant à faire face à la variabilité durant la première année de vie peut également être inférée à partir de la réorganisation perceptive. D'abord, l'adaptation graduelle du système perceptif à la langue ambiante indique que l'enfant peut faire face à la variabilité, en supposant que la réorganisation perceptive résulte de centaines de répétitions de sons de la parole produits par différents locuteurs dans divers contextes. En accord avec cet argument, une étude récente a démontré le rôle de l'analyse distributionnelle du signal de la parole dans la réorganisation perceptive (Anderson, Morgan, & White, 2003). Les chercheurs ont examiné si la fréquence relative des sons de la parole dans une langue en affecte l'ordre d'acquisition. La prédiction voulait qu'une diminution de sensibilité survienne d'abord pour les contrastes à haute fréquence. À l'aide de la procédure CHP, les chercheurs ont examiné la

capacité d'enfants de 6.5 et 8.5 mois apprenant l'anglais à distinguer des contrastes consonantiques non-natifs, plus précisément entre le contraste Hindi de plosives coronales rétroflexe versus dentale ([ʈɜ] versus [ʈ̪ɜ]) et le contraste Salish de plosives dorsales vélaire versus uvulaire ([k̓ɜ] versus [q̓ɜ]). Tel que prédit, les enfants de 6 mois ont distingué les deux contrastes, mais ceux de 8 mois étaient moins sensibles au contraste coronal que dorsal, ce dernier étant moins fréquent en anglais. Enfin, le fait que les enfants traitent les contrastes phonémiques différemment des contrastes non-phonémiques dans les études portant sur la réorganisation perceptive (par ex., Werker & Tees, 1984) suggère qu'ils perçoivent la similitude et la différences entre des stimuli expérimentaux avec lesquels il ne possèdent aucune expérience.

Un autre type d'évidences provient d'études comparatives entre l'humain et l'animal. L'une de ces études s'est penchée sur la capacité de la caille japonaise à normaliser la variabilité contextuelle (Kluender, Diehl, & Killeen, 1987). L'étude visait précisément à entraîner les oiseaux avec les consonnes [b], [d], [g] suivies de 12 voyelles, pour ensuite vérifier leur capacité à reconnaître ces consonnes présentés dans huit nouveaux contextes vocaliques. Une étude plus récente a observé l'impact de consonnes précédentes sur la normalisation de consonnes chez l'étourneau (Lotto, Kluender, & Holt, 1997). Lors d'une épreuve de reconnaissance de syllabes CV variant selon le lieu d'articulation ([ba] - [ga]) et précédées de [al] ou [ar], les chercheurs ont observé chez l'oiseau le même effet de compensation perceptive observé chez l'adulte. Dans l'ensemble, ces études suggèrent que les enfants apprennent à traiter la variabilité contextuelle durant la première année de vie à partir d'un mécanisme général à la base de la constance perceptive.

Les données rapportées dans cette première section suggèrent que les habiletés de perception initiales et leur développement durant la première année de vie font appel à des capacités auditives et des mécanismes d'apprentissage généraux. Aucune évidence directe n'indique toutefois comment les enfants structurent le signal de la parole durant l'acquisition phonétique. La manifestation d'une réorganisation perceptive dès l'âge de 4 mois remet en doute un mécanisme complexe basé sur l'extraction de caractéristiques phonétiques, suggérant la possibilité d'accomplir la tâche à partir des patrons continus du signal acoustique. Enfin, bien que l'invariance perceptive pour les sons de la parole semble une

propriété perceptive générale et une capacité précoce chez l'enfant, *comment* ces derniers traite la variabilité du signal reste à découvrir. À la recherche d'une solution, la prochaine section présente les modèles de l'invariance chez l'adulte, mais d'abord les modèles de la perception de la parole chez l'enfant, décrivant la façon dont chacun aborde le mécanisme d'apprentissage, l'unité de traitement, et la façon de résoudre le problème de la variabilité.

1.2 La perception de la parole en théorie

1.2.1 Modèles du développement de la perception de la parole

WRAPSA

L'un des premiers modèles de la perception de la parole chez l'enfant se nomme WRAPSA (*Word Recognition and Phonetic Structure Acquisition*) (Jusczyk, 1986, 1993). Ce modèle propose d'expliquer le processus de reconnaissance de mots, mais également l'évolution des capacités de perception de la parole chez l'enfant. La version adulte de WRAPSA identifie les mots à partir du signal acoustique, ceci en trois étapes. D'abord, le signal sortant du système auditif périphérique accède aux *routines de traitement analytique* pour fournir une représentation spectrographique du signal. Des détecteurs de caractéristiques acoustiques sensibles à certaines régions spécifiques de fréquences et propriétés spectro-temporelles (transitions formantiques par exemple) effectuent cette opération. L'information sortante passe ensuite par le *schème de pondération*, qui élimine les détails non pertinents et accentue les caractéristiques (fermeture de plosive ou glissement de transition par exemple) sur la base desquelles des énoncés peuvent être comparées et des décisions catégorielles peuvent être prises. À l'étape suivante d'*extraction de patrons*, l'entrée est segmentée en unités de la taille de mots. À ce moment, les représentations provisoires des éléments lexicaux ne sont pas décomposées en segments phonétiques. Les caractéristiques acoustiques saillantes sortant du schème de pondération sont plutôt intégrés en patrons syllabiques. Enfin, la sortie de l'extracteur de patrons est acheminée au lexique, pour constituer une nouvelle entrée lexicale ou s'apparier à un item lexical existant. La version initiale de WRAPSA (Jusczyk, 1986) supposait que le processus d'appariement s'effectue à partir d'items lexicaux prototypes mémorisés en termes de propriétés acoustiques. Des recherches subséquentes sur la mémoire ont mené l'auteur du modèle à favoriser un lexique basé sur des exemplaires, dans lequel

l'activation d'un item lexical s'effectue par le calcul d'une moyenne locale des traces multiples de cet item produit dans divers contextes.

En tant que modèle développemental, WRAPSA n'effectue pas toutes ces opérations au départ. Les routines de traitement et les détecteurs auditifs sont présumés innés, ces composantes étant décrites en termes de propriétés auditives générales plutôt que de processeurs phonétiques spécialisés. Le schème de pondération, également présent dès la naissance, demeure non spécifié en ce qui concerne la structure de la langue maternelle, jusqu'à l'accumulation suffisante d'expérience avec cette langue. Les caractéristiques acoustiques importantes qui permettent de distinguer les mots sont graduellement identifiées par l'extraction des régularités de la langue ambiante et la réduction de la sensibilité aux détails sans importance et non systématiques. Tout comme la discrimination, la normalisation de la variabilité constitue dans ce modèle une capacité élémentaire, qui repose sur l'existence de standards acoustiques/phonétiques.

SARAH

Le modèle SARAH (*Syllable Acquisition, Representation, and Access Hypothesis*) (Mehler, Dupoux, & Segui, 1990) a également été proposé pour rendre compte des capacités perceptives initiales de l'enfant et de la transition vers le traitement lexical adulte. Dans la version adulte de SARAH, la perception de la parole implique trois niveaux de traitement. Le signal de la parole est d'abord segmenté au *niveau syllabique*, selon les cadres syllabiques de la langue en question, et analysé en termes de séquences de segments phonétiques prototypiques. La sortie des analyseurs syllabiques, particulièrement en début de mot, est utilisée au *niveau lexical* afin d'activer une cohorte d'items lexicaux candidats. Enfin, le *niveau phonologique* correspond à une banque de phonèmes n'entretenant aucun rôle direct avec l'accès lexical et la reconnaissance du mot, mais contraignant plutôt la réalisation de surface des sons de la parole produits dans différents contextes, ceci à partir d'un ensemble de règles transformationnelles.

SARAH postule un certain nombre de composantes innés en termes de contenu représentatif et de processus. La syllabe constitue l'unité de traitement de base dès le début de l'acquisition du langage afin de construire le lexique mental et assure l'accès lexical au stade

adulte. Le filtre syllabique segmente le signal de la parole en séquences légales (par exemple, CV, CVC, V), alors que l'analyseur phonétique calcule la représentation phonétique universelle sous-jacente en termes de caractéristiques telles que définies par les théories traditionnelles des caractéristiques distinctives (par exemple, Chomsky et Halle, 1968). Le modèle vient également pourvu d'un détecteur de frontière de mots qui utilise la représentation syllabique et l'information acoustique (indices prosodiques et phonotactiques) pour associer le signal aux items lexicaux possibles. L'exposition à la langue maternelle permet à l'analyseur phonétique d'éliminer les caractéristiques non pertinentes aux catégories natives des sons de la parole, au filtre syllabique de n'accepter éventuellement que les séquences permissibles, et au détecteur de frontière de mots de maintenir les stratégies de segmentation utiles. Enfin, les unités de la parole sont stockées dans un format abstrait indépendant du locuteur et de la durée. Tout comme WRAPSA, SARAH postule l'invariance perceptive sans en décrire le mécanisme.

PRIMIR

Un autre modèle plus récent tente d'expliquer la façon dont le traitement de la parole se rapporte à l'acquisition du langage : PRIMIR (*Processing Rich Information from Multidimensional Interactive Representations*) (Werker & Curtin, 2005). Au stade adulte, le modèle comporte plusieurs niveaux de représentations qui interagissent afin de regrouper diverses sources d'informations. Le niveau de *perception générale* contient l'information détaillée du signal. Les catégories phonétiques y sont définies en termes de segments sensibles au contexte. Les propriétés acoustiques et phonétiques (c.-à-d. articulatoires) se regroupent pour définir l'inventaire segmental d'une langue, classification qui se base sur la similarité des signaux d'entrée. Les catégories phonétiques sont transmises au niveau de la *forme du mot*, où les items lexicaux sont extraits et groupés en classes selon leur ressemblance phonétique. L'étape suivante implique le niveau des *mots significatifs*, qui relie les mots à leur signification.

Le principal postulat développemental de PRIMIR veut que les enfants en bas âge soient outillés de mécanismes d'apprentissage généraux leur permettant de calculer les indices de probabilités transitoires et distributionnelles du signal de la parole. PRIMIR postule également des extracteurs de caractéristiques acoustiques/phonétiques innés sous la forme de

biais auditifs. Les catégories phonétiques spécifiques à une langue émergent en fonction de l'expérience et de la capacité de l'enfant à intégrer l'information provenant de différentes sources. Les phonèmes sont acquis plus tard sur la base des catégories perceptives générales, des classes de mots, d'un réseau suffisamment dense au niveau des mots significatifs, et finalement par l'exposition au langage écrit. Enfin, PRIMIR supposent comme les autres modèles l'existence d'invariants acoustiques/phonétiques.

Comme le démontre la description de ces modèles, la façon dont l'enfant catégorise les sons de la parole produits par différents locuteurs et dans différents contextes phonétiques demeure jusqu'à maintenant sans réponse. Ces modèles, bien qu'inspirés des théories de l'invariance chez l'adulte, stipulent un mécanisme inné de normalisation basé sur les standards acoustiques et phonétiques sans en spécifier les opérations impliquées. La prochaine section propose une brève incursion en territoire adulte, ce qui pourrait permettre de mieux comprendre le mécanisme à la base de l'invariance perceptive chez l'enfant.

1.2.2 Théories de la perception de la parole chez l'adulte

Invariance acoustique/auditive

Un premier type d'approche à l'invariance veut que des propriétés standards correspondant aux événements phonétiques se trouvent dans le signal de surface de la parole. Ceci suppose que le système auditif est sensible à ces propriétés et les utilise afin d'identifier les sons de la parole. Cette approche vise donc en premier lieu à identifier les indices acoustiques potentiels, puis à examiner leur impact sur la perception de la parole chez l'auditeur.

L'une des premières tentatives visant à tester l'approche acoustique porte sur les propriétés de signal indiquant le lieu d'articulation des consonnes plosives en anglais. Dans une série d'études, Blumstein et Stevens (1979; Stevens & Blumstein, 1978) ont observé que le spectre à court terme du dégagement articulatoire dans des syllabes CV et VC peut être utilisé pour classifier les plosives bilabiales, alvéolaires et vélaires, et ce, malgré divers contextes vocaliques et conditions de voisement. D'autres façons d'observer le signal acoustique ont également été proposées. Au lieu de l'approche statique tout juste décrite, Kewley-Port (1989) a proposé d'examiner les caractéristiques variables dans le temps liés au premières 40msec de syllabes CV afin d'indiquer le lieu d'articulation de consonnes voisées.

Un corpus de données produit par trois locuteurs dans huit contextes vocaliques a été utilisé. Les résultats montrent un accord inter-juge élevé pour l'identification des consonnes à partir de l'inspection visuelle des spectres continus. Sur le plan de la perception auditive, Blumstein et Stevens (1981) ont montré que les auditeurs adultes peuvent identifier le lieu d'articulation des consonnes voisées et non voisées dans divers contextes vocaliques à partir du début de syllabes CV. Ceci suggère que l'information spectrale en début de syllabe permet à l'auditeur d'identifier le lieu d'articulation de consonnes plosives produites par de multiples locuteurs dans divers contextes.

La théorie quantique de la parole (Stevens, 1972) propose que la relation entre la production de la parole et le signal acoustique est non linéaire, et que cette non-linéarité peut être exploitée pour percevoir les événements phonétiques. Selon ce point de vue, le changement continu d'un paramètre articulatoire résulte en un signal stable jusqu'à l'atteinte d'une certaine valeur, après quoi une perturbation maximale du signal se produit. L'appareil vocal produirait ainsi des zones de stabilité acoustique, correspondant aux caractéristiques phonétiques, ainsi que des zones d'instabilité entre deux configurations articulatoires 'quantiques', correspondant aux frontières phonétiques. Par exemple, la production d'une fricative en variant le lieu de constriction le long du tractus vocal produit éventuellement un changement brusque des basses fréquences formantiques, qui correspond à un changement de catégorie phonétique (Stevens, 1989).

La nature quantique de la parole pourrait également s'observer au niveau auditif. Un exemple classique se rapporte à l'effet du 'centre de gravité' lors de la perception vocalique (Chistovich & Lublinskaya, 1979). Tel que décrit dans Fant (1986), un positionnement avant de la langue résulte en des valeurs éloignées de F1 et F2, alors qu'un positionnement arrière résulte en formants étroitement espacés. Perceptivement, une différence plus petite ou plus grande que 3.5 barks entre F1 et F2 détermine respectivement si deux formants sont traités en tant que patron unique (en termes de valeur moyenne des formants) ou en tant que deux formants distincts. La position de la langue et la différence de valeurs formantiques qui en résultent peuvent varier jusqu'à un certain degré sans affecter la qualité de la voyelle, mais après une certaine valeur (3.5 barks), une autre voyelle est perçue. La constance perceptive

relèverait ainsi à la fois d'un processus passif, durant lequel l'auditeur capte les patrons acoustiques standards, et d'un processus actif, qui requière une transformation du signal.

Le système d'équations de locus (Sussman, McCaffrey, & Matthews, 1991) propose également une approche active à l'invariance, en supposant que la pente de F2 entre le début et le point médian de la voyelle révèle de façon invariante le lieu d'articulation de consonnes plosives en anglais. En ce qui concerne les voyelles, Miller (1989) a décrit le système vocalique de l'anglais américain (en se basant sur les données de Peterson & Barney, 1952) comme un ensemble de points dans un espace tridimensionnel où chaque axe représente la valeur logarithmique d'un rapport formantique, log (F2/F1) par exemple. Dans le même esprit, un modèle de production de la parole simulant la voix d'enfants et de locuteurs adultes masculins et féminins s'est montré adéquat pour séparer les 10 voyelles du français dans un espace tridimensionnel utilisant différents rapports de formants, et ces paramètres se sont ensuite montrés efficaces pour l'identification des voyelles chez l'adulte (Ménard, Schwartz, Boë, Kandel, & Vallée, 2002). Au contraire d'une invariance acoustique 'pure', ces approches impliquent une analyse perceptive et certaines transformations du signal afin de réduire ou éliminer la variabilité de la parole.

Invariance motrice

Les théories motrices de la perception de la parole proposent que les auditeurs perçoivent la parole en recrutant leurs connaissances sur la façon dont celle-ci est produite. Le signal brouillé de la parole permettrait à l'auditeur de récupérer les gestes articulatoires du locuteur, unités de base de la perception de la parole, qui elles seraient invariantes. Le cas de la plosive [d] vient à l'appui des théories motrices. Lorsque produits dans différents contextes vocaliques, les patrons acoustiques de la consonne révèlent des pentes variables de transition F2 qui recouvrent d'autres plosives (Delattre, Liberman, & Cooper, 1955). L'absence d'invariance acoustique n'empêche pas l'auditeur de percevoir le même [d], suggérant que celui-ci se base sur le même degré de constriction et de relâchement du bout de la langue (la lame) sur l'arête alvéolaire observé durant la production de [d] peu importe le contexte. Il s'agit du type de connaissance qui permettrait à l'auditeur d'établir la correspondance entre les sons perçus et produits. En d'autres termes, le cerveau, à l'écoute d'un énoncé, activerait le substrat neuronal nécessaire à la production de cet énoncé.

La théorie-moteur de la perception de la parole (Liberman et al., 1967) postulait au départ que l'objet de la perception se rapportait aux commandes motrices envoyées aux muscles articulatoires. Dans sa plus récente version (Liberman & Mattingly, 1985), les unités de la perception de la parole se rapportent aux structures de contrôle des gestes articulatoires. L'invariant se situerait donc au niveau abstrait de l'intention du locuteur et des cibles qu'il se fixe plutôt qu'au niveau concret de l'articulation de surface.

En contraste avec la théorie-moteur, la théorie du Réalisme direct (*Direct Realism*) (Fowler, 1986; Fowler & Rosenblum, 1986) postule que l'auditeur perçoit directement les énoncés du locuteur par la poursuite auditive des gestes articulatoires à partir du signal de la parole. Cette théorie s'inspire de l'approche écologique de la perception visuelle (Gibson, 1979), selon laquelle les objets et les événements sont perçus en captant l'information invariante qui les représente et qui est contenue dans l'environnement. À l'opposé de la théorie-moteur, où l'invariance fait appel à des algorithmes complexes de type analyse-par-synthèse (Halle & Stevens, 1962), la perspective écologique ne fait pas intervenir de processus cognitifs de haut niveau. L'observateur perçoit directement l'objet distal par le stimulus proximal. Dans le cas de la vision, lorsqu'un faisceau de lumière est réfléchi sur un objet, il acquiert la structure de cet objet, atteint ensuite la rétine de l'observateur, et impartie sa structure au système perceptif. Quant à la parole, les événements distaux qui structurent le signal acoustique se rapportent aux gestes phonétiques. Lorsque le signal acoustique atteint le système auditif, ce dernier capte directement les manœuvres articulatoires associées aux gestes du locuteur, sans processus de reconstruction. La perspective écologique souligne l'importance du mouvement de l'observateur dans l'invariance visuelle. Dans le cas de la parole, l'auditeur n'a pas à se déplacer. Il pourrait plutôt devoir porter attention au mouvement du signal reflétant le geste articulatoire, qui lui fournirait les invariants de la parole.

1.2.3 Résumé et critique

Les trois modèles de la perception de la parole chez l'enfant affichent certaines différences quant aux mécanismes d'apprentissage et de traitement qu'ils proposent. Avec SARAH, l'acquisition des catégories phonétiques rappelle un processus de maintenance (voir Figure 1.1). La structure sous-segmentale, entièrement spécifiée de façon universelle à la naissance,

conserve seulement les caractéristiques distinctives propres aux catégories phonétiques de la langue maternelle. À l'opposé, WRAPSA et PRIMIR relèvent d'un processus d'adaptation qui module une gamme de frontières auditives innées au son de la structure ambiante. Quant à l'unité de traitement de la parole, PRIMIR se fonde sur les propriétés acoustiques et articulatoires du signal, alors que WRAPSA et SARAH possèdent un cadre syllabique dont la composition segmentale n'émerge que plus tard. Malgré leurs divergences, les trois modèles affrontent similairement le problème de la variabilité en considérant innée la constance perceptive, sans en détailler le mécanisme.

Les théories de la perception de la parole chez l'adulte proposent pour leur part deux mécanismes afin de résoudre le problème de variabilité, localisant la constance perceptive au niveau acoustique ou articulatoire. L'approche motrice se distinguent de l'approche acoustique en ce que l'objet de la perception de la parole est l'événement distal (au niveau de la production) et non proximal (le signal acoustique). En ce qui a trait aux approches motrices, la théorie-moteur considère l'événement distal comme une structure privée et abstraite correspondant à l'intention du locuteur et que le stade d'exécution embrouille. Au contraire, l'objet de la perception selon la théorie du Réalisme direct se rapporte au geste manifeste, défini comme une action publique et observable et directement accessible à partir d'un signal acoustique structuré.

Malgré l'existence de patrons acoustiques standards reflétant certaines caractéristiques phonétiques dans certaines langues, les invariants découverts jusqu'à maintenant ne s'observent pas dans toutes les langues (Keating, 1985; Ladefoged, 1980; Port, 1981). D'autre part, bien que l'approche motrice vise à résoudre le problème de la variabilité, elle demeure vague quant à la façon dont le signal transportent les invariants, ainsi que sur la façon dont les gestes peuvent être récupérés par le système perceptif. Finalement, et plus important encore, les approches acoustiques et motrices s'intéressent avant tout à la perception de la parole chez l'adulte, négligeant la question de l'invariance dans la perception précoce de la parole.

Une dernière avancée des approches motrices proposent d'aborder l'invariance d'un point de vue développemental. Situé quelque part entre le réalisme direct et la théorie-moteur,

la Phonologie Articulatoire (*Articulatory Phonology*) (Browman & Goldstein, 1986, 1992, 1995) conçoit la production de la parole comme un système moteur dynamique. L'unité de base de la parole demeure le geste articulatoire, lui-même défini en tant que système dynamique spécifique dont les paramètres régissent les actions de l'appareil vocal. L'état d'équilibre constitue le paramètre principal du système et correspond à la cible visée par le geste. Une catégorie phonétique ne se définit plus comme un objet holistique ou un groupement de caractéristiques phonétiques statiques, mais comme une constellation de gestes effectués de façon séquentielle et qui se chevauchent parfois. Dans ce système, les invariants correspondent aux valeurs des paramètres de commande qui spécifient les gestes. De ce point du vue, la perception de la parole représente également un système dynamique, un système qui est sensible aux informations acoustiques reflétant les gestes articulatoires, qui à leur tour indiquent la cible invariante du locuteur. En lien avec cette perspective, la prochaine section traite de l'acquisition des catégories phonétiques et de la question de l'invariance à partir d'un cas particulier de sons paroliers, les tons lexicaux.

1.3 L'acquisition des tons lexicaux

1.3.1 Phonologie des tons et système de tons mandarins

Les tons lexicaux sont employés dans les langues tonales afin de distinguer la signification des mots, partageant ainsi la fonction linguistique des voyelles et des consonnes. Les langues tonales, parlées par plus de la moitié de la population mondiale (Fromkin, 1978), diffèrent entre elles en termes du nombre et du type des contraste tonaux, des leurs positions possibles dans un mot, et de leur unité porteuse, c.-à-d. l'élément auquel un ton s'associe dans la structure phonologique (Gussenhoven, 2004).

Le chinois mandarin possède quatre tons, deux tons ponctuels (Haut et Bas) et deux tons modulés (Montant et Descendant). Les tons ponctuels, ou de niveau, sont habituellement conçus en termes de patrons statiques, en ce qu'ils contiennent peu de mouvement dans leur réalisation de surface, au contraire des tons modulés, ou dynamiques (Abramson, 1962). Le mandarin possède également un ton neutre, non spécifié phonologiquement (Duanmu, 2000) et pouvant prendre place sur les syllabes non-initiales d'un mot. L'unité porteuse de tons en mandarin est généralement considérée comme étant la partie voisée de la syllabe (la voyelle

ou la rime) (par ex., Howie, 1974; Lin, 1995), bien que pour certains, le domaine des tons s'étend sur la syllabe entière (Xu, 1998).

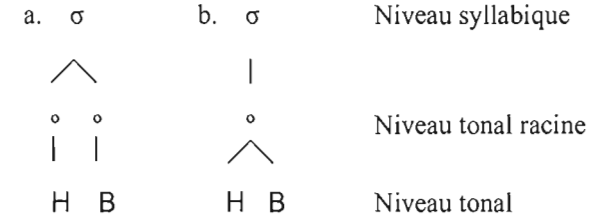
La représentation phonologique sous-jacente des tons implique diverses caractéristiques distinctives telles que ±Haut et ±Contour (Wang, 1967). Les tons ponctuels en mandarin se distinguent entre eux par la caractéristique /±Haut/, alors que les tons modulés sont spécifiées par des séquences de cibles, le ton Montant avec /-H +H/et le ton Descendant /+H -H/. Cette représentation statique est représentée en (1), où les points correspondent aux cibles tonales /+H/ et /-H/, qui se manifestent respectivement par [H] et [B] (pour 'bas') (adapté de Gussenhoven, 2004) :

(1)

a. ma	b. ma	c. ma	d. ma
H	BH	B	HB
'mère'	'chanvre'	'cheval'	'gronder'

Afin de rendre compte du comportement différent des tons modulés dans différentes langues, Yip (1989) distingue les 'groupements de tons' des 'contours de tons' en ajoutant un niveau racine entre le niveau tonal et le niveau syllabique. Le groupe de tons dans (2a) peut se décomposer au niveau racine en deux éléments distincts, /+H -H/ par exemple, alors que le contour de tons dans (2b) ne contient qu'un élément (±Contour par exemple), de la même façon que les éléments de diphtongues peuvent être séparés dans certaines langues (par exemple, hawaïen, comparable à 2a) et non dans d'autres (par exemple, néerlandais, comme dans 2b) (Gussenhoven, 2004).

(2)



Selon un autre point de vue, les tons modulés impliquent des cibles dynamiques plutôt qu'une séquence d'éléments statiques (Xu, 1997). Alors que la production de tons de niveau requière l'atteinte d'un état stable, la production des tons modulés requière l'atteinte d'une vitesse (positive pour Montant et négative pour Descendant). Les tons mandarins de niveau impliquent ainsi les cibles /+H/et /-H/, alors que les tons dynamiques sont spécifiés par une combinaison de cibles statiques et dynamiques (/-H+S/ pour Montant et /+H-S/ pour Descendant, où H = hauteur et S = pente).

1.3.2 Phonétique des tons

Le principal corrélat acoustique des tons est la fréquence fondamentale du signal de la parole (F0), qui correspond au taux de vibration des cordes vocales durant la phonation en général et la production tonale en particulier (Abramson, 1962; Chao, 1933; Howie, 1976). À l'opposé d'autres catégories phonétiques, le cas particulier des tons lexicaux permet de simplifier l'étude du développement de la perception de la parole à une seule dimension acoustique.

Bien que F0 soit habituellement considérée l'indice principal de la perception tonale chez l'adulte, d'autres caractéristiques acoustiques semblent jouer un rôle dans l'identification et la discrimination des tons. Certains chercheurs ont exploré des sources d'informations extrinsèques, ou comment les tons sont perçues relativement les uns aux autres (par ex., Leather, 1983; Xu, 1994). La majorité s'est toutefois penchée sur les propriétés intrinsèques des tons, telles que la qualité de la voix dans les langues tonales avec phonation lexicale (le green mong par exemple) (Andruski & Ratliff, 2000; Maddieson & Hess, 1986); la hauteur et le contour en thaï et mandarin (Abramson, 1978; Gandour, 1978; Massaro, Cohen, & Tseng, 1985); la hauteur et la direction en cantonais, mandarin, taïwanais, et thaï (Gandour, 1983); la direction et le contour en yorúbà (Hombert, 1976); la durée et l'amplitude en mandarin (Whalen & Xu, 1992); le moment du point de flexion et la pente de descente initiale pour les tons Montant et Bas en mandarin (Shen & Lin, 1991); la hauteur, la direction et le contour en cantonais (Gandour, 1979), et la hauteur moyenne, la direction, la pente et le point final en thaï et yorúbà (Gandour & Harshman, 1978).

Andruski et Costello (2004) ont proposé de modéliser la réalisation de surface des tons à partir d'équations linéaires de type $a + bx$, où le coefficient a correspond à l'ordonnée à

l'origine et *b* à la pente, qui représentent respectivement la hauteur initiale et la direction de F0. Une ordonnée à l'origine haute ou basse et une pente nulle devraient ainsi caractériser les tons mandarins Haut et Bas, alors qu'une ordonnée à l'origine basse ou haute combinée à une pente positive ou négative devraient définir les tons Montant et Descendant. Le panneau gauche de la figure 1.2a illustre les patrons de F0 des quatre tons mandarins produits en isolation par un locuteur masculin adulte natif du mandarin (données de Xu, 1997).

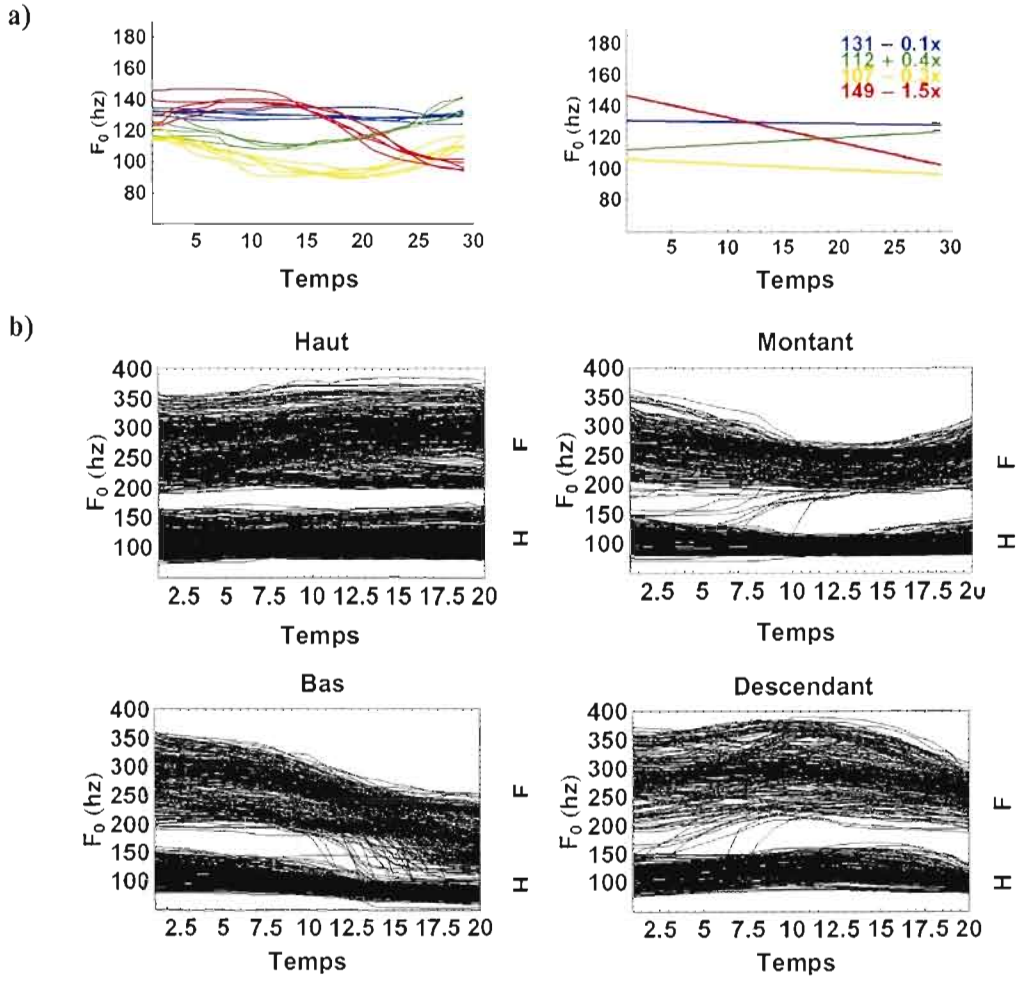


Figure 1.2 **a)** 20 exemplaires de tons Haut (bleu), Montant (vert), Bas (jaune) et Descendant (rouge) produits isolément par un locuteur masculin (panneau gauche) et leur approximation linéaire (panneau droit). **b)** Tons produits par un locuteur masculin et un locuteur féminin en parole continue avec contextes tonaux et focus variables. (Données de Xu, 1999).

Tel qu'illustré dans le panneau droit de la figure 1.2a, les polynômes de premier degré représentent adéquatement ces productions. Cependant, le modèle linéaire s'avère insatisfaisant pour les tons produits en contexte continu. La figure 1.2b contient 2800 exemplaires des quatre tons produits par un locuteur masculin et un locuteur féminin en parole continue avec contextes tonaux et focus prosodique variables (données de Xu, 1999). Un simple coup d'œil à ces graphiques suffit pour voir que les équations linéaires des patrons de F0 résultent en une pente nulle pour les tons Montant et Descendant et une pente positive et négative pour les tons Haut et Bas, le contraire de ce que prédit le modèle.

Les équations linéaires, plutôt que de caractériser la réalisation de surface des tons, semblent plutôt ici refléter les cibles tonales sous-jacentes, comme le propose le modèle d'approximation de cibles de la production tonale (*Target Approximation model of tonal production*) (Xu & Wang, 2001). Dans ce modèle, les coefficients linéaires *a* et *b* correspondent respectivement aux paramètres de commandes articulatoires statiques et dynamiques spécifiant les cibles sous-jacentes associées aux catégories tonales. L'efficacité avec laquelle les coefficients d'équations de degré plus élevé peuvent représenter les patrons de surfaces des tons mandarins sera abordée plus loin (1.3.4). La prochaine partie revient au sujet principal de cette thèse, l'acquisition phonétique.

1.3.3 Développement de la perception des tons

Peu d'études se sont penchées sur le développement de la perception des tons lexicaux chez l'enfant. La littérature comporte cependant plusieurs études relatives à la perception de la fréquence fondamentale de stimuli non verbaux, de la musique et de la parole. Par exemple, Trehub (1973) a examiné la sensibilité d'enfants âgés de 5 à 16 semaines pour des tonalités de niveau à l'aide de la procédure HAS. Les résultats montrent que les enfants n'ont pas discriminé les paires 1000-2000hz, 100-200hz et 200-1000hz de fréquences à ondes carrées (sans harmoniques) présentées à 26dB au-dessus du niveau de bruit ambiant. Dans sa conclusion, l'auteure explique l'absence de réponse en fonction de failles méthodologiques probables et indique des façons d'améliorer les procédures d'étude de la perception chez l'enfant. Des études subséquente utilisant d'autres méthodes plus sophistiquées ont montré que la discrimination des basses fréquences n'atteint pas pleine maturité avant l'âge de 8 à 10

ans (Olsho, Koch, & Halpin, 1987), mais aussi que les enfants de 7 à 9 mois perçoivent des différences de fréquence aussi minimes que 29 Hertz (Sinnott & Aslin, 1985).

La recherche sur la perception de F0 en musique révèle que les enfants de 5 à 11 mois peuvent distinguer des séquences tonales ou des mélodies dont le contour de F0 diffère (Trehub, Bull, & Thorpe, 1984). Thorpe (1986) a également étudié la capacité d'enfant de 7 à 10 mois à percevoir des contrastes de contours de F0. Les enfants ont été entraînés à répondre à un changement de séquences de tons montants (avec divers intervalles de 6-2, 4-2, 4-1 semitons) à des séquences de tons descendants. Les résultats indiquent que les enfants peuvent distinguer les changements directionnels, même si le contour n'implique qu'un semiton. Les enfants de 9 à 11 mois sont également en mesure de catégoriser des séquence de F0 malgré des variations perceptibles de clef ou de grandeur d'intervalle, suggérant qu'ils peuvent extraire la direction des contours de fréquence fondamentale (Trehub, Thorpe, & Morrongiello, 1987). Ces études suggèrent que les enfants en bas âge sont sensibles à l'information relative au mouvement de la fréquence fondamentale en musique.

En ce qui concerne la parole, les évidences suggèrent que les enfants sont sensibles à la structure d'intonation de leur langue maternelle (e.g., Kaplan, 1969; Morse, 1972). Jusczyk et collègues (1993) ont montré que les enfants de 6 mois apprenant l'anglais écoutent plus longuement des listes de mots anglais que norvégiens (suite au filtrage passe-bas des stimuli), les premières listes différant des secondes en ce qui a trait au contour de F0 des syllabes finales (Haugen & Joos, 1972) et la hauteur de F0 des syllabes accentuées (Peters, 1997). Les enfants qui apprennent l'anglais sont sensibles aux patrons d'accent prédominant des mots anglais entre 6 et 9 mois (Jusczyk, Cutler, & Redanz, 1993). Au cours de cette période, ils commencent également à utiliser l'accent pour segmenter les mots dans le discours continu (Jusczyk, Houston, & Newsome, 1999), comme le font les adultes (Cutler & Norris, 1988). Le rôle de la fréquence fondamentale dans l'acquisition du langage s'observe également dans la capacité d'enfants aussi jeune que 6 mois à utiliser leur sensibilité à la prosodie pour marquer des unités syntaxiques (Hirsh-Pasek et al., 1987; Nazzi, Nelson, Jusczyk, & Jusczyk, 2000; Soderstrom, Seidl, Nelson, & Jusczyk, 2003).

En lien avec les tons lexicaux, Harrison (2000) a étudié la perception tonale chez l'enfant de 6 mois apprenant le yorúbà, une langue africaine à trois tons de niveau : Haut, Moyen et Bas. Le but était de vérifier l'hypothèse selon laquelle les tons lexicaux et les contrastes de voisement possèdent le même élément laryngeal sous-jacent, tel que proposé par la théorie de la Phonologie du Gouvernement (par ex., voir Kaye, Lowenstamm, & Vergnaud, 1985). Comme l'adaptation perceptive aux contrastes natifs de VOT semblent se produire entre 6 et 8 mois (Eilers, Gavin, & Wilson, 1979), Harrison a prédit qu'au même âge, les enfants apprenant le yorúbà pourront identifier les contrastes tonaux de leur langue maternelle. Il a d'abord testé la discrimination des trois tons chez deux adultes parlant le yorúbà à l'aide de versions synthétiques de la syllabe [ki] produites en isolation par un locuteur masculin et desquelles l'information relative à la durée a été éliminée. Les résultats indiquent que la frontière perceptive des tons Haut versus non-hauts se situe entre 190 et 210 Hertz. Les adultes ne pouvaient pas distinguer les tons Moyen et Bas, en raison selon l'auteur de l'absence de l'information cruciale dans les stimuli pour établir cette distinction.

Ensuite, à l'aide d'une procédure de renforcement visuel, Harrison (2000) a entraîné six enfants apprenant le yorúbà et six enfants apprenant une langue non tonale à répondre à une paire de syllabes de tons Bas et Haut. Il a ensuite vérifié la capacité des enfants à discriminer entre les paires de tons Haut, Moyen et Bas avec la syllabe synthétique utilisée précédemment avec les adultes. Les résultats montrent que cinq sur six des enfants apprenant le yorúbà ont distingué la paire 190-210 (la frontière adulte entre les tons Haut et non-hauts), alors que seulement un enfant de l'autre groupe le pouvait. De plus, les enfants des deux groupes ne pouvaient pas distinguer d'autres paires de tons qui diffèrent de 20 Hertz situées entre 140 à 220 Hertz. Ces résultats suggèrent que les enfants apprenant une langue tonale sont sensibles aux variations de fréquences linguistiquement significatives dès l'âge de 6 à 8 mois. Cependant, l'étude de Harrison peut être critiquée sur une base méthodologique, notamment pour la petitesse de l'échantillon de participants et la nature des stimuli utilisés.

Dans sa thèse de doctorat, Mattock (2004) a exploré si une réorganisation perceptive des tons survient durant la petite enfance, comme pour les voyelles et les consonnes. En utilisant la procédure CHP, elle a examiné chez des enfants de 6 et 9 mois apprenant l'anglais ou le chinois (mandarin ou cantonais) la capacité de distinguer entre des contrastes tonaux et entre

la version musicale de ces tons. La prédiction voulait que les enfants apprenant l'anglais distinguent les tonalités musicales à 6 et 9 mois, mais les contrastes verbaux à 6 mois seulement, et qu'au contraire les enfants apprenant le chinois distinguent les stimuli verbaux ou non verbaux peu importe leur âge. Les stimuli utilisés pour l'entraînement et la phase de test impliquaient cinq exemplaires des tons Montant, Descendant et Bas en thaï, transportés par la syllabe [ba] et produits par un locuteur féminin. Dans une condition, les enfants devaient distinguer le contraste facile Montant et Descendant. Une seconde condition impliquait le contraste moins saillant Bas versus Montant. Les résultats révèlent qu'à 6 mois, les enfants des deux langues distinguaient les versions verbales et musicales des contrastes tonaux faciles et difficiles. Alors que les enfants chinois de 9 mois conservaient cette habileté, les enfants anglais du même âge ne pouvaient plus distinguer le contraste verbal difficile, suggérant que la réorganisation perceptive s'opère quelque part entre 6 et 9 mois.

En résumé, ces études suggèrent que les enfants peuvent traiter la fréquence fondamentale du signal acoustique à partir d'un jeune âge, et ce, pour les signaux non verbaux, la musique et la structure prosodique de la langue. Concernant le développement de la parole, l'acquisition des tons lexicaux se produirait entre 6 et 8 mois. Les études existantes sur la perception tonale impliquent des stimuli produits dans un environnement hautement contrôlé. Dans Mattock (2004) par exemple, les tons ont été produits en isolation par un seul locuteur. Les données de production chez l'adulte révèlent toutefois que les tons produits en conditions naturelles, par exemple en parole continue et par plusieurs locuteurs, résultent en patrons de F0 beaucoup plus variables que ceux produits en isolation ou par un seul locuteur (Xu, 1997, 1999). La parole adressée aux enfants se composent essentiellement d'énoncés contenant plusieurs mots (Shi, Morgan, & Allopenna, 1998) produits par de multiples locuteurs. L'acquisition tonale fait donc face au problème de variabilité.

1.3.4 Acquisition des tons et problème de la variabilité

Dans les langues tonales, la source principale de variabilité de F0 provient des tons lexicaux (Xu, 2001). D'autres sources de variabilité linguistique incluent les règles Sandhi, qui se rapportent à divers processus phonologiques (Peng, 2000). En mandarin par exemple, un ton Bas devient Montant lorsqu'il est suivi d'un autre ton Bas. Les patrons de F0 varient également en fonction du contenu segmental de la syllabe (Gussenhoven, 2004) et de la

réalisation simultanée de fonctions linguistiques de plus haut niveau, tel le focus prosodique et le type de phrases (Gauthier, Shi, & Xu, 2009; Xu, 1999). Une grande partie de la recherche sur la variabilité des tons se concentre sur la variabilité interlocuteurs. La fréquence fondamentale de femmes adultes se situe à environ ¾ d'octave plus haut que celle des hommes (Hillenbrand et al., 1995; Peterson & Barney, 1952), et la production de F0 varient grandement avec la croissance du tractus vocal durant l'enfance (Ménard, Schwartz, & Boë, 2004). Une autre source importante de variabilité de F0 provient du contexte phonétique, c.-à-d. des tons précédents ou suivants le ton cible dans un énoncé. En mandarin, la variabilité contextuelle provient majoritairement de l'effet résiduel de la tonalité précédente (Xu, 1997). La figure 1.2b montre les contours de F0 de 2880 exemplaires des quatre tons mandarins produits par un homme et une femme en parole continue et impliquant tous les contextes de tons possibles et diverses conditions de focus prosodique (données de Xu, 1999). La Figure 1.3 illustre les mêmes données dans différents espaces tonaux bidimensionnels, où les axes

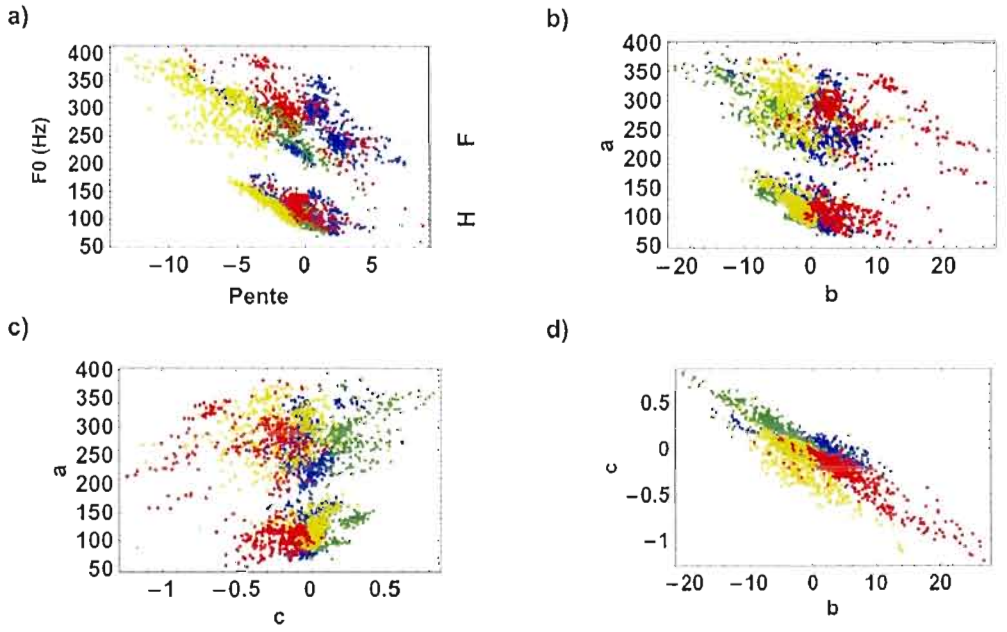


Figure 1.3 Divers espaces tonaux contenant chacun 2880 exemplaires de tons Haut (bleu), Montant (vert), Bas (jaune) et Descendant (rouge), produits par un locuteur masculin et un locuteur féminin en parole continue avec contextes tonaux et focus prosodique variables (données de Xu, 1999). **a)** Hauteur versus pente (coefficients *a* et *b* d'équations polynômiales de premier degré). Hauteur (*a*) versus coefficients *b* **(b)** et *c* **(c)** d'équations polynômiales de second degré. **d)** Coefficients quadratique (*c*) versus linéaire (*b*).

correspondent aux valeurs de diverses caractéristiques acoustiques reliés à la perception des tons mandarins (voir 1.3.2), et où chaque point représente un exemplaire tonal. Les caractéristiques acoustiques de chaque exemplaire sont estimées par les coefficients d'équations polynômiales de premier et second degré obtenue par une méthode des moindres carrés et fournissant une approximation continue des patrons de F0. Par exemple, les coefficients *a* et *b* d'équations linéaires $a + bx$ représentent respectivement la hauteur initiale (*Pitch height*) et la pente (*Slope*) de F0. Tel qu'illustré dans la Figure 1.3.a, ces deux caractéristiques ne peuvent ni séparer les quatre tons ou normaliser les différences inter-genres (M et F). Les trois autres figures illustrent diverses combinaisons des coefficients d'équations quadratiques de la forme $a + bx + cx^2$, où *a* représente la hauteur initiale et *b* et *c* la forme du contour, et plus précisément +/-*b* la hauteur du sommet ou de la vallée et *c* le moment du point de flexion (c.-à-d. vitesse nulle). Une combinaison de la hauteur initiale (indiquée par 'a' sur l'ordonnée des Figures 1.3.b et 1.3.c) avec l'une ou l'autre des composantes spécifiant le contour ('b' dans la Figure 1.3.b et 'c' dans 1.3.c) ne peuvent une fois de plus normaliser les genres. Cependant, la hauteur versus le coefficient linéaire *b* sépare les catégories Haut/Descendant de Bas/Montant (Figure 1.3.b), alors que la hauteur versus le coefficient quadratique *c* semble établir les contrastes Montant/Bas et Haut-Descendant. Enfin, les coefficients linéaires et quadratiques des contours de F0 séparent le mieux les quatre catégories tonales (Figure 1.3.d).

L'étude des tons lexicaux dans diverses langues a permis d'identifier la fréquence fondamentale comme dimension acoustique principale à la base de la perception tonale. D'autres études se sont penchées sur les manifestations acoustiques des tons produits dans diverses conditions et indiquent toutefois que la fréquence fondamentale est sujette à diverses sources de variabilité. Ceci soulève la question de l'invariance tonale chez l'adulte, et dans le cadre de cette thèse, la question de savoir comment l'enfant apprenant une langue tonale acquière ces catégories phonétiques à partir du signal acoustique. L'analyse présentée dans cette section suggère que l'information relative au mouvement de la fréquence fondamentale permet de réduire l'impact de plusieurs sources importantes de variabilité, compatible avec la théorie dynamique de l'invariance. Le rôle de l'information dynamique dans l'acquisition des

catégories phonétiques reste toutefois à déterminer, tout comme la capacité de l'enfant en bas âge à traiter ce type d'information.

1.4 Buts et hypothèses

Cette thèse étudie les mécanismes à la base du développement phonologique initial. Elle vise précisément à caractériser l'acquisition des catégories phonétiques durant la première année de vie, en abordant un type particulier de sons paroliers, les tons lexicaux, et une dimension spécifique du signal de la parole, la fréquence fondamentale. Trois études sont présentées afin de vérifier les hypothèses selon lesquelles 1) la poursuite auditive des patrons continus du mouvement de F0 permet de contraster les quatre tons mandarins, et 2) l'enfant peut percevoir cette information dynamique et l'utiliser afin de normaliser le signal de la parole. Deux études de modélisation simulent l'apprentissage perceptif du système de tons lexicaux mandarins par le biais de réseaux neuronaux artificiels non supervisés. La troisième étude lance deux groupes d'enfants à la poursuite auditive de patrons prosodiques continus, examinant leur capacité à détecter des variations subtiles de vitesse de F0.

Dans la première étude (Chapitre 2), des cartes auto-organisées (*Self-Organizing-Maps)* sont entraînées avec un corpus de tons produits par trois locuteur masculins, en parole continue et contenant tous les contextes de tons possibles. La seconde étude (Chapitre 3) présente quatre simulations impliquant un degré croissant de variabilité, utilisant la parole produite par quatre locuteurs masculins et quatre locuteurs féminins dans tous les contextes de tons possibles et impliquant diverses conditions de focus prosodique. En supposant une segmentation syllabique et une sensibilité aux propriétés distributionnelles et dynamiques du signal de la parole, il est prédit que 1) le réseau neuronal peut catégoriser les quatre tons mandarins à partir du signal continu de la parole, et que 2) la vitesse de la fréquence fondamentale, ou profil de vélocité (représenté par la première dérivée de patrons de F0) révèle les propriétés invariantes des tons de façon supérieure à F0. Un résultat positif indiquerait que la poursuite du mouvement acoustique reflétant les gestes articulatoires constitue une stratégie simple et efficace pour l'acquisition des tons, possiblement aussi pour d'autres types de catégories phonétiques comme les voyelles et les consonnes, et que l'analyse des caractéristiques phonétiques peut survenir à une étape ultérieure.

Afin de soutenir le rôle potentiel de l'information acoustique dynamique dans l'acquisition de catégories phonétiques, la troisième étude vérifie à l'aide d'une procédure de regard préférentiel la capacité d'enfants de 4 et 8 mois à traiter les profils de vélocité de F0. Les stimuli consistent en des séquences de non-mots produits par un locuteur féminin inconnu des participants et transportant des patrons d'intonation possibles et impossibles, ces derniers violant la contrainte articulatoire de la vitesse maximale de changement de F0. La distinction des patrons possibles et impossibles indiquerait chez l'enfant la capacité de percevoir le mouvement acoustique. De plus, une préférence pour les sons possibles, malgré l'absence d'expérience avec le locuteur expérimental, indiquerait non seulement la capacité de traiter l'information dynamique, mais de l'utiliser afin de normaliser le signal de la parole.

2 L'ACQUISITION DES CATÉGORIES PHONÉTIQUES PAR LA POURSUITE DU MOUVEMENT ACOUSTIQUE

2.1 Résumé de la publication en français

Cette étude explore la façon dont l'enfant peut apprendre les catégories phonétiques de sa langue maternelle à partir du signal hautement variable de la parole. Elle vise précisément à modéliser l'apprentissage non supervisé des tons lexicaux en mandarin par le biais de réseaux neuronaux artificiels sous forme de cartes topographiques auto-organisées. Lors de la première simulation, l'entraînement consiste à présenter à un réseau neuronal les patrons de fréquence fondamentale (F0) de taille syllabique, et à un second réseau les patrons de vélocité associés (D1, la première dérivée de F0) produits par de multiples locuteurs en énoncés continus. Avant l'apprentissage, aucune opération n'a été réalisée afin de réduire la variabilité du signal, ou encore pour inclure au corpus de données des caractéristiques abstraites comme la hauteur ou la pente de F0 pouvant simplifier la tâche des réseaux neuronaux. Lors de la tâche de rappel, les résultats montrent que les patrons de F0 atteignent un degré relativement élevé de catégorisation des quatre tons. Toutefois, la performance de catégorisation du réseau entraîné à partir de D1 s'avère presque parfaite. L'examen détaillé des classes de patrons de vélocité formées par le second réseau révèle que D1 a éliminé la variabilité du signal pour directement refléter les caractéristiques du geste articulatoire impliqué dans la production tonale. Lors d'une seconde simulation impliquant un nouveau réseau neuronal, l'apprentissage des patrons de D1 développés durant la première simulation résulte en quatre prototypes entretenant une relation un à un aux quatre tons. Les résultats de cette étude sont discutés dans le cadre des théories de l'acquisition du langage, de la perception de la parole et de la production de la parole.

2.2 Learning phonetic categories by tracking movements

2.2.1 Abstract

We explore in this study how infants may derive phonetic categories from adult input that are highly variable. Neural networks in the form of self-organizing maps (SOMs, Kohonen, 1989, 1995) were used to simulate unsupervised learning of Mandarin tones. In simulation 1,

we trained the SOMs with syllable-sized continuous F_0 contours, produced by multiple speakers in connected speech, and with the corresponding velocity profiles (D1). No attempt was made to reduce the large amount of variability in the input or to add to the input any abstract features such as height and slope of the F_0 contours. In the testing phase, reasonably high categorization rate was achieved with F_0 profiles, but D1 profiles yielded almost perfect categorization of the four tones. Close inspection of the learned prototypical D1 profile clusters revealed that they had effectively eliminated surface variability and directly reflected articulatory movements toward the underlying targets of the four tones as proposed by Xu and Wang (2001). Additional simulations indicated that a further learning step was possible through which D1 prototypes with one-to-one correspondence to the tones were derived from the prototype clusters learned in Simulation 1. Implications of these findings for theories of language acquisition, speech perception and speech production are discussed.

Key words: category formation, infant speech perception, language acquisition, unsupervised learning, self-organizing-maps, target approximation, lexical tone, contextual tonal variation, theories of speech production and perception

2.2.2 Introduction

The task of learning the sounds of the native language is daunting. Infants do not receive explicit language instructions, nor are they able to make inquiries about the structure that they are learning. They must discover the phonetic categories of their native language from the speech input of the surrounding speakers. The task is further complicated by the fact that they do not know how many categories to discover along any particular input dimension. To make things worse, the input they receive is highly variable. That is, there is lack of invariant acoustic manifestation of phonetic categories. A classical example was discussed by Peterson and Barney (1952), who demonstrated how between-category formant frequencies show great overlap in the American English vowel space when produced by multiple speakers. Many sources of variability have been studied since, including coarticulation, spoken rhythm and dialectal variations. Phonetic categories other than vowels have also been shown to be subject to different sources of variability. Liberman (1970) pointed out how English /b, d, g/ produced in the same vowel context show continuously changing 2nd formant transition slope patterns with identifiable category boundaries separating the three sounds. Subsequent

perception studies have shown good agreement between these acoustic characteristics and the perception of stop consonants (Menon, Rao, & Thosar, 1974; Ohde, 1988). However, these general slope patterns can change drastically in certain vowel contexts. For example, the second formant slope for /d/ is positive before /i/ and negative before /u/ (Delattre et al., 1955). The variable second transition slope patterns for /d/ overlap to some degree with the slope patterns of /b, g/ in certain vowel contexts.

The variability problem is just as severe when it comes to lexical tones. In many languages, words are distinguished from one another not only by consonants and vowels, but also by pitch patterns that occur during the voiced sound (Yip, 2002). In Mandarin, for example, the syllable /ma/ can mean "mother", "hemp", "horse" or "to scold" depending on whether its pitch pattern is high-level (High tone), rising (Rising tone), low-dipping (Low tone) or falling (Falling tone). The primary acoustic correlate of tones is F_0, i.e., the fundamental frequency of voice (Abramson, 1962; Chao, 1933; Howie, 1976). Although other phonetic/prosodic cues have been suggested to contribute to the perception of tones (e.g., duration, amplitude (Whalen & Xu, 1992)); voice quality for languages with lexical phonation types (Andruski & Ratliff, 2000; Maddieson & Hess, 1986), F_0 has been consistently shown to be the dominant cue in adult tone perception (e.g., Klein, Zatorre, Milner, & Zhao, 2001; Whalen & Xu, 1992). Figure 2.1a shows the (time-normalized) F_0 contours of five tokens and their means of the four Mandarin tones produced in citation form by a male speaker (data from Xu, 1997). As can be seen, when produced in isolation by a single speaker, the tones are well separated even when time-normalized. They become much less separated, however, when spoken in connected speech and when uttered by different speakers. Figure 2.1b-c shows the means and distributions of the same four tones spoken in connected utterances by three male speakers (Xu, 1997). The extensive overlap between the tones comes from two major sources[3]. The first is the difference in the pitch range of individual speakers and the second the variability introduced by tonal context in connected speech (Shen, 1990; Xu, 1994, 1997). Similar variability has been found in other tone

[3] There are many other sources of variability in tonal realization, as discussed in detail in Xu (2001; 2005). However, most of the other sources of variability are kept constant in the data shown in Figure 2.1.

languages such as Thai, Vietnamese and Yoruba (Gandour, Potisuk, & Dechongkit, 1994; Han & Kim, 1974; Laniran & Clements, 2003).

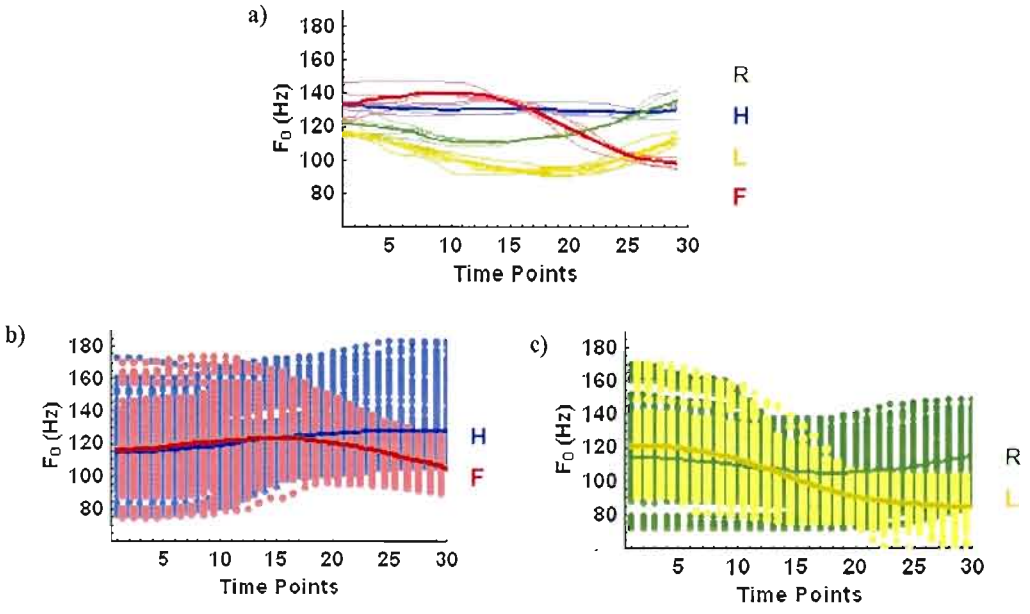


Figure 2.1 Tones produced **a)** in citation form by one speaker and **b-c)** in connected speech by 3 speakers. Thick lines correspond to means while pale background to the distribution of High (blue), Rise (green), Low (yellow) and Fall (red) (data from Xu, 1997).

While tonal variability is similar to segmental variability in nature, tones typically involve a single primary acoustic dimension, namely, F_0. This contrasts with the multiple acoustic dimensions such as formants or spectral peaks required for characterizing vowels and consonants. The variability problem with tones is therefore at least limited to a single dimension, which makes them ideal for testing hypotheses that involve detailed mechanisms of phonetic acquisition. In the present study, we will therefore use lexical tone as a probing tool to find a breaking point for understanding how infants could develop phonetic categories from adult speech input that is highly variable.

Although there has been much research on tone perception by adults, most studies have only investigated the perception of tones produced in isolation (e.g., Abramson, 1978; Gandour, 1983; Karlgren, 1962; Liang, 1963; Massaro et al., 1985; Moore & Jongman, 1997;

Shen & Lin, 1991; Whalen & Xu, 1992). Studies on speech input to infants have shown, however, that about 90% of parental speech to infants is multi-word utterances (e.g., Brent & Siskind, 2001; Shi et al., 1998; van de Weijer, 1998). In addition, infants generally hear multiple speakers in their daily life. Strikingly, young learners in tone languages seem to have their preliminary tonal categories in place before the first year of life. In fact, their earliest comprehension vocabulary and discriminative abilities already demonstrate certain knowledge of tonal categories. Perception studies by Harrison (2000) and Mattock (2004) indicate that tone-learning infants attend more closely to phonemic tones than to non-significant pitch variations at six months, and that their tonal perception is influenced by the phonemic systems of their ambient language at nine months. Moreover, given that these studies used stimuli with some degree of variability, the results also show that infants can handle certain variability in tonal perception. Although there is still no direct evidence about the age at which infants can deal with between speaker and contextual variability in tonal perception, research on infant speech perception has shown that perceptual normalization of segmental variability in vowels and consonants has already happened in infants before the onset of speech production (e.g., Jusczyk et al., 1992; Kuhl, 1976b, 1979, 1983; Kuhl & Miller, 1982). Since speech input to infants consists mainly of multi-word utterances by multiple speakers, tone learning must also involve processes that can not only effectively resolve the speaker and contextual variability, but also discover the number of tonal categories as well as the invariant characteristics of each category. But the question is, of course, how can infants do it?

To understand how it is possible for infants to discover tonal categories despite substantial variability and inter-tonal overlap, it is necessary to understand the nature of the variability. Of the two sources of variability mentioned earlier, the nature of cross-speaker difference in pitch range is quite transparent. The length and thickness of the vocal folds vary extensively across gender, age and individuals. The mean pitch therefore differs from speaker to speaker (Zemlin, 1988). The nature of contextual variability is more complex, but much has been learned about it in recent research (e.g., Gandour et al., 1994; Xu, 1997, 1999). For example, much of the variability in both Mandarin and Thai is induced by the preceding tone, i.e., due to carry-over effect, although the following tone exerts anticipatory effect in some

contexts. For example, if a High tone in Mandarin is preceded by a Low tone, it will be realized with a rising contour in the earlier part of the tone. Nevertheless, for all the four tones, regardless of what the preceding tone is, the F_0 contours of the syllable associated with the tone all gradually converge over time to an asymptote that is characteristic of the underlying tone: high-level for High, low-rising for Rise, low-level for Low and high-falling for Fall (Xu, 1997, 1999). Based on these findings, Xu and Wang (2001) proposed a theoretical model of contextual tonal realization: the Target Approximation (hereafter TA) model to account for the contextual variability of tones in production.

In the TA model, surface F_0 patterns are characterized as asymptotic movements toward underlying pitch targets defined as simple linear functions. These targets can be either static or dynamic. Static targets are specified by relative pitch height (e.g., [high], (Bradlow, Pisoni, Yamada, & Tohkura)) and dynamic targets by both the relative height and velocity of the pitch movement (e.g., [rise], [fall]). The pitch target therefore is the articulatory goal associated with the lexical tone. The articulation of both static and dynamic targets is subject to the physical constraints of (1) the maximum speed of pitch change due to the properties of the laryngeal muscles and mechanical characteristics of the larynx (Sundberg, 1979; Xu & Sun, 2002) and (2) the coordination of the larynx and other articulators (Kelso, 1984; Kelso, Saltzman, & Tuller, 1986; Xu & Wang, 2001). The maximum speed of pitch change has been demonstrated to be rather slow so that pitch movements are frequently made as fast as possible during regular speech (Xu & Sun, 2002) but still leading to recurrent undershoot (Xu, 1999). This means that surface F_0 contours consist mostly of movements toward one tonal target or another. The constraint of coordination of the larynx with other articulators has been argued to result in full synchrony between laryngeal and supra-laryngeal movements, so that the F_0 movement toward each tonal target is made only within the syllable that the tone is associated with (Xu & Wang, 2001).[4]

Based on the TA model, an intriguing prediction can be made. That is, despite the extensive contextual tonal variability as well as the tonal overlap due to pitch range

[4] The TA model has been used to explain various tonal and segmental data in both tonal and non-tonal languages (e.g., Chen & Xu, 2006; Xu & Liu, 2007; Xu & Xu, 2005). It has also been tested in speech synthesis (Prom-on, Xu, & Thipakorn, 2009; Sun, 2002).

differences across speakers, it is possible to infer the underlying pitch targets from the manners of F_0 movements even without context and speaker information, assuming that syllabic segmentation has been done. This can be achieved by taking the first derivative of F_0 (henceforth D1), which is the velocity of F_0 movement. D1 reflects the characteristics of F_0 movement toward the underlying pitch target. Moreover, as there exist pitch range variations due to speaker differences and intonational factors (Xu, 2005), the transformation of F_0 to D1 automatically eliminates most of these pitch range differences. For example, suppose the surface F_0 contour of a tone is represented by a polynomial of the form:

$$y = a + bx + cx^2 + \ldots + mx^n \qquad (2.1).$$

Taking the first derivative of (1) reduces it to:

$$y' = 0 + b + 2cx + \ldots + nmx^{n-1} \qquad (2.2).$$

The transformation turns a, the y-intercept of the polynomial, uniformly to 0, thus normalizing the initial F_0 height, which contains information both about the speaker and about the preceding tone. While both kinds of information are useful, they are not directly relevant to the tone to be recognized. Although Xu (1994) has shown the usefulness of contextual information for the recognition of severely distorted tones when they are produced in a prosodically weak position (the second syllable in a trisyllabic word) in Mandarin, it is not known whether context information is always indispensable. A major question we pose in the present study is therefore: Is the information about movements toward underlying tonal targets as represented by D1 sufficient for the categorization of the four Mandarin tones produced in connected speech by multiple speakers? A positive answer would point to a powerful tool that listeners may have in their possession for disentangling the vast amount of variability without the help of contextual information and complex normalization schemes. More importantly, for infants who are born into a Mandarin-speaking community with presumably no pre-endowed tone categories, a positive answer would mean that they are actually able to use this powerful tool for deriving from the adult input the underlying pitch targets associated with the lexical tones even if they have not yet developed effective

strategies for taking advantages of the contextual information and for normalizing speaker differences.

In speech perception research, it is often assumed that some kind of feature extraction needs to take place in order to recognize a sound as belonging to one category or another. Similarly, in tone perception studies, proposed solutions typically try to single out an acoustic feature such as height or slope of F_0 contours corresponding to each tone (Abramson, 1978; Gandour, 1983; Massaro et al., 1985; Shen & Lin, 1991; Wang & Li, 1967). Thus for both segmental and tonal perception, there is a popular assumption that some kind of preprocessing is done to single out certain abstract features in order to perform categorization. Note that, however, preprocessing, even if it does occur, would be just as difficult as the categorization task itself, because it would still need to first resolve the variability problem. If, instead, phonetic categories could be discovered by directly tracking continuous movements in the acoustic signal, the need for feature extraction would be greatly reduced. Thus, another major question we will ask in the study is this: Is it possible to derive phonetic categories directly from continuous signal input without extraction of abstract features?

It has been known for a long time that as they grow older, infants become less sensitive to non-native phonemic contrasts which are non-phonemic in their native language (Werker & Lalonde, 1988; Werker & Tees, 1984), as well as to sounds of foreign languages that are similar to those in their first language (Best, 1993). Such a reduction in sensitivity becomes even more extreme in adults (Best, McRoberts, & Goodell, 2001). At the same time, however, it has been demonstrated that even adults have not totally lost their ability to discriminate sounds that bear little resemblance to any sound in their native language (Best et al., 2001) or to establish a new category division in sounds that belong to the same category in their native language (Bradlow et al., 1997). These observations have lead to perception theories such as the Functional Reorganization Model (Werker & Tees, 1984), the Native Language Magnet Theory (NLM: Kuhl, 1991) and the Perceptual Assimilation Model (PAM: Best, 1995). Little is known, however, about how these behavior patterns are linked to the core mechanisms of the learning process itself. The third question we will ask in the present

study is therefore: Is there a possible link between the decline in sensitivity to within-category differences and the core mechanisms of learning phonetic categories?

2.2.3 Methodology

To test the possibility that D1 can be used in the perception and the learning of tones in Mandarin Chinese, we use a self-organizing neural network known as the self-organizing map (SOM) (Kohonen, 1989, 1995). The SOM is a statistical pattern recognition device using unsupervised learning methods for discovering the structure of high dimensional data. It is a particular case of neural map for which the basic idea comes from the discovery of topographically organized projections from the periphery to cortical areas in the brain (Kohonen, 1982). Information encoding by topographic maps has been observed in many regions of the brain, including some areas of the auditory cortex (Crottaz-Herbette & Ragot, 2000; Pantev et al., 1994; Seldon, 1985; Wessinger, Buonocore, Kussmaul, & Mangun, 1997). Because of its visualization properties, the SOM is useful for exploring the internal properties of the data as well as for modeling the place coding of sound properties in brain topographic maps. The model is also consistent with an inductive account of speech perception development. Recent research has shown that infants are sensitive to the statistical distributional properties of speech sounds in the input (Maye et al., 2002). Similarly, the SOM decodes systematic statistical patterns found in input distributions. The structural and functional properties of the SOM, combined with the simplicity of its algorithm, thus provide an attractive method for revealing invariants in the speech signal in general, and for modeling the learning of tone categories by naïve learners in particular. The detailed algorithm of the model is presented in Appendix 1.

Our simulations attempt to verify how continuous F_0 and D1 perform as input to the SOM and how D1 function as a normalizing parameter for both between speaker and contextual variability. Two simulations were performed to test the effectiveness of category formation through unsupervised learning, with either syllable-sized F_0 profiles obtained from a natural data set (Xu, 1997) or the syllable-sized velocity profiles (D1) derived from those F_0 profiles as input. Simulation 1 used a large receptive mapping area (with many units) for training and testing. Simulation 2 used a much smaller mapping area but with prototypes

developed in Simulation 1 as training input and the same F_0 and D1 profiles as in Simulation 1 as testing input.

Simulation 1

Input coding. The input corpus contains 1800 exemplars of the four Mandarin tones produced in connected utterances by 3 adult male speakers (data from Xu, 1997). Each stimulus corresponds to the first or second syllable of disyllabic 'mama' produced in the middle of a carrier sentence which had either high or low pre-target F_0 offset and post-target F_0 onset. Each input token is a 30 data point vector composed of equal-distanced discrete values taken from a syllable-sized time-normalized F_0 curve (for the exact F_0 extraction procedure, see Xu, 1997). The data are first transformed from Hertz to the Bark scale according to:

$$F_{0\,bk} = 7 \,.\, \mathrm{Log}(F_{0\,hz}/650+(1+ (F_{0\,hz}/650)^2)^{1/2}) \qquad (2.3).$$

The Bark scale is a frequency scale corresponding to human auditory perception. It is logarithmic at high frequencies but linear in low frequencies. Thus the transformation has no other impact than rescaling the pitch patterns in a way that facilitates future comparisons between F_0 and D1 mappings. In the input corpus, F_0 values range from 50 to 180 Hertz and D1 from –13 to 9. By using barks (0.5 to 2) F_0 and D1 are more comparable in scale. The D1 profiles are generated according to:

$$D1 = 0.5\ (F_{0\,hz}\ (t+1) - F_{0\,hz}\ (t-1)) \qquad (2.4),$$

which yields input vectors of 28 dimensions representing the discrete first derivatives of F_0 patterns.

Learning phase. During the adaptation process, the SOM implements a regression algorithm to map a continuous input distribution $P(x)$, $x_i \in X$, onto a discrete output space — a 10 x 10 map consisting of 100 units. The training corpus contains 900 stimuli, which are randomly presented to the network for 100 times. Each time the neighborhoods on the map are shifted to better fit the data.

Testing phase. During the recall task, new exemplars are used to verify the network's capability to generalize to novel data. The trained network assigns each input pattern, from a new set of 900 tokens, to a single unit using the transmission rule described in Appendix 1. The testing corpus, which contains as much variability as the learning one, is presented in an orderly fashion. The procedure involves presenting, in order, all exemplars of High, i.e., tone 1, (240 tokens), Rise - tone 2 (240), Low - tone 3 (180) and Fall - tone 4 (240) tones. The Low category contains fewer exemplars because a tonal variant of this tone due to a sandhi rule has been removed from the training and testing sets (Low tone becomes Rise when followed by another Low tone in Mandarin. Cf. Xu, 2001).[5]

Output coding. The trained networks are squared arrays of 10 x 10 processing units, each one being tuned to a particular subset of input patterns. During the testing phase, the number of input patterns projected onto each unit is indexed into a global firing frequency matrix. Units which fire at least once during recall according to the transmission rule (see Appendix 1) are considered as operable units. If a unit never fires during recall, it is considered non-operable. The number of activations of each unit for each class of input patterns is also indexed into four tone firing frequency submatrices, the sum of which corresponds to the global matrix. The proportion between tone firing frequencies and the global firing frequency of a unit yields the tone probability for this unit, i.e. its probability to be activated by each class given the testing corpus. Tone probabilities give rise to the distinction between categorized and ambiguous units. Units that have a tone probability equal to or above 68% are considered as categorized and are labeled with that particular tone. Units without such a majority class are considered as ambiguous. Such units respond to multiple tones, but none of the tones is dominant. This criterion is decided based on the central limit theorem. That is, plus and minus one standard deviation from the mean includes 68% of the responses to a particular tone.

---

[5] Tonal variation due to sandhi is a problem beyond the scope of the current project, because as far as surface acoustics is concerned, the sandhi-derived Rising tone resembles the underlying Rising tone so closely that listeners do not hear them as different tones (Peng, 2000; Wang & Li, 1967). This tonal sandhi alternation for Tone 3 involves context-sensitive rule learning at a morpho-phonological level. It is a learning process entirely different from the type of learning that we are testing in the current study.

Measures

Quantitative criteria. Performance and reliability measures are first used to assess the global properties of the maps. The first performance measure is categorical error. It corresponds to the proportion of the network which responds to more than one class, i.e., the number of ambiguous units on the total number of operable units. The second performance measure is classification error. It corresponds to the probability of the network to respond ambiguously during recall. The test tokens which land on ambiguous units are considered errors. The classification error is thus the number of error tokens divided by the total number of input tokens in the testing corpus. For example, if half of the testing corpus activates ambiguous units during recall, the classification error would correspond to 450/900 = 0.5. The performance measures help to quantify the clustering properties of the trained maps and they reflect the amount of category information carried by the input distributions.

The most common measures of reliability assessment of the trained SOM are the quantization and the topology errors. The quantization error, which evaluates the precision of the mapping, is given by:

$$e_q = 1/n \, (\Sigma \, ||x_i - r_v||) \qquad (2.5),$$

where x_i corresponds to the input pattern and r_v to the best matching unit (BMU) for that pattern. The equation gives the sum of the distance between each input pattern and its BMU divided by the total number of stimuli. It thus yields the average distance between input vectors and their BMU's receptive field center. The second reliability measure is the topology error, which evaluates the topographical organization of the map. It is given by:

$$e_t = 1/n \, (\Sigma \, d(x_i)) \qquad (2.6),$$

where d = 0 if the first and second BMUs for a given input are next to each other and d = 1 if they are not. The equation thus indexes 1 every time topology is not respected and divides the final amount by the total number of stimuli. Reliability measures are usually used to ensure the good functioning of the SOM so as to validate the conclusions inferred from other results about a data set. In the present study, they also act as a window on F_0 and D1 input distributional properties.

The preceding measures are simple scalar summaries for describing the clustering and distributional properties of the maps. A more detailed analysis of groups of units is useful for observing within and between-category map properties. In this regard, the between-category assessment of each condition can be expressed in terms of confusion patterns between each tone and will be presented in the form of confusion matrices. Finally, the rate of success measured for each tonal category is obtained by the sum of input tokens which activate corresponding labeled units divided by the amount of input tokens belonging to this category. For example, if all High input patterns activate High categorized units, the rate of success for High is 240/240 = 1.0.

Visualization of the maps. Projection techniques are used for graphically revealing the distributional and neighboring structure of the trained maps. Traditional ways to visualize the state of the SOM include the Sammon's mapping and other derived techniques. For example, the u-matrix is a regular grid of neurons between which the relative distance is represented in tones of grey; the lighter the color, the closer neurons are to each other. Another popular technique is the data histogram, which shows how many stimuli belong to a cluster defined by each neuron. Visualization can also be done by projecting the weight vectors into a color space in which similar units are assigned similar colors (e.g., Kaski, Venna, & Kohonen, 1999; Varfis, 1993). In this study, the coloring of the maps is done by representing tone categories with four distinct colors produced with the CMYK color system. The High tone is represented by blue, specified by a mix of cyan and magenta in the vector [1,1,0,0]; Rise = [1,0,1,0]; Low = [0,0,1,0] and Fall = [0,1,1,0]. Each map unit is thus described by a four-dimensional vector where the last element (black) remains null and where the other elements are specified in terms of the firing probabilities for each tone. If each category is well separated in the data, the color map should be divided into regions by classes. When a unit responds to more than one tone, the colors associated with each tone are mixed to yield 'impure' colors. Figure 2.2a shows the legend for interpreting the color maps.

A more conventional way to visualize the final state of the network is the phoneme map (Kohonen, 1989, 1995). This technique assigns each processing unit a label corresponding to the majority class of that unit. While the color map produces a clear display of whether

regions of clusters are formed, the phoneme map reveals more precisely the confusion areas. Figure 2.2b shows an idealized phoneme map of 4 units.

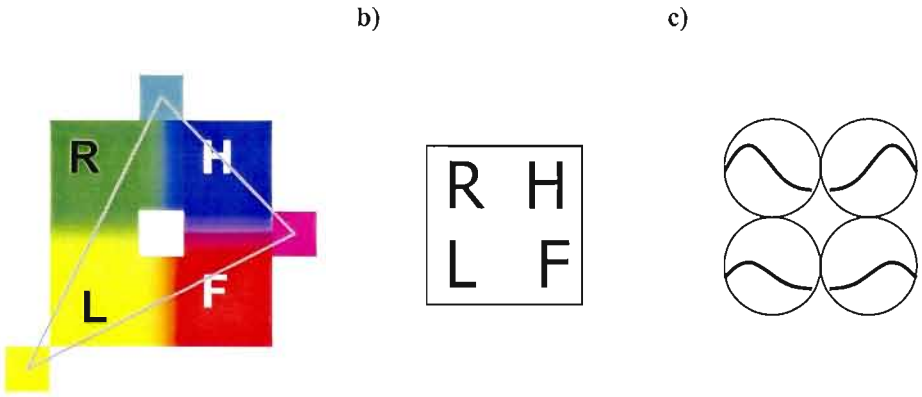


Figure 2.2 **a)** Color legend of the color map using the CMYK color system (see text for details); **b)** Idealized 2 x 2 phoneme map; **c)** Idealized 2 x 2 internal map.

Finally, the internal maps can be used to infer important characteristics and subtle details of a data set. An internal map is a graphical display of the connections which link the input space to the output space. More precisely, it is a projection of the receptive field center of the output units onto the input space. For example, the internal map of a network with *n* output units connected to a two-dimensional input space shows *n* data points on a plan, each of which corresponds to an output unit's receptive field center. In this study the input space as well as each output unit's receptive field center are 28 or 30-dimensional. The internal maps of F_0 and D1 thus project the prototype vector of each unit in a frequency or velocity by time space for visualizing the patterns developed during training. The prototype vectors can be observed separately as in Figure 2.2c, where four neighboring units display similar temporal patterns. The F_0 and D1 internal maps are presented later in a different style, overlaying the prototypes of each class on a single graph for a more direct comparison of groups of map units. Before engaging in such a detailed analysis, we consider whether or not D1 is a strong normalization procedure.

2.2.4 Results and discussion

In this section, the results and their interpretations are presented with respect to different aspects of the trained maps. The first part describes and compares the maps' global properties for both F_0 and D1 conditions. The second part focuses on the comparisons between groups of map units formed by the F_0 and D1 training corpus. Finally, the description of individual map units is presented mainly for the D1 condition.

Global map and input distributional properties

The global maps show directly on the trained network, rather than in the input space, whether topologically ordered categories are present in the data.

Performance results. Table 2.1 shows categorization errors (column 2) and classification errors (column 3) for F_0 and D1. The categorization error is larger for F_0 than for D1 (0.20 > 0.03), indicating that a majority of the D1 map units are category-specific while a larger portion of the F_0 map contains ambiguous units. Figure 2.3a shows how ambiguous units form widely spread confusion areas on the F_0 map. In contrast, in Figure 2.3b the D1 map shows a cleaner division of regions by classes, thus better representing the categories. These results suggest that the input distribution of D1 contains more categorical information than does the F_0 distribution.

Table 2.1 Categorization and classification errors of the performance measures for F_0 and D1 conditions.

	Performance measures	
	Categorization error	Classification error
F_0	0.20	0.22
D1	0.03	0.03

Unlike the categorical error measure, which assesses the units' responses, the classification error measure represents the percentage of new tokens that are ambiguously classified. The results (column 3 in Table 2.1) indicate a higher proportion of tokens being ambiguously classified in the F_0 map (0.22) than in the D1 map. The performance with D1

(0.03) is again nearly error-free, as in the case of the categorical error measure. In terms of input properties, this means that the density of the F_0 input distribution is greater than that of the D1 distribution in overlapping regions, i.e., more tokens are present in the overlapping region in the F_0 input space.

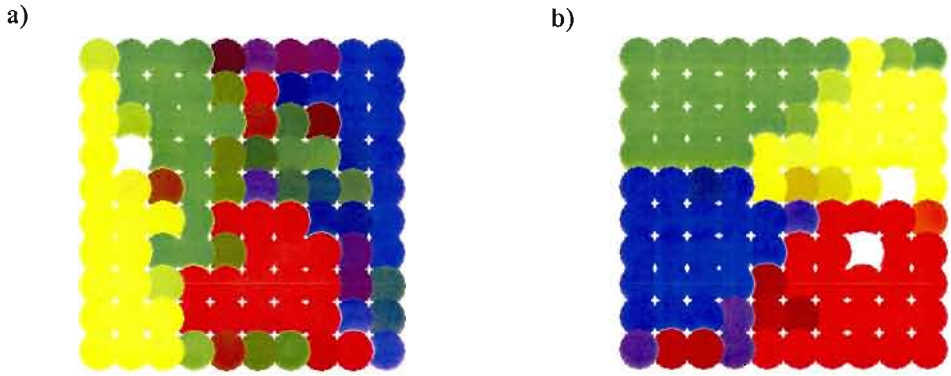


Figure 2.3 Color maps of **a)** F0 and **b)** D1 after training.

Reliability results. The quantization and topology error for each map are given in Table 2.2. The quantization error is higher for D1 than for F_0 (0.24 > 0.12), indicating that the average distance between D1 input patterns and their BMU is twice that of F_0's. These results, combined with the performance results, show that minimization of error accomplished by training does not necessarily give a better recognition rate. In the present context, category separation is more important than error minimization, and the result suggests that quantization error being too low might actually impede the ability of the network to detect between-category differences. This is because a lower quantization error corresponds to a higher number of units each having a small receptive field, which in turn might indicate a more compact global density function of the input distribution. The more compact a distribution is, the harder it may be to distinguish between neighboring data points which in fact could belong to different classes. According to this argument, D1 better represents tonal categories because it stretches the input space in such a way as to enhance the between-category contrasts.

Table 2.2 Quantization and topology errors of the reliability measures for F_0 and D1 conditions.

	Reliability measures	
	Quantization error	Topology error
F_0	0.12	0.009
D1	0.24	0.005

The topology conserving property of the SOM indicates whether or not the input distributions under study possess intrinsic organization, i.e., neighborhood structure. The F_0 topology error is higher than that of D1 (0.009 > 0.005), suggesting that the F_0 data set contains more discontinuities and that velocity profiles form more coherent clustering of tonal categories. D1 thus seems better suited than F_0 for topographical representation. The topology error can also indicate the proportion of the input tokens corresponding to a particular category which might switch class due to only small variations. In this sense, the F_0 system is more sensitive to noise than D1.

Summary of global results. The analyses in this section first showed that more distinct clusters are formed on the D1 map than on the F_0 map and that the probability for these clusters to be activated during recall was much higher for D1 than for F_0. The quantization and topology errors indicated that D1 was a more reliable cue for tone recognition and better suited for topographical representation of the input.

Groups of map units and input manifold properties

In this section we examine more specific aspects of the maps and of the data sets.

Between-category ambiguity. The phoneme maps in Figure 2.4 show the categorized (single labels) and ambiguous units (multiple labels) of F_0 and D1. The F_0 map contains more ambiguous units, which also show greater diversity among the categories they confuse. For the 20 ambiguous units of F_0, four units confuse tones H and R, three confuse tones H and F, two confuse tone R and L, four confuse tones R and F, and seven units confuse more than two categories. In contrast, the D1 map contains fewer confused units. Specifically, two units confuse tones R and L and a single unit confuses tones H and F. Overlap is thus present

between each tonal category in F_0 while the D1 map shows a more systematic confusion pattern. Another observation is that D1 overlapping regions are situated between the two classes they separate, while this is not always the case with F_0. For example, the HF cluster in the middle top region of the F_0 map appears between the High and Rise classes and the HR cluster (low right region) is between High and Fall regions. The D1 map thus seems more adequate for characterizing the boundaries between tonal categories.

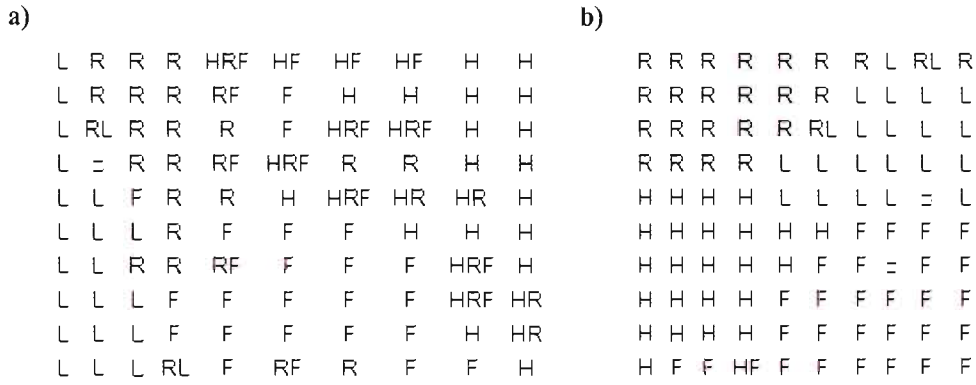


Figure 2.4 Phoneme maps of **a)** F_0 and **b)** D1: categorized (single label) and ambiguous (multiple label) units. Squared label correspond to non-operable units.

For revealing more subtle confusion patterns between each tone, Table 2.3 presents a confusion matrix where the rows correspond to the speakers' intended targets and the columns to the majority class of the units responding to the input patterns. The last column shows the number of tokens for each category which activate ambiguous units. For example, in the F_0 condition, of the 240 intended High tone targets, 162 are classified correctly to the High units, 6 misclassified to Rise units and 72 landed on ambiguous units. As can be seen, the total number of misclassification is lower for D1 (27) than for F_0 (55). Also, the darker elements of the matrices, which show higher probability misclassification patterns, are in greater number for the F_0 map, in which tone R gets mostly misclassified to tones H, L and F and tone F to tones H and R. The D1 condition appears to show only a single misclassification pattern between tones H and F.

Table 2.3 Confusion matrix for F_0 and D1 conditions.

		Confusion matrix				
		High	Rise	Low	Fall	Ambiguous
F_0	High	162	6	0	0	72
	Rise	**11**	157	**8**	**6**	58
	Low	0	0	172	0	8
	Fall	**11**	**12**	1	160	56
D1	High	219	1	2	**8**	10
	Rise	2	226	3	0	9
	Low	0	0	172	1	7
	Fall	**9**	0	1	224	6

Within-category rate of success. In the D1 confusion matrix, the number of tokens assigned to the corresponding majority class is overall higher than in F_0, as shown by each matrix diagonal element. This agrees with other results gathered so far. The Low tone in F_0, however, behaves differently, which leads us to consider each category in detail. The within-category rate of success for each tonal category is shown in Table 2.4. The rate of success corresponds to the number of times categorized units respond to a corresponding intended target divided by the total number of tokens of this category in the testing corpus. For example, of the 240 High tones presented to the network in the F_0 condition, 162 are projected onto a High unit, yielding a rate of success of 162/240 = 0.68 for the High category.

Table 2.4 Within-category rate of success for F_0 and D1 conditions.

	Rate of success	
	F_0	D1
High	0.68	0.91
Rise	0.65	0.94
Low	0.96	0.96
Fall	0.67	0.93

These results show that in the F_0 condition, High, Rise and Fall tones share a similar rate of success of about 66%, while the Low tone enjoys a success rate of 96%. Learning from F_0 information thus allows the network to only recover the Low tone with a high level of accuracy. In contrast, the results from the D1 condition indicate that every category shares a similar high rate of success which varies between 91 and 96%. Together with the confusion pattern results, the rate of success brings further evidence that D1 better represents tonal categories.

Simulation 2: Modeling the abstraction of categories after clustering formation

The results of Simulation 1 suggest that the D1 profiles of the four Mandarin tones provide sufficient information for a naïve system with no pre-existent tone categories to develop distinct cluster regions for the four tonal categories with well-defined boundaries in between. To answer question 3 raised in the Background, we test in a new simulation whether the learning system is able to further abstract from the learned maps four distinct categories. Based on the assumption that the new process simulated is neurologically linked to that of the first simulation, the neural map consists of the same number of units as the number of clusters learned in simulation 1.

Methodology. Instead of the 10 x 10 array used in Simulation 1, the neural map is now a two-dimensional array of 2 x 2 units. The training corpus now consists of 100 input profiles of F_0 or D1, each corresponding to a unit prototype vector developed in Simulation 1. During training, the learning parameter is kept the same as in Simulation 1, but the neighborhood function has been adjusted to better fit the size of the new map, reducing its values by a factor of 10 (i.e. 100->1 becomes 10->0.1). This is reasonable given that smaller radius is more appropriate for a four-unit network, as opposed to the 100-unit network in Simulation 1. The testing phase, as well as the output coding, are identical as in Simulation 1 and the same measures are applied for assessing the trained maps.

Results. Table 2.5 shows performance measures for F_0 and D1 in Simulation 2. The categorical error (i.e., the percentage of ambiguous units) and classification error (percentage of tokens landing on ambiguous units) are high for F_0, but the performance of D1 is much more successful. Although the perfect performance of D1 may be partially related to an

artifact of the performance measures, we conducted the same analysis for Simulation 2 as for Simulation 1 to maintain consistency across simulations.

Table 2.5 Performance measures for F_0 and D1 conditions in Simulation 2.

	Categorization error	Classification error
F_0	1.00	1.00
D1	0.00	0.00

Figure 2.5 shows phoneme maps of F_0 and D1 after training. As can be seen, the categorization with the F_0 input is poor, with all four units responding confusingly to multiple categories. The D1 input, on the other hand, resulted in four units representing four distinct tones, suggesting successful abstraction of four tonal categories from the well-delineated clusters and neighborhoods developed during the previous training using 100 units.

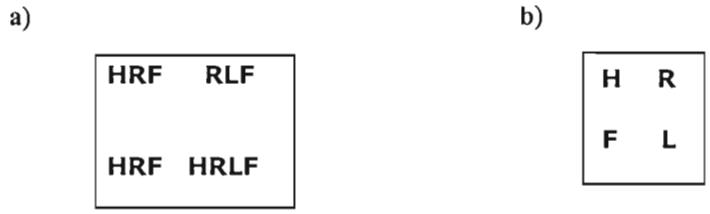


Figure 2.5 Phoneme maps of **a)** F_0 and **b)** D1 after training in Simulation 2. The phoneme map of F_0 shows four ambiguous units while the phoneme map of D1 shows four categorized units.

2.2.5 General discussion

At the outset of the study we raised three questions: (a) Is it possible for a perceptual system to derive phonetic categories directly from continuous signal input without extraction of abstract features? (b) Is the information about movement toward underlying tonal targets as represented by D1 sufficient for the categorization of the four Mandarin tones produced in connected speech by multiple speakers? (c) Is there a possible link between the decline in sensitivity to within-category differences and the core mechanisms of learning phonetic categories?

To answer the first question, we used the self-organizing map method that is biologically plausible because it is based on the discovery that there are topographically organized projections from the periphery to cortical areas in the brain (Kohonen, 1982). The biological plausibility of our methodology is further enhanced by the naturalness of the input that we used. In all the simulations, time-varying continuous trajectories were used directly as input, with no further extraction of more abstract features. The only preprocessing involved are (a) the extraction of continuous F_0 trajectories from the acoustic signal, (b) conversion of F_0 trajectories to continuous velocity profiles, and (c) division of the trajectories into syllable-sized chunks. Both (a) and (b) are highly plausible in human perception, given what is known about pitch perception and velocity processing and representation in human brains (Gandour, 1983, 2000; Seldon, 1985). The division of continuous profiles into syllable-sized chunks, i.e., (c), seems reasonable on a number of considerations. The syllable appears intuitively salient, as evidenced by the facts that many, including the Chinese writing systems represent speech directly at the level of the syllable (Chao, 1968; DeFrancis, 1984), and that linguistic theories typically treat the syllable as a level of representation (e.g., Chomsky & Halle, 1968; Prince & Smolensky, 1993). More importantly, experimental research on speech perception has shown that the syllable is the perceptual units in very young infants (Bertoncini & Mehler, 1981; Bijeljac-Babic et al., 1993; Jusczyk & Derrah, 1987). There has also been accumulating evidence for the syllable as a critical unit in speech production as summarized in recent theories of the syllable (Fujimura, 2000; Krakow, 1999; MacNeilage, 1998; Xu & Liu, 2006). It is therefore plausible to assume that some kind of division of the acoustic signal into syllable-sized chunks occurs in the brain during learning in infants and processing in adults.

As shown by the results of simulation 1, even the lowest performance, obtained with F_0 as input, achieved 80% correct categorization. Thus the answer to the first question is positive: It is indeed possible for a biologically plausible perceptual system to learn at least one type of sound categories directly from continuous signal input without extraction of abstract features. The implication is that phonological features such as those proposed by Jacobson, Fant and Halle (1967) and Chomsky and Halle (1968), while seemingly appealing to us as scientific observers, may not be the elements actually processed by the brain.

What is more likely to be processed, as suggested by the even better performance of D1 in Simulation 1 (97% correct categorization), is the actual movements themselves. And such high performance has provided a clear positive answer to the second question, namely, the information about movement toward underlying tonal targets as represented by D1 is sufficient for the categorization of the four Mandarin tones produced in connected speech by multiple speakers. To better understand why the D1 is so much more effective than F_0, we plotted in Figure 2.6 the internal maps, which show the prototypical F_0 and D1 profiles developed during training in Simulation 1 for each tone category. Two general patterns can be observed. First, the F_0 profiles show much larger within-category vertical spread than D1 profiles, and the spread is especially wide near the syllable onset. Second, the F_0 profiles show much less distinct movement patterns than D1 profiles. In fact, with the only exception of Low, F_0 profiles of each tone move in both overall directions: up and down. The D1 profiles, in contrast, display high consistency in terms of the overall direction of movement. And they differ within each tonal category mostly in magnitude of the movement.

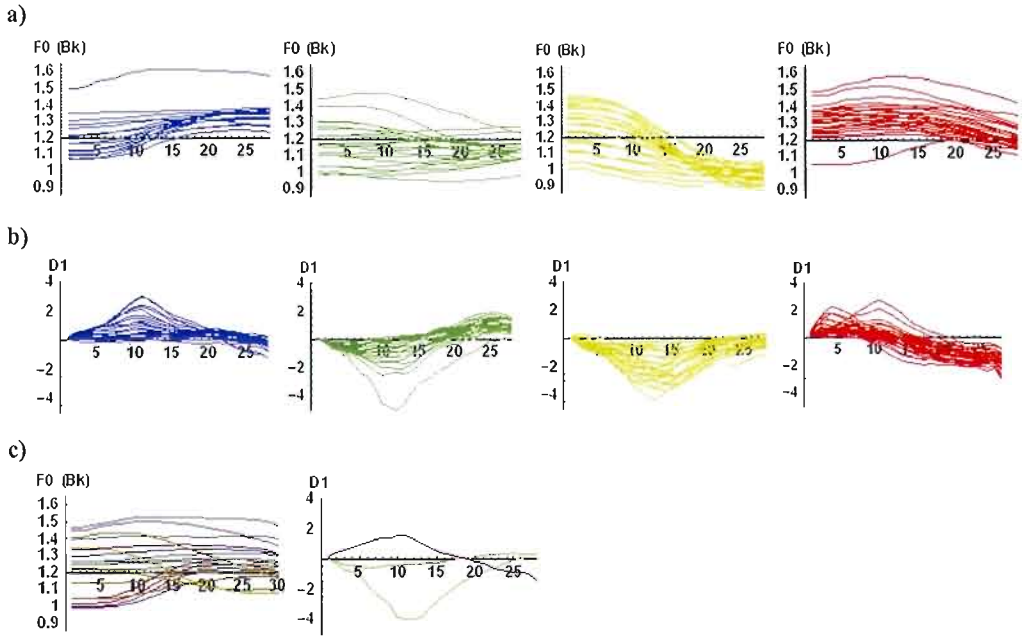


Figure 2.6 Internal maps of the four Tones in **a)** F_0 and **b)** D1 categorized units and **c)** F_0 and D1 ambiguous units.

The consistent D1 profiles seem to directly reflect the nature of the F_0 movements as characterized by the Target Approximation model (Xu & Wang, 2001) and by the velocity profiles of movements proposed by Nelson (1983). Considering the static tones, most of the High profiles increase their speed from 0 toward positive values, reaching peak velocity around the center of the syllable and finally slowing down toward the initial speed of 0 near the end of the syllable. The Low profiles show almost mirror images of the High profiles. Such unimodal velocity profiles fit the definition of a simple movement given by Nelson (1983), i.e., one that starts from one position and stops at another. A voluntary movement such as reaching satisfies this definition. It follows then that the movements involved in the High and Low tones are those toward a static F_0 height. The positive velocity profiles during High correspond to movements toward an above-average pitch height, and the negative velocity profiles during Low correspond to movements toward a below-average pitch height.

The D1 profiles of the dynamic tones present a different picture. Like the static tones, the D1 profiles of Rise and Fall both increase their speed from 0 at syllable onset toward a negative/positive value. But instead of continuing with the initial direction, the D1 profiles reverse their directions, cross the zero speed line and continue until a high velocity is reached near the end of the syllable. In other words, the Rise/Fall velocity profiles indicate rapid initial F_0 movement toward a relatively low/high F_0, followed by another movement in the opposite direction toward the zero line, thus indicating a movement toward an initial static height per Nelson's (1983) definition. But the movements afterwards no longer fit Nelson's definition. Rather, the fact that D1 reaches a high (positive or negative) value near the end of the syllable suggests that the high velocity itself is the final goal of Rise and Fall. In other words, the targets of these tones are dynamic, i.e., with a simple linear function as their goal, as postulated in the Target Approximation model (Xu & Wang, 2001).

Based on the above understanding, the prototypes developed in Simulation 1 actually contain three apparently inappropriate ones. One in Rise (with a very low valley) that should belong to the Low tone, and two in the Fall tone (with a very low valley) that should belong to the High tone. Indeed, the further categorized D1 profiles developed in Simulation 2 (Figure 2.7) seem to have fully eliminated those deviant prototypes.

The direct characterization of articulatory movement toward underlying tonal targets is not the only benefit of D1 profile as input to a learning system. It also has the benefit, as explained in the Background, of immediately removing most of the individual differences in terms of their idiosyncratic pitch ranges as well as much of the influence of the preceding tone. Variability due to both of these sources can be clearly seen in the large vertical distribution of initial F_0 in the upper row of Figure 2.6. Such variability is virtually absent in

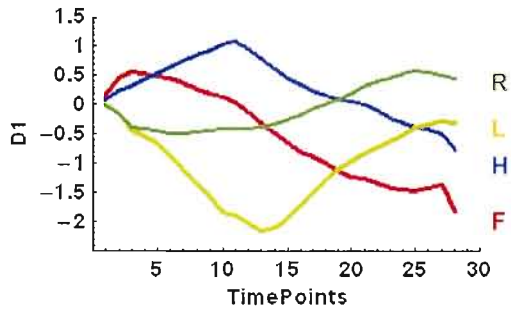


Figure 2.7 Velocity profiles for the four units after Simulation 2.

the D1 profiles in the lower row of the figure. As explained in the Background, the differentiation process eliminates from a function the term specifying its y-intercept, thus removing most of the speaker-related variability and much of the context-related variability.

The results of Simulation 1 suggest that with sufficient number of receptive units in a neural network, variability in D1 profile can be handled by developing topographical clusters with well-defined borders, each corresponding to a tonal category. While the formation of clusters in Simulation 1 may seem to support the hypothesis of the exemplar theory that linguistic categories can be represented by probabilistic exemplar clouds (e.g., Pierrehumbert, 2002), the results of Simulation 2 suggest that exemplar clouds do not need to be always maintained. That is, a further learning step can take place in which the learned profile clusters can be further reduced to even more ideal prototypes with one-to-one correspondence to the tonal categories. Such direct representations achieved through a two-stage learning process suggest the possibility of drastically reducing the number of neurons needed to represent phonetic categories in the later stage.

An implication of such more economical representation is that it could naturally lead to the behavior patterns described by speech perception theories such as the Functional Reorganization Model (Werker & Tees, 1984), the Native Language Magnet Theory (Kuhl, 1991) and the Perceptual Assimilation Model (Best, 1995), namely, as children mature, their sensitivity to sounds of foreign languages is reduced, unless those sounds bear little resemblance to any sound in the children's native language. Thus, the results of Simulation 2 suggest a positive answer to the third question of the present study, i.e., there is indeed a possible link between the decline in sensitivity to within-category differences and the core mechanisms of learning phonetic categories. It is conceivable that the increased efficiency in a native language as a child matures is related to the precedence given to the mapping of incoming speech sounds to the prototypes formed in the second stage of learning. Such precedence would bias the incoming foreign sounds toward the closest prototype whenever possible. But the precedence can be also softened if a new round of cluster formation is made to happen, as during extensive perceptual training (Bradlow et al., 1997) or during any focused second language learning.

Perhaps the biggest surprise to us was the success of the tonal category formation without direct assistance of any contextual information. This is because we have seen in previous research that the availability of contextual information to the listener is critical for recognizing tones severely distorted by tonal contexts (Xu, 1994). We note, however, a major difference between the input data used in the present study (from Xu, 1997) and those used in Xu (1994). That is, the target tones in the latter were produced in the middle syllable of trisyllabic words, which is known to be a prosodically weak position (Chao, 1968; Shih, 1993), leading to much heavier contextual distortion than in the data used in the present study. Furthermore, the tonal information Xu (1994) was further degraded by the voiceless initial consonants in the target syllable, which both hide and perturb the F_0 contours that were already quite short in duration due to the prosodically weak position (Xu, Xu, & Sun, 2003). What the findings of the present study demonstrate is that as long as the contextual distortion is not too severe and sufficient amount of F_0 movements toward the tonal target is available in the input (i.e., not hidden by voiceless consonants), a learning system can successfully derive the tonal categories without directly processing contextual information. It would be

interesting in future investigations, of course, to explore how direct processing of contextual information can be used for helping the recognition of speech sounds that have been more severely distorted by contexts.

Finally, the present findings also have implications for another long-standing debate over the nature of speech perception, i.e., whether it is the auditory patterns (e.g., Diehl & Kluender, 1989) or articulatory gestures (Liberman & Mattingly, 1985) that are the distal objects of speech perception. While the auditory accounts may have difficulty explaining how variability with apparent articulatory sources can be effectively processed without referring to the articulatory movements, the motor theory may have difficulty explaining how infants who cannot yet speak can develop perceptual phonological categories that are articulatory in nature. The learning simulations in the present study suggest that by tracking the velocity profile of an acoustic parameter that closely reflects the underlying articulatory movements, variability due to both individual difference and contextual variations can be drastically reduced. And, the remaining variability, being articulatorily lawful, can be effectively handled by a neural network through unsupervised learning. This finding is reminiscent of the direct realist view of speech perception (Fowler, 1986) which postulates that the objects of speech perception are articulatory gestures as opposed to auditory properties in the form of distinctive features. The direct realist view also postulates that speech perception is done by tracking articulatory movements. As we have seen, the learned prototypical velocity profiles in the present study directly reflect movements toward underlying targets that are either static or dynamic in terms of both acoustic patterns and articulatory states. It is therefore conceivable that a further learning step for the infants is to derive those targets from categorized velocity profiles. Once stored in the brain, infants may then use those targets as articulatory goals when they babble and learn to speak themselves. This understanding therefore allows the possibility that speech production and perception are closely linked to each other but not necessarily always in lockstep.

2.2.6 Concluding remarks

Given that the speech input to infants is highly variable, and that infants are not typically told what the meaningful sounds are in a language, one of the greatest puzzles about human speech is how an infant can discover the sound categories of the ambient language from adult

input. In the present study, we investigated the possibility that infants can derive phonetic categories directly from the time-varying acoustic signals produced by adults without having to extract abstract features from the signal. To this end, we explored the hypothesis of the Target Approximation model of tone production (Xu & Wang, 2001) that the consistency of lexical tones produced in connected speech in a language like Mandarin lies in the continuous articulatory movement toward the underlying targets of the tones, as reflected in the F_0 trajectories during the syllable. We also explored the possibility that the velocity profiles (D1) represent more directly (than F_0) articulatory movements toward the underlying targets of the lexical tones, and as such they can significantly reduce the amount of variability due to speaker difference and tonal context. We tested these possibilities with a self-organizing topographical neural network using syllable-sized F_0 and D1 profiles as input. Although the debate persists in the field of language acquisition about the role of feedback during learning, our simulations demonstrate that guided feedback is not needed for the learning system to successfully derive the tonal categories. Testing results showed that while F_0 gave reasonably good performance, the prototypical D1 profile clusters developed through training yielded virtually perfect tone recognition without the help of any contextual information or pre-abstracted features. Further simulation showed that the learned D1 clusters, through additional learning, can be developed into even more ideal prototypes that have one-to-tone correspondence to the tones. These findings not only point to a possible way via which infants can develop phonetic categories through unsupervised learning, but also may lead to answers various theoretical questions about language acquisition, speech perception and speech production.

Acknowledgment

Part of the results of the study was reported at the 149th meeting of the Acoustical Society of America, 2005 and the ISCA Workshop on Plasticity in Speech Perception, 2005. We thank the support of a FCAR (FQRSC) scholarship to the first author, the funding from SSHRC, NSERC and FQRSC to the second author, and the support from NIH Grant DC006243 to the third author.

2.2.7 Appendix I – The SOM algorithm

Architecture. The SOM maps a high-dimensional input space onto a discrete lower dimensional array of topologically ordered processing units. A 1-dimensional SOM is illustrated in Figure 2.8 (adapted from Ritter & Schulten, 1986). The input and output layers are fully interconnected to each other. Output space N is a lattice on which units are labeled by a position vector r indicating their physical position on that lattice (filled dots on the vertical line). Input space X is mapped on output space N by a set of adaptive receptive field centers, or connection weights $w_r \in X$ (empty dots on horizontal lines) for which corresponds a typical $x_i \in X$. The subset of X closer to a unit's receptive field center than to any other w_r constitutes the receptive field of that unit (vertical bold lines). In the present study, a two-dimensional map of 10 x 10 units is used.

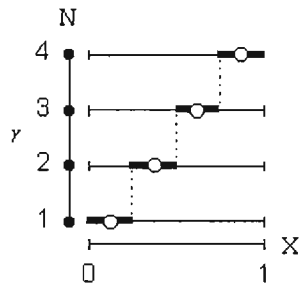


Figure 2.8 Architecture of a one-dimensional SOM: linear array N of 4 output units r (filled dots), their receptive field center (empty dots) and receptive field (bold horizontal lines) for input space X = [0,1] (adapted from Ritter & Schulten, 1986).

Transmission rule. The transfer function of the network contains two steps. First, the distance between the receptive field center of each unit to that of input vector x_i is evaluated according to:

$$u_r = (\Sigma (x_i - w_r)^2)^{1/2} \qquad (2.7)$$

where u_r represents the net value of each unit r. The unit with the shortest Euclidean distance between w_r to reference input pattern x_i is selected to be the winner according to:

$$v = \min (u_r) \qquad (2.8)$$

where v corresponds to the position of the winner, or Best-Matching-Unit (BMU). The net value is further transformed to yield the final response, given by the non-linear Gaussian function:

$$\eta_r = \text{Exp}(- ((v - r)^2 / \sigma)) \quad (2.9)$$

where η_r corresponds to each unit's activation. The Gaussian is peaked at v so the winning unit is the most activated ($\eta_v = 1$). Units falling into neighborhood radius σ get activated by means of afferent lateral activity, although to a lesser degree that depends on their position relative to the winner. The transmission rule can be conceived as a basic perceptual discriminative function that computes the distance between a perceived signal and a signal stored in a list of prototypes.

Learning rule. The SOM implements a regression algorithm for mapping an input distribution $P(x)$, $x_i \in X$, onto the output space. The lateral connections between output space nodes allow for topological ordering to be preserved in the map during the learning period. Receptive field centers w_r are adapted during a stochastic learning procedure in which a random sequence of data points x_i is presented repeatedly for a predefined number of times. Each time an input vector is presented, the winning unit and its neighbors shift their receptive field center toward the data point according to:

$$\Delta w_r = \alpha \, . \, \eta_r (x_i - w_r) \quad (2.10)$$

where η_r is the value outputed by the transmission rule and α is the learning step size. The weight matrix is then updated according to:

$$w_r (t+1) = w_r (t) + \Delta w_r \quad (2.11).$$

The learning rule can be conceived as a basic perceptual learning function that transforms the internal organization to reflect the environment characteristics.

Initialization of the map. The weight matrix is initialized as follows. Each receptive field center is set to correspond to a linear trajectory of the form ax + b, where the slope a = 0.00009 and the intercept b ranges from 0.03 to 1. This yields a map in which the minimum

and maximum values correspond to 0.03 and 1.99, which covers the input space (ranging from 0.54 to 1.95) without supposing a predefined number of categories. Other types of initialization schemes have been tried and made no difference in the final results, since whether the weights are bigger or smaller, they will respectively shrink or expand with the learning process to fit the input distribution.

3 L'ACQUISITION DES TONS LEXICAUX À PARTIR DU SIGNAL CONTINU DE LA PAROLE PAR LE BIAIS D'UN MODÈLE CONNEXIONNISTE NON SUPERVISÉ

3.1 Résumé de la publication en français

Les enfants acquièrent les catégories phonétiques de leur langue maternelle par simple exposition à la parole adulte. Toutefois, la façon dont ils réussissent à faire face à la variabilité inhérente au signal de la parole demeure inconnue, tout comme la façon dont ils parviennent à traiter de multiples fonctions linguistiques transmises simultanément et partageant la même dimension acoustique. Cette étude vise à modéliser l'acquisition des tons lexicaux chinois mandarin à l'aide de réseaux neuronaux artificiels de type non supervisés. Des cartes topographiques auto-organisées (Kohonen, 1989, 1995) ont été entraînées avec un signal continu de la parole lors de quatre simulations impliquant un degré croissant de variabilité. La première simulation impliquait le plus faible degré de variabilité, avec un corpus de 1 800 exemplaires de tons produits par trois locuteurs dans divers contextes tonaux. La simulation impliquant le plus haut degré de variabilité présentait aux réseaux 11 520 tons produits par quatre locuteurs et quatre locutrices dans tous les contextes tonaux et diverses conditions de focus prosodique. Dans le but de comparer la performance de l'information acoustique de surface et l'information dynamique sous-jacente dans la catégorisation des tons, chaque simulation présentait à un réseau les patrons de fréquence fondamentale (F0) des tons et à un autre réseau leurs profils de vélocité, c.-à-d. les premières dérivées de F0 (D1). Au total, la performance des réseaux entraînés avec F0 ou D1 se situe au delà du niveau de chance. Toutefois, face à un degré élevé de variabilité, seuls les réseaux entraînés avec D1 offrent une performance similaire à celle de l'adulte. Ces résultats montrent que, malgré l'importante variabilité induite par le contexte, les multiples locuteurs et la présence de fonctions linguistiques compétitives (ici les tons et le focus), un simple mécanisme d'apprentissage inductif peut extraire les catégories tonales directement à partir du signal continu de la parole, sans l'aide d'une supervision externe ou de rétroaction. Cette étude suggère également que le signal continu de la parole contient suffisamment d'information catégorielle et que l'information acoustique dynamique peut être utilisée pour résoudre le problème de la variabilité.

3.2 Simulating the acquisition of lexical tones from continuous dynamic input

3.2.1 Abstract

Infants develop phonetic categories by simply being exposed to adult speech. It is not known, however, whether this is achieved by directly processing the speech signal, or by filtering the signal through a set of innate distinctive features that define phonological contrasts of all languages. In a neural network simulation of phoneme acquisition, robust tonal categorization is achieved by topographically tracking continuous fundamental frequency contours and their velocity profiles. This result suggests that continuous speech signal carries sufficient categorical information that can be directly processed, and that innate distinctive features are not needed for discovering phonetic categories from adult speech.

3.2.2 Introduction

Before reaching one year of age, human infants have developed the ability to process speech sounds specific to their native language (Werker & Tees, 1984). This predates their understanding of word pairs containing minimal contrasts (e.g., *bear-pear*) (Caselli et al., 1995), and thus cannot be attributed to lexical pressure. A long-standing debate regarding this development concerns the knowledge infants are born with. According to the nativist hypothesis, infants are endowed with innate phonetic detectors in the form of distinctive features, e.g., a set of articulatory-acoustic properties with binary values (Jakobson et al., 1967), and they learn the sounds of their native language by setting the values of the distinctive features based on the speech input. An empiricist view is that infants are born with general auditory mechanisms to process all speech sounds of human languages (Aslin, Werker, & Morgan, 2002), and that later they develop the ability to analyze the distributional properties of the input speech, which enable them to establish native phonetic categories (Maye et al., 2002).

No empirical research has explicitly tested whether distinctive features are necessary as a prerequisite for learning phonetic categories. Here we investigate with self-organizing neural map simulations (Kohonen, 1995) whether phonetic categories can be discovered through unsupervised learning by directly processing continuous speech signals, with no mediation of distinctive features. Our experiments simulate the acquisition of lexical tones in

Mandarin Chinese, a language that uses tones to distinguish words that are otherwise identical in their vowel and consonant composition. Like segments, tones have been assumed to be represented by innate distinctive features (Goldsmith, 1990). Because tones typically involve a single primary acoustic dimension, namely, F_0, they are ideal for testing hypotheses that involve detailed mechanisms of phonetic acquisition. Given that infants receive no instructions from adults, unsupervised networks are powerful for revealing mechanisms of language learning (Shi et al., 1998).

Mandarin has four tones: High, Rise, Low and Fall, which are carried by the fundamental frequency (F_0) of the vocal fold vibration. The F_0 values of a tone, however, vary extensively due to at least three sources (Figure 3.1). First, cross-speaker variability arises from differences such as age, gender and idiosyncrasies (Hillenbrand et al., 1995; Peterson & Barney, 1952). Second, contextual variability arises from neighbouring sounds affecting one another (Ohman, 1966). In particular, the F_0 pattern of any tone depends much on that of the preceding tone (Xu, 1997). Infant-directed speech consists primarily of multiword utterances (Shi et al., 1998), leading to considerable contextual variability. Finally, variability in F_0 comes from its use to not only distinguish words, but also to encode information such as focus, which can introduce F_0 variations with magnitudes similar to those of tones (Xu, 1999).

Despite the extensive variability, tonal perception appears early in infancy (Mattock, 2004). A key to infants' strategies may lie in the understanding of the mechanism of contextual variability. That is, the patterns of variations show that the talkers' articulatory strategy remains the same: to approach a constant tonal target starting from the initial state due to the preceding tone (Xu & Wang, 2001). Such a strategy would result in velocity profiles that directly reflect the nature of the tonal targets, as they have been shown to reveal the dynamics of skilled actions such as jaw movements during speech (Nelson, 1983). Furthermore, taking the derivative of a curve results in the removal of all its constant term(s), thus eliminating any overall height differences such as those due to cross-speaker variability. We thus tested the hypotheses that (a) tonal categories can be discovered by processing continuous pitch movement patterns without the mediation of distinctive features, and that (b)

the velocity of F_0 (i.e., the first derivatives of F_0 patterns, henceforth D_1), better reveals the invariant properties of tones than F_0 itself.

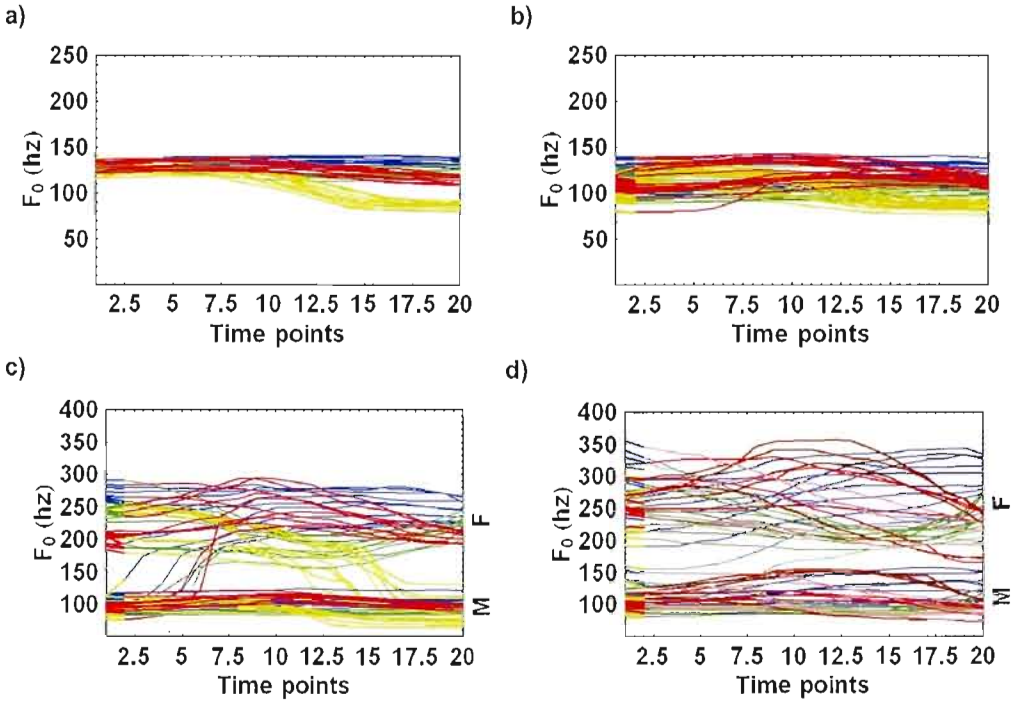


Figure 3.1 Variability in tones. **a)** F_0 (in hertz) of 40 repetitions of the four Mandarin tones (**High**, **Rise**, **Low**, **Fall**) by one male speaker with identical preceding tone (High) and identical focal status (neutral). **b)** Contextual variability: F_0 of 80 repetitions of the four tones by one male speaker with different preceding tones (specified by syllable onset color) and identical focal status (neutral). **c)** Contextual and speaker variability: F_0 of 80 repetitions of the four tones by one female and one male speakers with different preceding tones and identical focal status. **d)** Contextual, speaker and focal variability: F_0 of 80 repetitions of the four tones by one female and one male speakers with different preceding tones and variable narrow focus (dark=on-focus, medium=neutral focus, pale=post-focus).

3.2.3 Method

The neural networks were presented with learning material of increasing variability to compare the performance of F_0 and D_1 in tonal categorization. The input corpuses for training and testing contained a large number of exemplars of the four Mandarin tones produced in connected utterances (data from Xu, 1997, 1999). In Simulation 1, the input corpus contained 1800 exemplars produced by three adult male native Mandarin speakers. Each stimulus

corresponded to the first or second syllable of disyllabic word 'mama' produced with varying tones in the middle of a carrier sentence which had either high or low pre-target F0 offset and post-target F0 onset. In Simulations 2, 3 and 4, the input corpora respectively contained 1440, 5760 and 11 520 tones from 3840 declarative sentences produced by four adult male and four adult female native Mandarin speakers. Sentences were formed of a subject, verb and object and contained five syllables, each word corresponding to one or two consonant-vowel (CV) syllable(s), where C was a sonorant (/m,n/), except when the Low tone occurred on the fourth syllable, where C corresponded to /d/. The subject and object words were disyllabic and the verb was monosyllabic. The sentences were produced in various focus conditions: (a) neutral focus, (b) focus on word 1, (c) focus on word 2, and (d) focus on word 3 (e.g., 'maomi mo maomi' (kitty touches kitty)) in response to the following wh-questions: 'What is Kitty doing?', 'Who is stroking Kitty?', 'What is Kitty doing to Kitty?', 'What is Kitty stroking?'). Since the tone on the first and last syllables was kept constant to High, these syllables were removed from the input corpus. The second, third and fourth syllables contained varying tones (H, R, L, F on the 2nd syllable, H, R, F on the 3rd syllable and H, L on the 4th syllable).

Input tokens were multidimensional vectors composed of equal-distanced discrete values from syllable-sized F_0 curves. F_0 input vectors were first transformed from hertz scale to semitone scale according to:

$$F_0st = 12 \log_2 (F_0hz) \qquad (3.1).$$

The velocity profiles of F_0 were generated according to:

$$D_{1i} = (F_0st_{i+1} - F_0st_i) / (T_{i+1} - T_i) \qquad (3.2),$$

where T represents time, which yields the discrete first derivatives of F_0.

The neural maps were squared arrays of processing units, the number of which was determined as a function of input corpus size. The maps contained 144 units (12 x 12) in Simulations 1 and 2, 400 (20 x 20) units in Simulation 3, and 900 (30 x 30) units in Simulation 4. Initial connection weights were set to correspond to linear trajectories of the form $ax + b$, where minimum and maximum weight values fell within the range of the input space, without supposing a predefined number of categories. During training, input tokens

were randomly presented to the networks. The learning step size decreased linearly from 0.7 to 0.01 and the neighbourhood function included all units at training onset, decreased exponentially, and only included the best matching unit (BMU) at the end of training (i.e., the closest unit in terms of Euclidean distance). The testing phase presented the training corpus and a new set of input tokens to verify the networks' capacity to generalize to novel data. During testing, each input token was assigned to the BMU. Units that responded to a single category at least 68% of the time during testing were labelled according to that category. The performance was assessed in terms of percentage of correctly classified input tokens

To better understand the outcome of the categorization process, quantitative coloring of the neural maps was obtained by associating the distinct tonal categories with distinct colors produced with the CMYK color system. The High tone, represented by blue, was specified by a mix of cyan and magenta in the vector [1,1,0,0] (Rise = green [1,0,1,0]; Low = yellow [0,0,1,0]; Fall = red [0,1,1,0]). Each map unit was then associated to a four-dimensional vector, the values of which were specified according to the unit firing probabilities for each tonal class during testing (the fourth element K (black) was kept null). In consequence, units responding to a single tone are represented by a saturated color while units sensitive to multiple tones are represented by 'impure' colors. Learned categories are thus shown on the maps as regions of distinct colors.

3.2.4 Results

Overall, the networks trained with either F_0 or D_1 yielded above chance level performance, but those trained with D_1 yielded better performance than those trained with F_0 (Figure 3.2a). In Simulation 1 the input exemplars had been produced in different tonal contexts by three male speakers. Reasonably high overall rate of success was achieved with F_0 (mean = 0.84, standard deviation = 0.09) but D_1 profiles yielded almost perfect categorization (m. = 0.93, s.d. = 0.03), replicating the results we obtained in a previous study (Gauthier, Shi, & Xu, 2007a). In Simulation 2 input exemplars were produced by a single speaker in different tonal contexts with variable narrow focus (i.e., on-focus, neutral focus or post-focus). Relative to Simulation 1, the overall rate of success decreased for F_0 (m. = 0.72, s.d. = 0.11) and for D_1 (m. = 0.84, s.d. = 0.03), suggesting that focus-induced variability is more detrimental than cross-speaker (within-gender) variability.

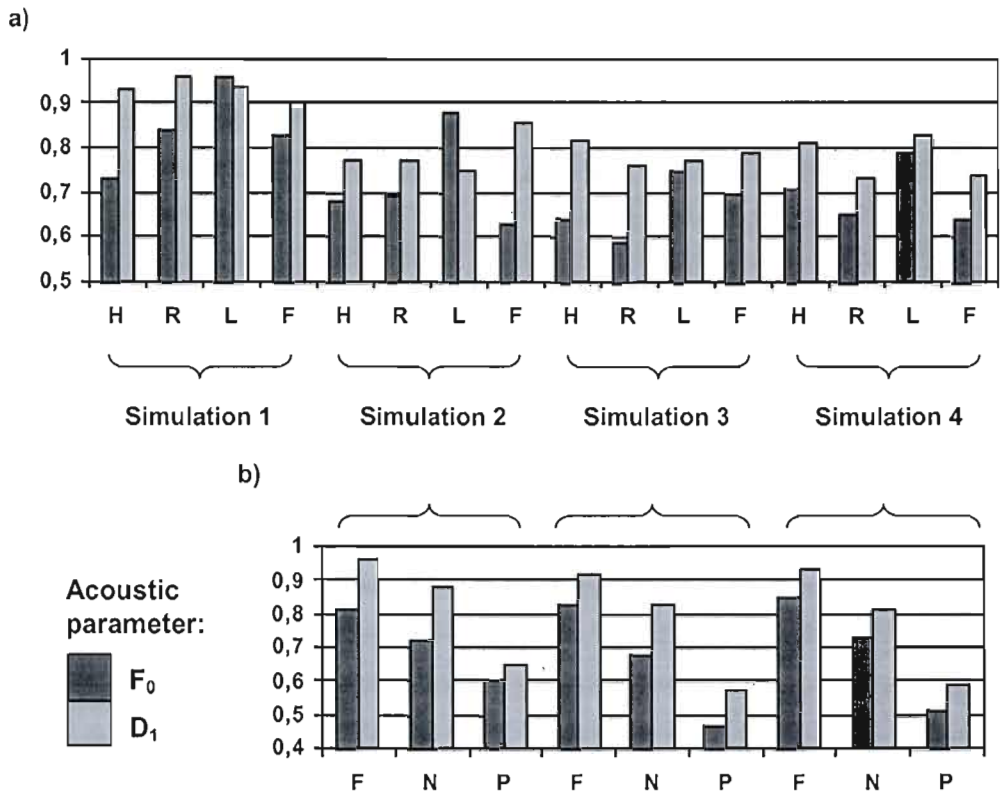


Figure 3.2 Tonal classification rate of success. **a)** Training and testing were done with data produced in different tonal contexts by four male speakers (Simulation 1), different tonal contexts with variable focus by one female speaker (Simulation 2), different tonal contexts with variable focus by four male speakers (Simulation 3), and different tonal contexts with variable focus by four male and four female speakers (Simulation 4). **b)** Results of Simulations 2, 3 and 4 expressed as a function of focal status.

Simulation 3 tested the combined impact of the aforementioned sources of variations. Input exemplars were produced by four male speakers in different tonal contexts with variable focus conditions. Although the performance of both networks declined, D_1 (m. = 0.79, s.d. = 0.03) still performed better than F_0 (m. = 0.67, s.d. = 0.07). Finally, Simulation 4 involved the highest amount of variability, with input exemplars produced by four male and four female speakers in different tonal contexts with variable focus conditions. Both F_0 (m. = 0.70, s.d. = 0.07) and D_1 (m. = 0.78, s.d. = 0.05) showed a similar decline in performance

relative to Simulation 1. But most of the errors in the simulation involved post-focus elements (Figure 3.2b). This is consistent with the fact that the pitch range of post-focus elements are severely suppressed as part of the encoding pattern of focus (Xu, 1999).

Tonal color maps for F_0 and D_1 networks of Simulation 4 were constructed by associating the four tonal categories with four distinct colors (Figure 3.3a). In Figure 3.3b, the F_0 color map exhibits no clear tonal organization. Some tonal categories are distributed to multiple clusters. Also, activations of units during testing are uneven across the map: some map units responded to many input tokens, while others to few or none at all (the larger a unit, the greater its firing probability). In contrast, the D_1 color map (Figure 3.3c) shows four well-separated regions – one corresponding to each tone. Activations of units are even: each responding to a comparable number of test tokens. D_1 is thus much more powerful than F_0 for normalizing and categorizing the Mandarin tone system.

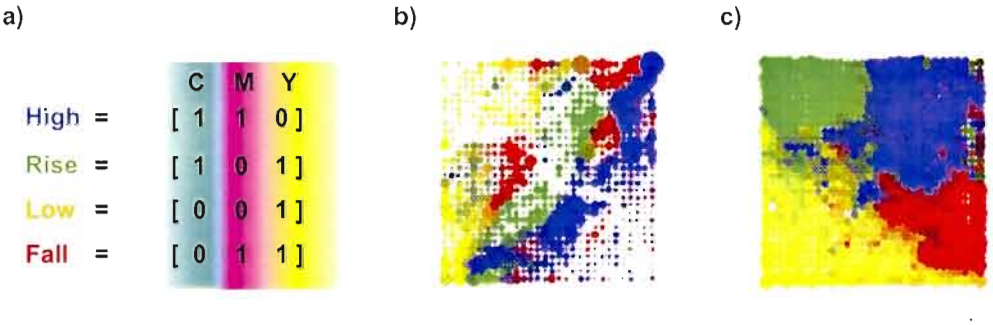


Figure 3.3 Color maps of trained networks in Simulation 4. **a)** Color legend for interpreting the color maps. Color maps for **b)** F_0 and **c)** D_1.

3.2.5 Discussion

In summary, our results show that a simple learning mechanism suffices for extracting tonal categories directly from the acoustic input, assuming that F_0 contours have already been segmented into syllable-sized chunks by a separate mechanism (Bertoncini & Mehler). This finding demonstrates that, while it is possible that abstract representations in terms of distinctive features may emerge at a later stage of learning, initial acquisition of phonetic categories can be achieved without the mediation of distinctive features.

In addition, our results show that velocity as input for phonetic categorization is more robust than F_0 in factoring out the variability in the input signal and revealing the invariant underlying phonetic targets. Furthermore, because velocity profiles capture the similarities between adult and infant speech production, it is plausible that perceptual representations of phonetic categories based on such dynamic information may be used to guide infants' own production of these categories. If that is the case, the velocity patterns developed through perceptual acquisition could form a direct link between perception and production, the two key components of human speech communication.

Acknowledgment

This work was supported by funding from SSHRC, NSERC and FQRSC to the second author and supported in part by a NIH Grant to the third author.

4 LA PERCEPTION DU MOUVEMENT ACOUSTIQUE CHEZ L'ENFANT PRÉVERBAL

4.1 Résumé de la publication en français

De récentes études de modélisation par le biais de réseaux neuronaux artificiels suggèrent que l'acquisition des catégories phonétiques et la normalisation du signal de la parole chez l'enfant peut reposer sur le mouvement acoustique reflétant les gestes articulatoires. La présente étude explore la sensibilité d'enfants préverbaux envers l'information dynamique du signal de la parole relative aux mouvements articulatoires. La limite de vitesse imposée par les contraintes de l'appareil vocal sur les modulations de la fréquence fondamentale a été utilisée afin de tester la capacité d'enfants âgés de 4 et 8 mois à percevoir le profil de vélocité de contours d'intonation à l'aide d'une procédure de préférence visuelle. L'hypothèse voulait que les enfants puissent discriminer entre des sons possibles et impossibles sur le plan articulatoire et impliquant une différence subtile dans la vitesse de la fréquence fondamentale. De plus, puisque l'enfant est exposé dans son environnement aux sons de la parole qui reflètent les limites de l'appareil vocal sur la phonation, une préférence était prédite pour les stimuli respectant la contrainte de vitesse maximale de changement de fréquence fondamentale relativement aux stimuli qui violent cette contrainte. Enfin, deux groupes d'âge ont été testés afin d'observer l'évolution de la sensibilité pour le mouvement acoustique et vérifier si les enfants plus jeunes possèdent cette sensibilité, ou si au contraire ils requièrent une plus longue période d'exposition à la parole adulte. Les deux prédictions ont été vérifiées chez les deux groupes d'âge, suggérant que la poursuite auditive du mouvement acoustique reflétant les gestes articulatoires constitue un mécanisme utile pouvant supporter l'acquisition des catégories phonétiques.

4.2 The perception of acoustic movement in preverbal infants

4.2.1 Abstract

Recent neural network simulations have shown that phonetic categories can be learned from dynamic acoustic information reflecting motor speech movement (Gauthier et al., 2007a; Gauthier, Shi, & Xu, 2007b). The present study examined if preverbal infants are sensitive to

acoustical movement. The speed limit imposed by articulatory constraints on pitch modulation (Xu & Sun, 2002) was used for testing four- and eight-month-olds' ability to perceive velocity profiles of intonation patterns. The results show that infants are sensitive to the maximum speed of pitch change of an unknown speaker, indicating that they can use pitch velocity to normalize speech input variability. This suggests that early speech perception involves tracking continuous dynamic acoustic patterns that reflect articulatory movement. This perceptual ability may eventually lead to phonetic category learning.

4.2.2 Introduction

Infants possess acute abilities to perceive speech from the earliest stage of language development. They not only can categorize speech sounds (Eimas et al., 1971), but also preserve the sensitivity to fine-grained within-category acoustic details (McMurray & Aslin, 2005). During the second half of the first year of life, phonetic perception begins to be shaped towards the specific sound patterns of their ambient language (Mattock & Burnham, 2006; Polka & Werker, 1994; Werker & Tees, 1984). Determining the exact mechanism underlying initial phonetic categorization is one of the main research endeavours of the field in recent years.

Perceptual studies with infants, however, largely focus on testing one phonetic parameter at a time, using stimuli constructed to be free of other factors. Little work has been devoted to infants' ability to perceive speech sounds that are subject to different sources of variability. Natural speech input is highly variable due to factors such as speaker differences, phonetic contextual differences, etc. Such variability leads to substantial overlap of phonetic categories. For example, first and second formant frequencies that are classically considered as cues to vowels exhibit significant overlap between vowel categories due to speaker variations (Peterson & Barney, 1952).

Adults have no difficulty solving the variability problem during language comprehension. Do naïve language learners do this task? Previous work showed that infants as young as two months of age have the ability to filter out variation due to different speakers when perceiving target syllables (Jusczyk et al., 1992). Six-month-olds can perceive vowel categories despite speaker and pitch variation (Kuhl, 1979), and two- to four-month-olds can

categorize stop versus semivowel consonants based on formant transition cues even when these cues were much affected by the factor of speech rate (Eimas & Miller, 1980). These studies suggest that infants do perform normalization tasks. The mechanism underlying the normalization process in infants has, however, received little attention. The present study aims at understanding the possible computations in infants' ability to handle the variability problem.

Theories of adults' strategies to variability may shed light on our research with infants. According to the Direct Realism theory (Fowler, 1986), listeners use proximal invariants that reflect the distal stimulus during phonetic perception; articulatory gestures are encoded in the speech signal, and thus directly available to perception. The Dynamic Specification theory (Strange, 1989) was proposed to account for adults' ability to identify vowels based on segmental onset and offset transitional patterns (Strange, 1987). Formant transitions were also shown to cue consonantal contrasts in adults (e.g., Liberman et al., 1954). Infants can also perceived dynamic acoustic patterns. For example, six-month-old tone and non-tone learners both succeeded in distinguishing different contour tones in Thai (Mattock & Burnham, 2006). Even newborns can discriminate word lists differing in pitch contour (Nazzi, Floccia, & Bertoncini, 1998). In music perception studies with infants aged seven to 11 months, infants were able to discriminate directional pitch changes involving a single semitone difference (Thorpe, 1986) and extract directional cues for categorizing pitch sequences that vary irrelevantly in key or interval size (Trehub et al., 1987). These studies seem suggestive that infants might detect movement cues from pitch contours and use them to acquire intonations and tones. Such evidence, however, is only indirect with regards to the perception of movement information. The results are also compatible with the acoustical based approaches to the lack of invariance (Stevens, 1989). However, surface transition and contour patterns remain subject to between-speaker and contextual variability (Liberman et al., 1954). The type of evidence that would support the perception of movement needs to be beyond the surface acoustic patterns and directly refer to articulatory gestures.

Gauthier et al. (2007a; 2007b) recently attempted to tease apart the surface acoustic patterns versus underlying dynamic cues that can be derived from the surface pattern, using tones as the testing case. In an unsupervised learning paradigm, artificial neural networks

were trained with many exemplars of Mandarin tones produced by multiple speakers in continuous speech, with all possible tonal contexts and variable prosodic focus. In one condition, neural maps were trained with syllable-size fundamental frequency contours (F0), i.e., the surface acoustic patterns. In another condition, velocity profiles of F0 directly indicating patterns of articulatory movement were used, as represented by the discrete first derivatives of F0, i.e., $D1_t = 0.5\ (F0_{t+1}-F0_{t-1})$, where t represents the time points. When exposed to D1, the maps developed the categories corresponding to the Mandarin tones. In contrast, the trained maps exposed to raw F0 values showed more extensive between-category overlap. Importantly, learning D1 yielded distinct velocity profiles, each one representing a tone and characterizing the laryngeal gestures that overcome extensive input variability. These results suggest that velocity profiles lead to successful categorical learning. But whether infants compute D1 remains unknown.

To demonstrate infants' sensitivity to dynamic acoustic information reflecting articulatory gestures would require testing their perception of speech properties that involve the computation of D1. A basic but non-trivial regularity that humans are exposed to from birth is speech spoken by physically constrained production systems. Pitch can serve as an ideal test case. Laryngeal mechanics impose specific limits on F0 production such that pitch glides can only be achieved with a certain maximum speed (Ohala & Ewan, 1973). The maximum speed of pitch change in adult tonal production was recently determined (Xu & Sun, 2002), as shown in Eq. (1) and (2) for pitch rises and falls, respectively:

$$\text{MaxD1} = 10.8 + 5.6\,d \qquad (4.1)$$

$$\text{MaxD1} = 8.9 + 6.2\,d \qquad (4.2),$$

where the maximum speed of pitch change (MaxD1) is in semitones per second (st/*s*) and *d* the pitch excursion size in semitones. The minimum time required for a given pitch excursion can also be derived from these equations (minimum time = *d*/*s*). Thus, wider pitch excursions take a longer time. For example, raising pitch by four or 12 semitones can be achieved at a maximum speed of 33st/s and 78st/s, in a minimum time of 121 and 154 msecs. Although MaxD1 is often approached by adults (Xu & Sun, 2002), it is obviously not violated. Assuming that infants can perceptually derive this constraint based on their experience

hearing various intonation contours in the input, a perceptual boundary should exist between possible versus impossible D1 patterns for speech, arguably for cueing speech from non-speech auditory input.

In this study we used the MaxD1 boundary to examine infants' ability to distinguish between subtle changes in pitch velocity. Stimuli involved pitch contours with D1 patterns that approached and were within the MaxD1 constraint versus those just above the MaxD1 limit. If infants have the ability to compute D1, they should show the discrimination of subtle D1 changes, and furthermore, a preference for adult productions that respect MaxD1 given their accumulated experience of hearing intonation patterns produced by humans. A preference was expected for the less swift contours over the more swift but physiologically impossible contours in our experiment as previous work showed that infants prefer to listen to speech versus non-speech analogue (Vouloumanos & Werker, 2004). Such a preference would provide particularly strong support for the ability to track D1. To explore the age at which the ability to track D1 emerges, we compared four- and eight-month-old infants.

4.2.3 The experiment

Forty-six monolingual French-learning infants participated in the study. Six did not complete the experiment due to crying or fussing. The final dataset contained 20 infants for each of the two age groups (mean age of younger group = 4;5 and of older group = 7;9), half boys and girls.

An adult female speaker produced nonsense trisyllables /malama/ and /lamala/, respectively carrying /High Low High/ (/HLH/) and /LHL/ intonation patterns, and the trisyllable /mamama/ carrying both intonation patterns, for a total of four sequences. The speaker was instructed to produce the high and low pitches as differently and fast as possible, so that the MaxD1 was approached in the large pitch sweeps. The average duration was 993 msecs for the sequences, 285 msecs for the first syllable, 267 msecs for the second syllable, and 441 msecs for the third syllable. Each recorded sequence was modified using the Praat software, first by simplifying the natural pitch curves with a frequency resolution of two semitones. This reduced the number of pitch dots from over a 100 to 5, making subsequent manipulations more manageable. Possible stimuli were then created by resynthesizing the

pitch contour using linear predictive coding and replacing the natural pitch curve with the simplified one. Impossible stimuli were created in the same way, although the velocity profiles of part of the second syllable of each sequence was increased before the resynthesis. Thus, although both possible and impossible stimuli were modified from the natural production, the impossible tokens were slightly beyond the articulatory constraint. Both possible and impossible stimuli sounded natural, as judged by five adults coming from different linguistic backgrounds.

To test infants' subtle discrimination ability, we kept the difference between possible and impossible versions very small. Only the first half of the second syllable's pitch change was made more abrupt because a slope increase of the whole syllable excursion size would result in too large a perceptual difference. For each modified syllable, Table 4.1 contains the pitch excursion size and time from syllable onset to offset, the half value of this excursion size and the associated minimum excursion time, and the new excursion time for creating impossible stimuli. Figure 4.1 shows as an example this transformation for the pitch of lamala (LHL) sequence.

Table 4.1 Excursion size (column 2) in semitones and excursion time (column 3) in milliseconds of syllable onset to offset pitch change of second syllables; half excursion size (column 4) and associated minimum excursion time (column 5), and modified excursion time of half excursion size used for impossible stimuli (column 6).

Stimuli		Excursion size (st)	Excursion time (ms)	½ excursion size (st)	Minimum time for ½ excur. (ms)	Impossible time (min./4) (ms)
Lamala	LHL	12.9	258	6.5	138	34.5
Malama	HLH	10.6	153	5.3	127	31.75
Mamama	LHL	14.8	259	7.4	142	35.5
Mamama	HLH	13.8	280	6.9	134	33.5

A preferential looking procedure was used. In an IAC acoustic booth, the infant sat on his or her parent's lap in front of a central television screen and a hidden loud speaker that played the stimuli from the direction of the screen. The parent wore headphones to hear masking music. Each trial started when the infant looked at the screen and stopped after two

consecutive seconds of non-looking. A checkerboard was displayed on the screen during each trial, simultaneously with speech stimuli. A red light flashed on the screen between trials to attract the infant's attention. A researcher outside of the test room, who was blind to the auditory stimuli, pressed down a computer key whenever infant looked at the screen. The testing was run with an experimental software, which recorded all looking times online.

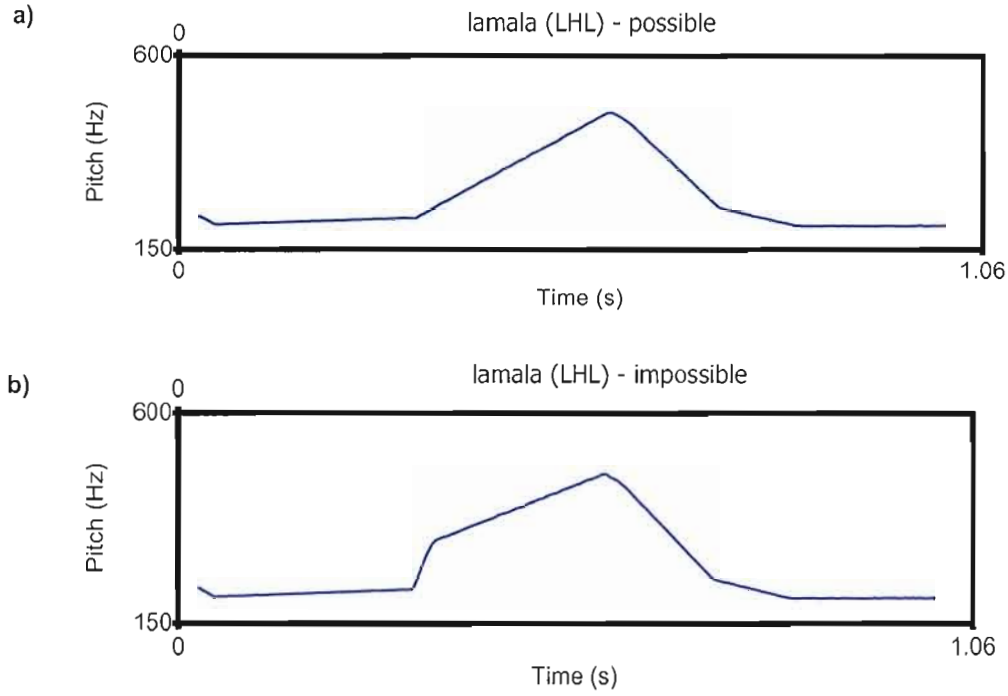


Figure 4.1 Simplified F0 contours of **a)** possible and **b)** impossible lamala (LHL) sequence.

Infants were presented with alternating trials of possible and impossible stimuli sequences of speech sounds. The auditory stimuli were presented in 17-second alternating trials of possible and impossible sequences, counterbalanced for the order of trial presentation (half of the infants began the testing by first hearing a possible trial and the other half first heard an impossible trial). Both possible and impossible trials contained modified versions of the natural pitch contours as described earlier. Each trial contained the four sequences, and each sequence was repeated two or three times, for a total of 11 exemplars, which were randomly ordered in a trial. The inter-stimulus interval of possible and impossible trials was

410 msecs, with slight adjustment to ensure that the total trial length being equal across all trials.

There were ten trials in total, five possible and five impossible. When piloting with adult listeners, discrimination was present for the initial four trials, after which they no longer could discriminate the two trial types, possibly reflecting perceptual adaptation. In light of this, we expected to observe looking difference in the initial four trials in infants. The total looking time across the four first trials containing the possible stimuli and those containing the impossible stimuli were calculated for each infant, and were analyzed in a 2x2x2 ANOVA, with Velocity (possible vs. impossible) as the within-subject factor and Age (4- vs. 8-month-old) and Order (possible first vs. Impossible first) as the between-subject factors[6]. The results revealed a significant main effect of Age, $F(1,38) = 8.589, p < 0.01$, with the younger group producing a longer overall listening time during the whole experiment than did the older group. No Velocity x Age interaction ($F(1,38) = 0.773, p = 0.385$), no main effect of Order ($F(1,38) = 0.696, p = 0.410$), no interaction of Velocity x Order ($F(1,38) = 0.002, p = 0.962$), and no three way interaction were observed ($F(1,38) = 2.471, p = 0.125$). Crucially, there was a significant main effect of Velocity, $F(1,38) = 7.609, p < 0.01$, indicating that both age groups can discriminate between possible and impossible speech sounds. Furthermore, infants showed a preference for possible sequences over impossible sequences, i.e., they listened longer to possible speech (Figure 4.2). Taken together, these results suggest that infants process velocity profiles and do so for normalizing the speech input. The means and standard errors of the means are shown in Table 4.2.

Table 4.2 Mean looking times and standard errors of the means.

Looking time	Possible		Impossible	
	Mean	Standard error	Mean	Standard error
Age				
4 month-old	22.585	2.056	20.415	1.849
8 month-old	17.240	1.880	13.040	1.325
Order				
Possible - Impossible	19.035	1.703	15.795	1.778
Impossible -Possible	20.790	2.352	17.660	1.831

[6] The same ANOVA analysis was performed for later trials and yielded no differential result, as expected based on adults' response.

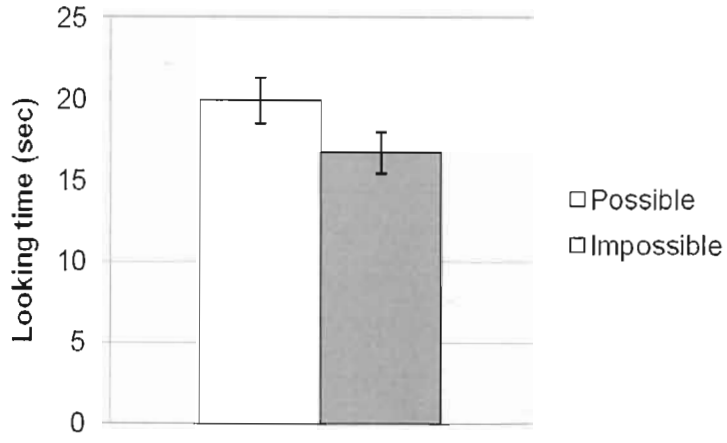


Figure 4.2 Infants' looking times while listening to speech stimuli respecting the constraint of maximum speed of pitch change (i.e., possible) versus those violating the constraint (impossible) (Mean and SEs).

4.2.4 Discussion

As predicted, infants listened longer to possible over impossible stimuli. This indicates that preverbal infants can distinguish between sequences of nonsense syllables that respect the maximum speed of vocal pitch change constraint from those which violate MaxD1. More importantly, infants' preference for pitch changes that are possible within a physical constraint of speech production suggests that D1 perception is present at an early stage of language acquisition. Infants' capacity to distinguish subtle speed changes as shown here indicates that they are able to process pitch velocity profiles well before one year of age, which in turn suggests that infants do analyze dynamic properties of the speech input that reflect articulatory gestures, consistent with the use of D1 for successful tonal learning shown in our previous simulation work (Gauthier et al., 2007a, 2007b). This is also consistent with recent brain imaging evidence showing activities in the superior temporal cortex (involved in auditory processing) during speech perception in 6-month-old infants, which also activated the motor cortex in the Broca's area, the left inferior frontal area associated with speech production in adults (Imada et al., 2006).

Another question that we raised concerns whether infants use D1 for normalizing the speech input. Given that the stimuli in this study were nonsense sequences produced by an

unknown speaker to the infants, the fact that they preferred possible over impossible D1 patterns generated from this novel speaker shows that they can potentially use D1 to normalize at least some amount of between-speaker variability. In fact, infants' preference suggests that they must have computed the D1 values of both types of stimuli and may have interpreted one of them as possible for the speech production system. This process of computing D1 is a means for normalizing absolute pitch differences between speakers. This suggests a useful strategy for learning speech sounds categories. Given that infants since birth are routinely exposed to speech sounds produced within the limit of human physical apparatus, the preference observed in this study suggest that experience hearing natural speech may allow infants to derive acoustic dynamic properties reflecting vocal movement. The fact that four- and eight-month-olds were equally good at discriminating subtle D1 variations and that both preferred listening to possible D1 patterns suggests that D1 tracking is already present at an early age, and that pitch movement information may be available for learning crucial linguistic structures such as intonation, focus, word-level prosody, and tones. We predict that infants may also be capable of tracking acoustic movement information reflecting other articulatory dimensions such as vocal track movement for vowel and consonant production. Future studies need to directly test D1 processing related to such articulatory parameters.

In sum, this study examined whether infants can process dynamic acoustic information that reflects articulatory gestures underlying speech sound categories. We approached this question by using an articulatory constraint related to the speed of vocal fold movement. Our results show that infants can discriminate between minimally contrasting pitch velocity profiles. Moreover, they prefer listening to velocity profiles that are possible for the production system over those that are impossible despite the fact that stimuli were nonsense sequences generated from an unknown speaker. These findings suggest that infants do process acoustic movement information that reflects articulatory gestures, and use it for normalizing and representing the speech input. We suggest that D1 processing may represent the key dynamic strategy for bridging the gap between production and perception in adults and in language acquisition.

Acknowledgment

This work was supported by funding from SSHRC, NSERC and CFI to the second author.

5 DISCUSSION

La première section de ce chapitre reprend les principaux résultats des deux études de modélisation en abordant en premier lieu la question de l'unité de traitement de la parole chez l'enfant, suivi d'une discussion sur la normalisation des tons lexicaux. Une description formelle du système de tons lexicaux mandarin est ensuite présentée. La deuxième section présente les résultats de l'étude comportementale et leurs implications pour les théories de l'invariance. Enfin, la troisième section présente une synthèse des trois études et propose l'esquisse d'un modèle d'acquisition phonétique, en abordant précisément la façon dont le mécanisme de l'invariance proposé peut rendre compte de la relation entre la perception et la production de la parole durant le stade initial de l'acquisition du langage.

5.1 Modélisation de l'acquisition des tons lexicaux en mandarin

5.1.1 Sur la question des caractéristiques phonétiques

L'une des questions soulevées dans cette thèse concernait la nécessité de postuler un mécanisme d'extraction des caractéristiques phonétiques chez l'enfant pour découvrir les sons de la parole. Pour tenter d'y répondre, les simulations proposaient de vérifier si les patrons continus d'une dimension acoustique particulière, la fréquence fondamentale (F0), pouvaient directement mener à l'acquisition des tons lexicaux.

Les simulations de la première étude et la première simulation de la seconde étude impliquaient des corpus de données contenant un degré modéré de variabilité. Les résultats montrent que les patrons de F0 de taille syllabique renferment une quantité raisonnable d'information catégorielle relative aux tons. Ceci suggère que l'extraction de caractéristiques phonétiques n'est pas nécessaire pour acquérir les sons de sa langue maternelle, du moins en ce qui a trait aux tons lexicaux mandarins. Lorsque les simulations impliquaient des degrés de variabilité plus élevés (seconde étude), les patrons de F0 ne suffisaient plus à la tâche. Une simple transformation des patrons de F0 en profils de vélocité préalable à l'apprentissage a cependant permis aux réseaux d'atteindre une performance similaire à celle de l'adulte. L'acquisition des catégories phonétiques à partir des patrons continus du signal de la parole s'avère une stratégie plus simple et directe, sans recours à l'extraction de caractéristiques

sommaires du signal. La représentation interne développée par ces réseaux a ensuite permis d'en extraire une représentation abstraite rappelant les caractéristiques phonétiques. Mais avant de discuter comment l'enfant peut former une représentation abstraite des tons suite à la catégorisation initiale de patrons de signal continus, la prochaine section décrit la nature de cette transformation et tente de répondre à la question principale de cette thèse.

5.1.2 Normaliser la parole par la poursuite du mouvement acoustique

Afin d'examiner si l'invariance perceptive précoce peut utiliser des stratégies similaires à celles de l'adulte, les simulations visaient à comparer l'efficacité de deux types d'information dans l'acquisition des tons lexicaux : le signal acoustique de surface et l'information dynamique sous-jacente du signal acoustique. Les réseaux neuronaux étaient donc exposés d'une part au stimulus proximal, c'est-à-dire les patrons de surface de la fréquence fondamentale, et d'autre part aux profils de vélocité de la fréquence fondamentale reflétant le geste articulatoire distal.

La vitesse s'est avérée une information utile pour caractériser les stratégies de contrôle moteur à la base de gestes articulatoires, telle que l'oscillation de la mâchoire durant la production de la parole (Nelson, 1983). Les mouvements du larynx durant la phonation impliquent également des profils de vélocité, notamment en ce qui a trait aux changements du taux de vibration des cordes vocales. Cette information est directement accessible à partir du signal de surface par le simple calcul de la première dérivée des patrons de F0.

Dans la simulation impliquant le plus haut et le plus naturel degré de variabilité (dernière simulation de la deuxième étude), le réseau entraîné à partir des patrons de F0 a réalisé un taux moyen de classification de 70% pour les quatre tons. Le stimulus proximal semble ainsi insuffisant pour la catégorisation tonale. Cependant, l'information dynamique a permis au réseau d'atteindre un taux de classification de 90% en contexte de variabilité modérée, et la supériorité de D1 sur F0 s'est avérée robuste malgré l'accroissement du degré de variabilité. De plus, au contraire des réseaux exposés à F0, les réseaux D1 ont atteint un taux de classification comparable à l'identification des tons chez l'adulte en conditions similaires. Dans la simulation impliquant le plus haut degré de variabilité, le réseau D1 pouvait comme

l'adulte mieux distinguer les tons accentués par focus prosodique que les tons sans focus, atteignant un taux de classification de 93%, pour 91% chez l'humain (Prom-on et al., 2009).

Le focus prosodique est une fonction linguistique qui utilise les patrons d'intonation pour souligner verbalement une partie d'un message. Tel que suggéré par une récente étude de modélisation simulant le développement de la perception de la parole, les profils de vélocité de F0 peuvent également s'avérer utile pour l'acquisition du focus (Gauthier et al., 2009). Dans cette étude, des réseaux neuronaux non supervisés ont été entraînés avec des patrons de F0 et D1 dans le but de distinguer diverses conditions focales. Les résultats montrent que les réseaux entraînés avec F0 offrent un patron de réponses erratiques, alors que la performance du réseau D1correspond à celle d'auditeurs adultes, tous deux pouvant mieux distinguer des énoncés de trois mots dont le focus se situe sur le premier ou le second mot que des énoncés dont le focus est neutre ou se situe sur le dernier mot.

Ces résultats illustrent l'impact important de diverses sources de variabilité sur le stimulus proximal, soulignant l'insuffisance d'une approche strictement acoustique à l'invariance dans la catégorisation tonale à partir de F0. Par ailleurs, la vitesse de la fréquence fondamentale, malgré la variabilité induite par de multiples locuteurs et contextes de production, permet de catégoriser les tons lexicaux et de distinguer les éléments d'autres types de fonctions communicatives comme le focus prosodique. En supposant que le système perceptif puisse effectuer la conversion des patrons de F0 en patrons de D1, ces résultats peuvent s'expliquer dans le cadre de l'invariance auditive (par ex. Ménard et al., 2002; Miller, 1989). Une simple transformation auditive des patrons de F0 en leurs profils de vélocité permettrait de former quatre catégories correspondant aux quatre tons, et de traiter les signaux de façon comparable à l'adulte faisant face à un degré similaire de variabilité. Cette explication ne peut toutefois rendre compte de la relation entre la perception et la production de la parole.

Une perspective écologique de la perception de la parole (Fowler, 1986) permet au contraire d'établir un tel lien, en interprétant les profils de vélocité de la fréquence fondamentale comme le résultat spécifique de la structuration par le geste articulatoire du signal acoustique et de sa dynamique. La poursuite du mouvement acoustique reflétant les

gestes articulatoires représente ainsi une stratégie utile non seulement pour l'adulte, mais aussi pour l'enfant afin de résoudre le problème de la variabilité. À l'instar de la théorie motrice de l'invariance (Liberman & Mattingly, 1985), le calcul de la première dérivée constitue une opération simple et ne nécessite aucun accès aux schèmes moteurs, de toute façon embryonnaires sinon inexistants chez l'enfant en bas âge. Les profils de vélocité, en représentant les gestes invariants impliqués dans la production tonale, permettrait cependant d'établir la relation entre la perception et la production de la parole, comme l'explique la prochaine section.

5.1.3 Nature des tons et caractéristiques phonétiques revisitées

Comme le suggèrent les études de modélisation, la représentation des sons de la parole par un ensemble de caractéristiques phonétiques ne précèderait pas mais résulterait plutôt de l'acquisition des catégories phonétiques. Les résultats des simulations montrent en effet que le signal continu de la parole s'avère suffisant pour la formation des catégories tonales. Les prochaines lignes décrivent comment un tel apprentissage peut mener à une représentation mature des tons basées sur les caractéristiques phonétiques, elles-mêmes entretenant un lien direct avec la production tonale.

Le type de réseau neuronal utilisé dans cette thèse, les cartes auto-organisées (Self-Organizing-Maps (SOMs), Kohonen, 1982, 1995), permet de projeter un espace stimulus multidimensionnel et continu sur une grille d'unités discrètes à plus faibles dimensions. La dimensionnalité intrinsèque d'un corpus de données est habituellement plus petite que sa dimensionnalité de surface en raison de la corrélation entre les données (Bishop, 1995). De façon similaire à une analyse en composantes principales, le SOM extrait automatiquement la dimensionnalité intrinsèque de l'espace stimulus, c.-à-d. les dimensions sur lesquelles les vecteurs-stimuli affichent la plus grande variance, pour directement refléter cette structure sur la carte neuronale une fois l'apprentissage complété (Kohonen, 1989). Dans cette thèse, et comme c'est généralement le cas avec le SOM, les données étaient comprimées dans un quadrillage neuronal bidimensionnel. Suite à l'entraînement des réseaux avec D1, l'analyse de la structure interne des cartes neuronales a montré que deux dimensions étaient suffisantes pour représenter le système de tons mandarins.

Dans la première simulation de la première étude, le nombre d'unités des cartes neuronales était considérablement plus grand que celui des catégories à découvrir. Cette première étape a permis d'observer la structure générale des données, de comparer l'organisation tonale à l'intérieur des réseaux F0 et D1, et de découvrir quatre groupements distincts d'unités répondant chacun à un des quatre tons lexicaux dans le réseau D1. Cette représentation détaillée (*fine-grained*) a également permis d'explorer la structure interne des tons et de caractériser la variabilité intra-catégorielle. La seconde simulation de la première étude visait à modéliser l'abstraction des catégories tonales en projetant les représentations développées dans le réseau initial sur une seconde carte auto-supervisée de quatre unités. La formation du nouveau SOM à partir de D1, mais non avec F0, a permis de dégager quatre profils de vélocité correspondant aux quatre tons.

Les quatre unités du réseau offre un portrait global (*coarse-grained*) du système de tons mandarins, où chaque profil de vélocité correspond à la moyenne d'un groupe d'unités de la carte neuronale initiale. Les profils schématiques de la Figure 5.1 montrent une approximation continue des profils développés par le réseau réduit de quatre unités (voir Figure 2.7). Ces courbes correspondent à des équations polynômiales de second degré dont les coefficients permettent de spécifier et de séparer les quatre catégories tonales, tel que prédit au Chapitre 1 (1.3.4). Le coefficient linéaire spécifie la hauteur du sommet ou de la

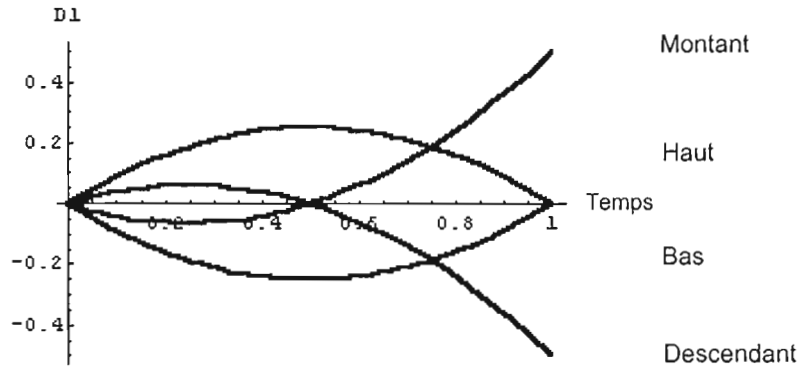


Figure 5.1 Profils de vélocité schématiques des quatre tons mandarins (Haut, Montant, Bas, Descendant).

vallée de la fonction, alors que le coefficient quadratique correspond au moment du point de flexion (c.-à-d. vitesse nulle). En termes de caractéristiques acoustiques, la hauteur spécifie la direction initiale de F0 (haut ou bas), et permet de distinguer entre eux les tons de niveaux et les tons dynamiques. Le point de flexion spécifie la pente de la direction initiale, distinguant les deux tons de niveau des deux tons dynamiques.

Ces deux caractéristiques représentent également les gestes invariants impliqués dans la production tonale. Tel qu'illustré dans la Figure 5.1, le ton Haut affiche une augmentation du taux de vibration des cordes vocales, un pic de vélocité au milieu de la syllabe, et une diminution de vitesse jusqu'à un taux de changement nul à la fin de la syllabe. Le profil du ton Bas constitue une image miroir du ton Haut. Ces patrons de vélocité reflètent un geste simple, qui consiste à partir d'une position pour s'arrêter à une autre, et dans le cas présent d'une valeur initiale vers une valeur finale de F0. Les deux tons dynamiques se distinguent également entre eux par leur direction initiale, et se distinguent des tons de niveau par un changement de vitesse intra-syllabique vers une direction opposée à la direction initiale, l'atteinte d'une vitesse nulle vers le milieu de la syllabe, et une vitesse positive ou négative à la fin de la syllabe.

Le système de tons lexicaux mandarin peut donc se résumer à partir de quatre profils de vélocité, et sa description quantitative sur le plan articulatoire fait appel à une représentation dynamique, où chaque profil correspond à une équation différentielle non linéaire de la forme $Dl = a\,(b - x)\,x$. Le système de tons mandarins correspond ainsi au système dynamique :

$$\begin{aligned} Dl_H &= a\,(b - x)\,x \\ Dl_R &= -a\,(b/y - x)\,x \\ Dl_L &= -a\,(b - x)\,x \\ Dl_F &= a\,(b/y - x)\,x \end{aligned} \qquad (5.1),$$

où a spécifie le sommet de la vélocité (positif ou négatif), b la vitesse finale (y indiquant le passage à zéro du profil de vélocité), et dont la solution mène à la réalisation de surface des tons. Ce modèle est compatible avec la théorie dynamique de la phonologie articulatoire (Browman & Goldstein, 1986, 1992, 1995), de même qu'avec le modèle d'approximation de cibles de la production tonale (Xu & Wang, 2001), dans lequel les coefficients linéaires a et b

correspondent aux paramètres de commande articulatoires statiques et dynamiques spécifiant la production tonale.

L'un des principaux avantages de l'apprentissage en deux étapes tel que proposé par les simulations est de ne pas avoir à postuler quatre catégories phonétiques abstraites a priori. L'acquisition des catégories phonétiques procède plutôt ainsi de la classification initiale des données brutes suivant leur similarité. La seconde étape permet de former les prototypes auditifs continus de ces catégories, à partie desquelles l'enfant peut éventuellement extraire les caractéristiques phonétiques en associant les profils perçus aux gestes produits durant ses propres vocalisation. Cette séquence d'acquisition est plausible du point de vue de l'apprentissage non supervisé chez l'enfant et sera discuté plus en détail sans la Section 5.3.

En résumé, les résultats des études de modélisation montrent qu'un simple dispositif de reconnaissance de patrons permet de traiter le signal acoustique de la parole en catégories linguistiquement pertinentes, en accord avec les mécanismes d'adaptation/induction et les capacités d'analyse distributionnelle observées chez l'enfant. Les résultats indiquent également que les patrons continus du signal de la parole permettent de former quatre catégories correspondant aux quatre tons lexicaux mandarins, éliminant le besoin de supposer un mécanisme inné d'extraction de caractéristiques phonétiques. Enfin, les simulations démontrent la supériorité de D1 sur F0 pour catégoriser les tons lexicaux, suggérant que la poursuite de l'information acoustique reflétant le mouvement articulatoire s'avère une stratégie utile que l'enfant en bas âge peut utiliser afin d'acquérir les catégories phonétiques de sa langue maternelle.

5.2 La perception auditive du mouvement chez l'enfant préverbal

Alors que les études de simulation démontraient l'utilité de la vitesse de la fréquence fondamentale pour révéler les invariants associés aux tons lexicaux, l'étude comportementale explorait si les enfants en bas âge sont sensibles à l'information dynamique du signal de la parole et peuvent normaliser la variabilité à partir de cette information. À l'aide d'une procédure de regard préférentiel, l'étude vérifiait la capacité des enfants à percevoir des variations subtiles de vitesse reflétant une contrainte de production. La tâche consistait plus précisément à distinguer entre eux des patrons d'intonation possibles et impossibles sur le

plan articulatoire, ces derniers violant à certains endroits la contrainte de vitesse maximale de changement de la fréquence fondamentale, ou MaxD1. Puisque l'environnement reflète les limites de l'appareil vocal durant la phonation, et comme l'enfant en bas âge manifeste une préférence pour les stimuli paroliers (Vouloumanos & Werker, 2004), la prédiction voulait que les enfants écoutent plus longuement les stimuli possibles, qui respectent la contrainte articulatoire. Enfin, des enfants de 4 et 8 mois ont participé à l'étude, ceci dans le but d'observer l'évolution de la sensibilité à D1 durant la première année de vie.

Les résultats montrent que les enfants préfèrent écouter les stimuli possibles. Le fait que ces stimuli différaient seulement en termes de changement de vitesse intra-syllabique indique que les enfants peuvent percevoir des variations subtiles de D1. Ceci suggère qu'ils perçoivent les profils de vélocité de la fréquence fondamentale de la parole et qu'ils peuvent calculer les premières dérivées de l'information spectrale continue. En outre, les deux groupes d'âge ont montré une réponse préférentielle pour les stimuli possibles, suggérant que la perception de D1 opère dès l'âge de quatre mois, et peut-être même avant.

Par ailleurs, le fait que les enfants aient écouté plus longuement les stimuli possibles suggère une préférence de leur part pour un signal de la parole marqué d'une contrainte articulatoire, celle de la vitesse maximale de changement de fréquence fondamentale. Cette préférence n'est pas étonnante étant donnée la fréquence élevée des sons possibles dans l'environnement de l'enfant et l'effet de familiarité généralement observé à cet âge dans ce type d'études. Cependant, le fait que les stimuli expérimentaux provenaient d'un locuteur inconnu aux enfants montre que ceux-ci doivent avoir calculé D1 et utilisé les profils de vélocité de la fréquence fondamentale afin de normaliser la variabilité interlocuteur.

En résumé, les résultats de l'étude comportementale indiquent que les enfants sont sensibles à la limite de vitesse de modulation de la fréquence fondamentale imposées par les contraintes articulatoires chez l'adulte, suggérant que la poursuite auditive du mouvement continu du signal de la parole reflétant les gestes articulatoires est une capacité qui émerge tôt durant la première année de vie.

5.3 Modèle de l'acquisition phonétique

Les études présentées dans cette thèse portaient sur la perception d'une dimension particulière du signal de la parole, la fréquence fondamentale, à partir de laquelle peut s'effectuer l'apprentissage d'un type précis de catégories phonétiques, les tons lexicaux. Le modèle proposé peut toutefois se généraliser au développement de la perception phonétique et s'appliquer à d'autres dimensions acoustiques ou sons de la parole. Le modèle d'apprentissage statistique implémenté dans les études de simulation rappelle la capacité du système auditif, chez l'enfant comme l'animal, à percevoir la distribution de propriétés phonétiques dans le signal de la parole (Maye et al., 2002; Pons, 2006). D'autres chercheurs ont modélisé l'acquisition d'autres types de catégories phonétiques à partir du signal de la parole par le biais de réseaux neuronaux artificiels (par exemple, Behnke, 1998; Vallabha, McClelland, Pons, Werker, & Amano, 2007). Les simulations proposées dans cette thèse supportent ainsi l'apprentissage distribué non supervisé en tant que mécanisme à la base de la réorganisation perceptive chez l'enfant (Aslin et al., 1983). Le modèle d'acquisition en deux étapes permet pour sa part de rendre compte de la formation de prototypes phonétiques, processus auparavant attribuée à un mécanisme propre à l'humain et spécifique au langage (Kuhl, 1991). La capacité de l'animal à former des prototypes phonétiques (Kluender et al., 1998) suggère toutefois qu'un mécanisme plus spécifique sous-tend l'acquisition des catégories phonétiques chez l'enfant. Le mode spécial de perception relèverait plutôt de la complexité des catégories développées chez l'enfant (e.g., Lotto, 2000), et plus précisément de la capacité des prototypes perceptifs formés à éventuellement établir le lien avec la production de la parole.

La nature de la relation entre la perception et la production demeure un problème majeur pour les théories de la parole. Plusieurs se sont penchés sur la façon dont la perception influence la production et *vice-versa* durant l'acquisition phonétique. L'idée selon laquelle la perception de la parole agit sur la vocalisation chez l'enfant en bas âge se résume par le concept de dérive du babillage (Brown, 1958), ou l'impact de l'environnement linguistique sur la production des sons de la parole, qui s'observe dès l'âge de 10 mois (deBoysson-Bardies, Halle, Sagart, & Durand, 1989). En ce qui a trait à l'influence de la production sur la perception, l'imitation apparaît comme un mécanisme important dans l'acquisition des

catégories phonétiques. Le nombre de mots que les enfants tentent d'imiter augmente vers l'âge d'un an (Vihman & Miller, 1988). L'imitation impliquerait d'abord la projection des représentations perceptives de la parole adulte sur les schèmes moteur vocaux, qui correspondent aux patrons de production que préfèrent les enfants (Vihman, 1986). La relation auditive-articulatoire se développerait ensuite par l'exploration des possibilités vocales durant le babillage (Vihman, 1993). L'enfant guideraient dès l'âge de 20 semaines ses propres vocalisations à partir des sons de la parole adulte stockés en mémoire perceptive (Kuhl & Meltzoff, 1996). L'imitation requiert chez l'enfant la capacité de comparer le signal de la parole entrant au signal sortant afin de juger du succès de sa production. La production vocale d'enfants en bas âge diffère toutefois considérablement du signal d'entrée de la parole adulte (Hillenbrand et al., 1995; Ménard et al., 2004; Peterson & Barney, 1952), du moins en ce qui a trait au signal de surface. Le développement du lien sensorimoteur exige donc de stocker l'information auditive-perceptive de façon indépendante du locuteur, mais aussi dans un format compatible avec les schèmes de production de l'enfant.

Les simulations de cette thèse suggèrent que malgré de multiples sources de variabilité dans le signal de la parole, la poursuite auditive du mouvement acoustique permet la catégorisation phonétique simplement par le calcul de la première dérivée de la fréquence fondamentale. L'étude de perception prête un appui comportemental à ce mécanisme, révélant la capacité d'enfants de 4 et 8 mois à distinguer les profils de vélocité du signal de la parole et à normaliser la variabilité interlocuteur en calculant D1. Les résultats indiquent également que les profils de vélocité de la fréquence fondamentale caractérisent les gestes invariants du larynx impliqués dans la phonation tonale, capturant ainsi la similarité entre la production adulte et celle de l'enfant. La boucle perception-production pourrait ainsi se fonder sur les traces dynamiques laissées par la parole adulte, à partir desquelles l'enfant en bas âge peut éventuellement guider son expérience vocale.

Le rôle du mouvement acoustique et de l'information dynamique dans la normalisation de la parole et l'acquisition des catégories phonétiques demeure une question empirique. Plusieurs évidences pointent toutefois dans cette direction. Les adultes peuvent par exemple identifier des voyelles produites dans divers contextes sur la base des transitions formantiques (Gottfried & Strange, 1980; Jenkins, Strange, & Edman, 1983; Strange, Verbrugge,

Shankweiler, & Edman, 1976), observation ayant mené au modèle de spécification dynamique de la perception vocaliques (Strange, 1987, 1989). De plus, l'étude acoustique du signal de la parole permet d'inférer les configurations et mouvements articulatoires associés à la production de consonnes à partir d'éléments précis du spectre continu de la parole (Stevens, 1993). Enfin, le taux de changement du second formant impliqué dans la production de diphtongues est essentiellement invariant (Gay, 1968) et s'avère pertinent dans l'identification de diphtongues en anglais (Gay, 1970). Les profils de vélocité de la fréquence fondamentale, combinés à la vitesse d'autres dimensions acoustiques et à d'autres propriétés dynamiques du signal, offre ainsi un cadre unificateur permettant d'appréhender la relation entre la perception de la parole, la production de la parole et l'acquisition du langage.

CONCLUSION

Trois études ont été proposées afin de tester l'hypothèse selon laquelle les enfants peuvent utiliser l'information dynamique du signal acoustique qui reflète les gestes articulatoires afin de normaliser la variabilité de la parole et acquérir les catégories phonétiques de leur langue maternelle. Deux études de modélisation proposaient de décrire de façon formelle le développement de la perception de la parole tel qu'il peut se produire chez l'enfant à l'aide d'un outil informatique simple et biologiquement plausible. Une étude de perception visait ensuite à appuyer les mécanismes proposés dans les simulations au niveau comportemental, en spécifiant si et à quel âge ceux-ci émergent chez l'enfant.

Les études de modélisation simulaient les mécanismes fondamentaux de perception, de normalisation et d'apprentissage impliqués dans l'acquisition des tons lexicaux. Les résultats démontrent qu'il est possible pour un réseau neuronal artificiel auto-organisé d'extraire le système de tons lexicaux chinois mandarin en détectant les régularités du signal de la parole, sans information préalable quant au nombre de catégories à apprendre. De plus, les simulations montrent que l'acquisition des tons lexicaux peut se réaliser en traitant des portions continues du signal de la parole, sans recourir aux propriétés acoustiques-phonétiques sommaires caractérisant chaque ton. Le corpus de stimuli utilisé pour entrainer les réseaux neuronaux impliquait toutefois de multiples sources de variabilité, faisant en sorte que le calcul des propriétés statistiques de l'information spectrale continue s'avère insuffisant pour la tâche. Suite à la performance sous-optimale des patrons de fréquence fondamentale de surface pour catégoriser les tons lexicaux, les simulations ont testé l'hypothèse selon laquelle l'invariance perceptive pour les sons de la parole se situe au niveau dynamique sous-jacent. Les résultats montrent que le mouvement acoustique du signal surface, tel que représenté par les profils de vélocité de la fréquence fondamentale, s'avèrent suffisant pour extraire la structure tonale à partir du signal de la parole malgré un degré important de variabilité. En outre, les quatre profils de vélocité prototypiques développés par le réseau, en plus de correspondre aux quatre tons mandarins, caractérisent de façon explicite les gestes invariants du larynx impliqués dans la production des tons. Enfin, l'analyse dynamique des profils de vélocité révèle que deux paramètres de contrôle déterminent les patrons de fréquence

fondamentale requis pour produire chaque ton, offrant un mécanisme précis pour rendre compte de la relation entre la production et la perception de la parole.

Si la vitesse constitue un indice robuste des tons lexicaux pour les réseaux neuronaux, la modélisation demeure hypothétique jusqu'à l'observation de la capacité humaine à effectuer les mêmes opérations. L'étude de perception a donc testé l'hypothèse selon laquelle les enfants en bas âge sont sensibles aux caractéristiques dynamiques du signal acoustique et peuvent normaliser la variabilité à partir de cette information. À l'aide d'une procédure de regard préférentiel, l'étude explorait la capacité d'enfants préverbaux à percevoir des variations acoustiques reflétant une contrainte articulatoire, la vitesse maximale de changement de la fréquence fondamentale. La tâche consistait pour 40 enfants de 4 et 8 mois à distinguer des patrons d'intonation possibles et impossibles sur le plan articulatoire. Les résultats montrent que les enfants des deux groupes d'âge peuvent distinguer des patrons de fréquence fondamentale qui diffèrent seulement en termes de changement de vitesse intra-syllabique. Ceci indique qu'ils peuvent percevoir des variations subtiles de vélocité telles que celles impliquées dans l'identité tonale, et qu'ils peuvent calculer les premières dérivées d'informations spectrales continues. Les résultats montrent également que les enfants ont plus longuement écouté les patrons d'intonation possibles, suggérant une préférence envers les stimuli verbaux qui respectent la contrainte articulatoire. Cette préférence, induite par la parole d'un locuteur inconnu, indique que l'enfant peut utiliser les profils de vélocité de la fréquence fondamentale pour normaliser la variabilité interlocuteur. Enfin, les résultats suggèrent que l'action normalisatrice des profils de vélocité peut opérer dès l'âge de quatre mois. Cette étude illustre comment le développement de la constance perceptive envers les catégories phonétiques peut s'effectuer par la poursuite auditive de patrons acoustiques dynamiques, stratégie grâce à laquelle l'enfant pourrait éventuellement superviser ses propres productions.

En conclusion, le modèle formel proposé a d'abord permis de formuler et d'évaluer des hypothèses spécifiques au sujet du rôle de l'information dynamique dans le développement phonétique. Ensuite, la découverte d'une sensibilité envers les profils de vélocité chez l'enfant en bas âge, en plus de s'ajouter à ses multiples capacités documentées par des années de recherche en perception développementale de la parole, supporte le rôle possible de la

poursuite auditive du mouvement acoustique reflétant les gestes articulatoires dans la catégorisation des sons de la parole. Enfin, en combinant l'approche comportementale et la modélisation, cette thèse a permis de mieux comprendre les mécanismes à la base de l'acquisition des catégories phonétiques et de caractériser la relation entre la perception de la parole, la production de la parole et l'acquisition du langage. D'autres études sont nécessaires pour explorer le rôle de l'information dynamique dans diverses tâches d'acquisition (des voyelles ou consonnes par exemple) et, dans la lignée de recherche actuelle visant à établir le fondement neurologique de la parole, afin de fournir l'évidence directe du traitement neuronal des profils de vélocité.

RÉFÉRENCES

Abramson, A. S. (1962). *The vowels and tones of standard Thai: Acoustical measurements and experiments* (Vol. 20). Bloomington, IN: Indiana University Research Center in Anthropology, Folklore, and Linguistics.

Abramson, A. S. (1978). Static and dynamic acoustic cues in distinctive tones. *Language and Speech, 21*(4), 319-325.

Anderson, J. L., Morgan, J. L., & White, K. S. (2003). A statistical basis for speech sound discrimination. *Language and speech, 46*(2-3), 155-182.

Andruski, J. E., & Costello, J. (2004). Using polynomial equations to model pitch contour shape in lexical tones: an example from Green Mong. *Journal of the International Phonetic Association 34*, 125-140.

Andruski, J. E., & Ratliff, M. (2000). Use of phonation type in distinguishing tone: The case of Green Mong. *Journal of the International Phonetic Association, 30*, 39-62.

Aslin, R. N., Pisoni, D. B., & Jusczyk, P. W. (1983). Auditory development and speech perception in infancy. In P. H. Mussen (Ed.), *Handbook of child phonology* (Vol. 2: Infancy and developmental psychobiology, pp. 573-687). New York: John Wiley & sons.

Aslin, R. N., Werker, J. F., & Morgan, J. L. (2002). Innate phonetic boundaries revisited. *Journal of the Acoustical Society of America, 112*(4), 1257-1260.

Behnke, K. (1998). *The acquisition of phonetic categories in young infants: A self-organizing artificial neural network approach.* Published doctoral dissertation, Universiteit Twente, Enschede, The Netherlands.

Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P. W., Kennedy, L. J., & Mehler, J. (1988). An investigation of young infants' perceptual representations of speech sounds. *Journal of Experimental Psychology: General, 117*(1), 21-33.

Bertoncini, J., & Mehler, J. (1981). Syllables as units in infant speech perception. *Infant Behavior and Development, 4*(3), 247-260.

Best, C. T. (1993). Emergence of language-specific constraints in perception of non-native speech: A window on early phonological development. In B. deBoysson-Bardies, S. Schonen, P. W. Jusczyk, P. McNeilage & J. Morton (Eds.), *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 289-304). Dordrecht: Kluwer Academic Publishers.

Best, C. T. (1995). A direct realist perspective on cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Theoretical*

and methodological issues in cross-language speech research (pp. 167-200). Timonium, MD: York Press.

Best, C. T., McRoberts, G. W., & Goodell, E. (2001). American listeners' perception of nonnative consonant contrasts varying in perceptual assimilation to English phonology. *Journal of the Acoustical Society of America, 1097*, 775-794.

Best, C. T., McRoberts, G. W., & Sithole, N. N. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance, 14*(3), 345-360.

Bijeljac-Babic, R., Bertoncini, J., & Mehler, J. (1993). How do 4-day-old infants categorize multisyllabic utterances? *Developmental Psychology, 29*(4), 711-721.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Clarendon Press.

Blumstein, S. E., & Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America, 66*, 1001-1017.

Blumstein, S. E., & Stevens, K. N. (1981). Phonetic features and acoustic invariance in speech. *Cognition, 10*, 25-32.

Bradlow, A. R., Pisoni, D. B., Yamada, R. A., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/ : IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America, 101*, 2299-2310.

Brent, M., & Siskind, J. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition, 81*(2), B33-44.

Browman, C. P., & Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook, 3*, 219-252.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica, 49*, 155-180.

Browman, C. P., & Goldstein, L. (1995). Dynamics and articulatory phonology. In R. F. Port & T. van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition* (pp. 175-194). Cambridge: MIT Press.

Brown, R. (1958). *Words and things*. Glencoe: Free Press.

Carrell, T. D., & Smith, L. B. (1979). Some perceptual dependencies between vowel color and pitch. *Journal of the Acoustical Society of America, 65*(S1), S7.

Caselli, M. C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., et al. (1995). A cross-linguistic study of early lexical development. *Cognitive Development, 10*, 159-199.

Chao, Y. R. (1933). Tone and intonation in Chinese. *Bulletin of the Institute of History and Philology, 4*, 121-134.

Chao, Y. R. (1968). *A grammar of spoken Chinese*. Berkeley, CA: University of California Press.

Chen, Y., & Xu, Y. (2006). Production of weak elements in speech - Evidence from f0 patterns of neutral tone in standard Chinese. *Phonetica, 63*(1), 47-75.

Chistovich, L. A., & Lublinskaya, V. V. (1979). The 'center of gravity' effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. *Hearing Research, 1*(3), 185-195.

Chomsky, N. A., & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.

Creelman, C. D. (1957). Case of the Unknown Talker. *Journal of the Acoustical Society of America, 29*(5), 655.

Crottaz-Herbette, S., & Ragot, R. (2000). Perception of complex sounds: N1 latency codes pitch and topography codes spectra. *Clinical Neurophysiology, 111*, 1759-1766.

Cutler, A., & Norris, D. G. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human perception and Performance, 14*, 113-121.

deBoysson-Bardies, B., Halle, P., Sagart, L., & Durand, C. (1989). A crosslinguistic investigation of vowel formants in babbling. *Journal of Child Language, 16*(1), 1-17.

DeFrancis, J. (1984). *The Chinese Language: Fact and Fantasy*. Honolulu: University of Hawaii Press.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America, 27*(4), 769-773.

Diehl, R. L., & Kluender, K. R. (1989). On the objects of speech perception. *Ecological Psychology, 1*(2), 121-144.

Duanmu, S. (2000). *The phonology of standard Chinese*. Oxford: Oxford University Press.

Eilers, R. E., Gavin, W., & Wilson, W. R. (1979). Linguistic experience and phonemic perception in infancy: A crosslinguistic study. *Child Development, 50*(1), 14-18.

Eilers, R. E., Wilson, W. R., & Moore, J. M. (1977). Developmental changes in speech discrimination in infants. *Journal of Speech and Hearing Research, 20*(4), 766-780.

Eimas, P. D. (1974). Auditory and linguistic processing of cues for place of articulation by infants. *Perception and Psychophysics, 16*, 513-521.

Eimas, P. D. (1975). Speech perception in early infancy. In L. B. Cohen & P. Salapatek (Eds.), *Infant perception: from sensation to cognition (vol. 2)*. New York: Academic Press.

Eimas, P. D., & Miller, J. L. (1980). Contextual effects in infant speech perception. *Science, 209*, 1140-1141.

Eimas, P. D., Siqueland, E. R., Jusczyk, P. W., & Vigorito, J. (1971). Speech perception in infants. *Science, 171*(3968), 303-306.

Elman, J., & McClelland, J. (1986). Exploiting lawful variability in the speech wave. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 360-385). Hillsdale: Lawrence Erlbaum.

Fant, G. (1986). Features: Fiction and facts. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 480-492). Hillsdale: Lawrence Erlbaum.

Flege, J. E. (1989). Chinese subjects' perception of the word-final English /t/-/d/ contrast: Before and after training. *Journal of the Acoustical Society of America, 15*, 67-83.

Fodor, J. A., Garrett, M., & Brill, S. (1975). Pi ka pu: The perception of speech sounds by prelinguistic infants. *Perception and Psychophysics, 18*, 74-78.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14*, 3-28.

Fowler, C. A., & Rosenblum, L. D. (1986). The perception of phonetic gestures. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor-theory of speech perception*. Hillsdale: Erlbaum.

Friederici, A. D., & Wessels, J. M. (1993). Phonotactic knowledge of word boundaries and its use in infant speech perception. *Perception & Psychophysics, 54*(3), 287-295.

Fromkin, V. (1978). *Tone: A linguistic survey*. New York: Academic.

Fujimura, O. (2000). The C/D model and prosodic control of articulatory behavior. *Phonetica, 57*, 128-138.

Gandour, J. (1978). The perception of tones. In V. A. Fromklin (Ed.), *Tone: A linguistic survey* (pp. 41-76). New York: Academic.

Gandour, J. (1979). Perceptual dimension of tones. In G. I. Liem (Ed.), *Southeast Asian linguistic studies: Vol. 3. Pacific Linguistics (Series C, No. 45)*. Canberra: Australia National University, Department of Linguistics.

Gandour, J. (1983). Tone perception in Far Eastern languages. *Journal of Phonetics, 11*, 149-175.

Gandour, J. (2000). Frontiers of brain mapping of speech prosody. *Brain and Language, 71*, 75-77.

Gandour, J., & Harshman, R. A. (1978). Crosslanguage differences in tone perception: a multidimensional scaling investigation. *Language and Speech, 21*(1), 1-33.

Gandour, J., Potisuk, S., & Dechongkit, S. (1994). Tonal coarticulation in Thai. *Journal of Phonetics, 22*, 477-492.

Gauthier, B., Shi, R., & Xu, Y. (2007a). Learning phonetic categories by tracking movements. *Cognition, 103*(1), 80-106.

Gauthier, B., Shi, R., & Xu, Y. (2007b). Simulating the acquisition of lexical tones from continuous dynamic input. *Journal of the Acoustical Society of America, 121*(5), EL190-EL195.

Gauthier, B., Shi, R., & Xu, Y. (2009). Learning prosodic focus from continuous speech input: A neural network exploration. *Language Learning and Development, 5*(2), 94-114.

Gay, T. (1968). Effect of speaking rate on diphthong formant movements. *Journal of the Acoustical Society of America, 44*(6), 1570-1573.

Gay, T. (1970). A perceptual study of American English Diphthongs. *Language and Speech, 13*(2), 65-88.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Goldsmith, J. A. (1990). *Autosegmental and metrical phonology*. Oxford: Blackwell Publishers.

Gottfried, T. L., & Strange, W. (1980). Identification of coarticulated vowels. *Journal of the Acoustical Society of America, 68*(6), 1626-1635.

Gottlieb, G. (1983). The psychobiological approach to developmental issues. In P. H. Mussen (Ed.), *Handbook of child psychology* (Vol. 2: Infancy and developmental psychobiology, pp. 1-26). New York: John Wiley & Sons.

Grieser, D., & Kuhl, P. K. (1989). Categorization of speech by infants: Support for speech-sound prototypes. *Developmental Psychology, 25*(4), 577-588.

Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge: Cambridge University Press.

Halle, M., & Stevens, K. N. (1962). Speech recognition: A model and a program for research. *IEEE Transactions on Information Theory, 8*(2), 155-159.

Han, M. S., & Kim, K.-O. (1974). Phonetic variation of Vietnamese tones in disyllabic utterances. *Journal of Phonetics, 2*, 223-232.

Harrison, P. (2000). Acquiring the phonology of lexical tone in infancy. *Lingua, 110*, 581-616.

Haugen, E., & Joos, M. (1972). Tone and intonation in East Norwegian. In D. Bolinger (Ed.), *Intonation* (pp. 414-436). Harmondsworth: Penguin Ltd.

Hillenbrand, J. (1983). Perceptual organization of speech sounds by infants. *Journal of Speech and Hearing Research, 26*(2), 268-282.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America, 97*(5), 3099-3111.

Hirsh-Pasek, K., Nelson, D. G. K., Jusczyk, P. W., Cassidy, K. W., Druss, B., & Kennedy, L. (1987). Clauses are perceptual units for young infants. *Cognition, 26*(3), 269-286.

Hombert, J. M. (1976). Perception of tones of bisyllabic nouns in Yoruba. *Studies in African Linguistics, Supplement 6*, 109-121.

Howie, J. M. (1974). On the domain of tone in Mandarin. *Phonetica, 30*, 129-148.

Howie, J. M. (1976). *Acoustical studies of Mandarin vowels and tones*. New York: Cambridge University Press.

Imada, T., Zhang, Y., Cheour, M., Taulu, S., Ahonen, A., & Kuhl, P. K. (2006). Infant speech perception activates Broca's area: a developmental magnetoencephalography study. *Infancy, 17*(10), 957-962.

Irwin, O. C. (1947). Infant speech: the problem of variability. *Journal of Speech Disorders, 12*, 173-176.

Jakobson, R., Fant, G., & Halle, M. (1967). *Preliminaries to speech analysis: The distinctive features and their correlates*. Cambridge: MIT Press.

Jenkins, J. J., Strange, W., & Edman, T. R. (1983). Identification of vowels in "vowelless" syllables. *Perception and Psychophysics, 34*(5), 441-450.

Johnson, K., & Mullennix, J. W. (1997). *Talker variability in speech processing*. New York: Academic Press.

Jusczyk, P. W. (1980). Discrimination of relative onset time of two-component tones by infants. *Journal of the Acoustical Society of America, 67*(1), 262-270.

Jusczyk, P. W. (1986). Towards a model for the development of speech perception. In J. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes*. Hillsdale: Erlbaum.

Jusczyk, P. W. (1993). From general to language-specific capacities: the WRAPSA model of how speech perception develops. *Journal of Phonetics, 21*, 3-28.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology, 29*(1), 1-23.

Jusczyk, P. W., Cutler, A., & Redanz, N. J. (1993). Infants' preference for the predominant stress patterns of English words. *Child Development, 64*(3), 675-687.

Jusczyk, P. W., & Derrah, C. (1987). Representation of speech sounds by young infants. *Developmental Psychology, 23*(5), 648-654.

Jusczyk, P. W., Friederici, A. D., Wessels, J. M. I., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language, 32*(3), 402-420.

Jusczyk, P. W., Goodman, M. B., & Baumann, A. (1999). Nine-month-olds' attention to sound similarities in syllables. *Journal of Memory and Language, 40*(1), 62-82.

Jusczyk, P. W., Houston, D., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology, 39*, 159-207.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language, 33*(5), 630-645.

Jusczyk, P. W., Pisoni, D. B., & Mullennix, J. W. (1992). Some consequences of stimulus variability on speech processing by 2-month-old infants. *Cognition, 43*(3), 253-291.

Jusczyk, P. W., Rosner, B. S., Reed, M. A., & Kennedy, L. J. (1989). Could temporal order differences underlie 2-month-olds' discrimination of English voicing contrasts? *Journal of the Acoustical Society of America, 85*(4), 1741-1749.

Kaplan, E. L. (1969). *The role of intonation in the acquisition of language.* Cornell University, New York.

Karlgren, B. (1962). *Sound and symbol in Chinese.* Hong Kong: Hong Kong University Press.

Kaski, S., Venna, J., & Kohonen, T. (1999). Coloring that reveals high-dimensional structures in data. In *Proceedings of the 6th International Conference on Neural Information Processing* (pp. 729-734). Australia: Perth.

Kaye, J., Lowenstamm, J., & Vergnaud, J.-R. (1985). The internal structure of phonological elements: a theory of charm and government. *Phonology Yearbook, 2*, 305-328.

Keating, P. A. (1985). CV phonology, experimental phonetics, and coarticulation. *UCLA Working Papers in Phonetics, 62*, 1-13.

Kelso, J. A. S. (1984). Phase transitions and critical behavior in human bimanual coordination. *American Journal of Physiology: Regulatory, Integrative and Comparative, 246*, R1000-R1004.

Kelso, J. A. S., Saltzman, E. L., & Tuller, B. (1986). The dynamical perspective on speech production: data and theory. *Journal of Phonetics, 14*, 29-59.

Kewley-Port, D. (1989). Time-varying features as correlates of place of articulation in stop consonants. *Journal of the Acoustical Society of America, 73*(1), 322-335.

Klein, D., Zatorre, R. J., Milner, B., & Zhao, V. (2001). A cross-linguistic PET study of tone perception in Mandarin Chinese and English speakers. *NeuroImage, 13*, 646-653.

Kluender, K. R., Diehl, R. L., & Killeen, P. R. (1987). Japanese quail can learn phonetic categories. *Science, 237*, 1195-1197.

Kluender, K. R., Lotto, A. J., Holt, L. L., & Bloedel, S. L. (1998). Role of experience for language-specific functional mappings of vowel sounds. *Journal of the Acoustical Society of America, 104*(6), 3568-3582.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics, 43*, 59-69.

Kohonen, T. (1989). *Self-organization and associative memory.* Berlin: Springer.

Kohonen, T. (1995). *Self-Organizing-Maps.* Berlin: Springer.

Krakow, R. A. (1999). Physiological organization of syllables: a review. *Journal of Phonetics, 27*(1), 23-54.

Kuhl, P. K. (1976a). Speech perception in early infancy: Perceptual constancy for vowel categories. *Journal of the Acoustical Society of America, 60*, S90.

Kuhl, P. K. (1976b). Speech perception in early infancy: The acquisition of speech sound categories. In S. K. Hirsh, D. H. Eldridge, I. J. Hirsh & S. R. Silverman (Eds.), *Hearing and Davis: Essays honoring Hallowell Davis*. St-Louis: Washington University Press.

Kuhl, P. K. (1977). Speech perception in early infancy: Perceptual constancy for the vowel categories /a/ and /ɔ/. *Journal of the Acoustical Society of America, 61*, S39.

Kuhl, P. K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *Journal of the Acoustical Society of America, 66*, 1668-1679.

Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior and Development, 6*(3), 263-285.

Kuhl, P. K. (1991). Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics, 50*(2), 93-107.

Kuhl, P. K., & Meltzoff, A. N. (1996). Infant vocalizations in response to speech: Vocal imitation and developmental change. *Journal of the Acoustical Society of America, 100*(4), 2425-2438.

Kuhl, P. K., & Miller, J. D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science, 190*, 69-72.

Kuhl, P. K., & Miller, J. D. (1982). Discrimination of auditory target dimensions in the presence or absence of variation in a second dimension by infants. *Perception and Psychophysics, 31*, 279-292.

Kuhl, P. K., & Padden, D. M. (1982). Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques. *Perception and Psychophysics, 32*, 542-550.

Kuhl, P. K., & Padden, D. M. (1983). Enhanced discriminability at the phonetic boundary for the place feature in macaques. *Journal of the Acoustical Society of America, 73*, 1003-1010.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science, 9*(2), F13-F21.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255*(5044), 606-608.

Ladefoged, P. (1980). What are linguistic sounds made of? *Language, 56*(3), 485-502.

Laniran, Y. O., & Clements, G. N. (2003). Downstep and high raising: Interacting factors in Yoruba tone production. *Journal of Phonetics, 31*, 203-250.

Lasky, R. E., Syrdal-Lasky, A., & Klein, R. E. (1975). VOT discrimination by four and six and a half month old infants from Spanish environments. *Journal of Experimental Child Psychology, 20*, 215-225.

Leather, J. (1983). Speaker normalization in perception of lexical tones. *Journal of Phonetics, 11*(4), 373-382.

Levitt, A., Jusczyk, P. W., Murray, J., & Carden, G. (1988). The perception of place of articulation contrasts in voiced and voiceless fricatives by two-month-old infants. *Journal of Experimental Psychology: Human perception and Performance, 14*(3), 361-368.

Liang, Z. A. (1963). Auditory perceptual cues in Mandarin tones. *Acta Physiologica Sinica, 26*, 85-91.

Liberman, A. M. (1970). Some characteristics of perception in the speech mode. *Perception and its Disorders, 48*, 238-254.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*(6), 431-461.

Liberman, A. M., Delattre, P., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America, 27*, 769-773.

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs, 68*(8), 13.

Liberman, A. M., Delattre, P. D., & Cooper, F. S. (1958). Some cues for the distinction between voiced and unvoiced stops in initial position. *Language and Speech, 1*, 153-167.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*(1), 1-36.

Lin, M. (1995). A perceptual study on the domain of tones in Standard Chinese. *Chinese Journal of Acoustics, 14*, 350-357.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America, 89*(2), 874-886.

Lotto, A. J. (2000). Language Acquisition as Complex Category Formation. *Phonetica, 57*(2-4), 189-196.

Lotto, A. J., Kluender, K. R., & Holt, L. L. (1997). Perceptual compensation for coarticulation by japanese quail (Coturnix coturnix japonica). *Journal of the Acoustical Society of America, 102*(2), 1134-1140.

MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. *Behavioral and brain sciences, 21*, 499-511.

Maddieson, I., & Hess, S. (1986). 'Tense' and 'lax' revisited: More on phonation type and pitch in minority languages in China. *UCLA Working Papers in Phonetics, 63*, 103-109.

Marean, G. C., Werner, L. A., & Kuhl, P. K. (1992). Vowel categorization by very young infants. *Developmental Psychology, 28*(3), 396-405.

Massaro, D. W., Cohen, M. M., & Tseng, C. (1985). The evaluation and integration of pitch height and pitch contour in lexical tone perception in Mandarin Chinese. *Journal of Chinese Linguistics, 13*, 267-290.

Mattock, K. J. (2004). *Perceptual reorganisation for tone: Linguistic tone and non-linguistic pitch perception by English language and Chinese language infants.* Unpublished doctoral dissertation, University of Western Sydney, Sydney.

Mattock, K. J., & Burnham, D. (2006). Chinese and English infants' tone perception: Evidence for perceptual reorganization. *Infancy, 10*(3).

Maye, J., & Weiss, D. (2003). Statistical cues facilitate infants' discrimination of difficult phonetic contrasts. In B. Beachley, A. Brown & F. Conlin (Eds.), *Proceedings of the 27th Annual Boston University Conference on Language Development* (pp. 508-518). Sommerville: Cascadilla Press.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*(3), B101-B111.

McMurray, B., & Aslin, R. N. (2005). Infants are sensitive to within-category variation in speech perception. *Cognition, 95*, B15-B26.

Mehler, J., Dupoux, E., & Segui, J. (1990). Constraining models of lexical access: The onset of word recognition. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 236-262). Cambridge: MIT Press.

Ménard, L., Schwartz, J.-L., & Boë, L.-J. (2004). The role of vocal tract morphology in speech development: Perceptual targets and sensori-motor maps for French synthesized vowels from birth to adulthood. *Journal of Speech, Language and hearing research, 47*(5), 1059-1080.

Ménard, L., Schwartz, J.-L., Boë, L.-J., Kandel, S., & Vallée, N. (2002). Auditory normalization of French vowels synthesized by an articulatory model

simulating growth from birth to adulthood. *Journal of the Acoustical Society of America, 111*(4), 1892-1905.

Menon, K. M., Rao, P. V., & Thosar, R. B. (1974). Formant transitions and stop consonant perception in syllables. *Language and Speech, 17*(1), 27-46.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America, 85*(5), 2114-2134.

Miller, J. L., & Eimas, P. D. (1979). Organization in infant speech perception. *Canadian Journal of Psychology, 33*(4), 353-367.

Moore, C. B., & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *Journal of the Acoustical Society of America, 102*(3), 1864-1877.

Morse, P. A. (1972). The discrimination of speech and nonspeech stimuli in early infancy. *Journal of Experimental Child Psychology, 14*(3), 477-492.

Nazzi, T., Floccia, C., & Bertoncini, J. (1998). Discrimination of pitch contours by neonates. *Infant Behavior and Development, 21*, 779-784.

Nazzi, T., Nelson, D. G. K., Jusczyk, P. W., & Jusczyk, A. M. (2000). Six-month-olds' detection of clauses embedded in continuous speech: Effects of prosodic well-formedness. *Infancy, 1*(1), 123-147.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America, 85*(5), 2088-2113.

Needham, A., Dueker, G., & Lockhead, G. (2005). Infants' formation and use of categories to segregate objects. *Cognition, 93*(3), 215-240.

Nelson, W. L. (1983). Physical principles for economies of skilled movements. *Biological Cybernetics, 46*(2), 135-147.

Ohala, J. J., & Ewan, W. G. (1973). Speed of pitch change. *Journal of the Acoustical Society of America, 53*(345).

Ohde, R. N. (1988). Revisiting stop-consonant perception for two-formant stimuli. *Journal of the Acoustical Society of America, 84*(4), 1551-1555.

Ohman, S. E. G. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America, 39*(1), 151-168.

Olsho, L. W., Koch, E. G., & Halpin, C. F. (1987). Level and age effects in infant frequency discrimination. *Journal of the Acoustical Society of America, 82*, 454-464.

Pantev, C., Bertrand, O., Eulitz, C., Verkindt, C., Hampson, S., Schuierer, G., et al. (1994). Specific tonotopic organizations of different areas of the human

auditory cortex revealed by simultaneous magnetic and electric recordings. *Electroencephalography and Clinical Neurophysiology, 94*, 26-40.

Peng, S.-h. (2000). Lexical versus 'phonological' representations of Mandarin Sandhi tones. In M. B. Broe & J. B. Pierrehumbert (Eds.), *Papers in laboratory phonology V: Acquisition and the lexicon* (pp. 152-167). Cambridge: Cambridge University Press.

Perkell, J. S., & Klatt, D. H. (1986). *Invariance and variability in speech processes.* Hillsdale: Lawrence Erlbaum.

Peters, A. (1997). Language typology, individual differences and the acquisition of grammatical morphemes. In D. Slobin (Ed.), *Cross-linguistic perspectives in language acquisition* (Vol. 4). Hillsdale: Erlbaum.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24*(2), 175-184.

Pierrehumbert, J. B. (2002). Word-specific phonetics. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology VII* (pp. 101-140). Berlin: Mouton de Gruyter.

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech, 46*(2-3), 115-154.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception and Psychophysics, 13*, 253-260.

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*(4), 1352-1361.

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson & J. W. Mullennix (Eds.), *Speaker variability in speech processing* (pp. 9-32). New York: Academic Press.

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance, 20*(2), 421-435.

Pons, F. (2006). The effects of distributional learning on rats' sensitivity to phonetic information. *Journal of Experimental Psychology: Animal Behavior Processes, 32*(1), 97-101.

Port, R. F. (1981). Linguistic timing factors in combination. *Journal of the Acoustical Society of America, 69*, 262-274.

Prince, A., & Smolensky, P. (1993). Optimality theory: Constraint interaction in generative grammar. In *Rutgers University Center for Cognitive Science Technical Report 2.*

Prom-on, S., Xu, Y., & Thipakorn, B. (2009). Modeling tone and intonation in Mandarin and English as a process of target approximation. *Journal of the Acoustical Society of America, 125*, 405-424.

Ritter, H., & Schulten, K. (1986). On the stationary state of Kohonen's self-organizing sensory mapping. *Biological Cybernetics, 54*, 99-106.

Seldon, H. L. (1985). The anatomy of speech perception: Human auditory cortex. In A. Peters & E. G. Jones (Eds.), *Cerebral cortex* (Vol. 4, pp. 273-327). New York: Plenum.

Shen, X. S. (1990). Tonal coarticulation in Mandarin. *Journal of Phonetics, 18*, 281-295.

Shen, X. S., & Lin, M. (1991). A perceptual study of Mandarin tones 2 and 3. *Language and Speech, 34*(2), 145-156.

Shi, R., Morgan, J. L., & Allopenna, P. (1998). Phonological and acoustic bases for earliest grammatical category assignment: A cross-linguistic perspective. *Journal of Child Language, 25*(1), 169-201.

Shih, C. (1993). Relative prominence of tonal targets. In *Proceedings of The 5th North American Conference on Chinese Linguistics*. Newark, Delaware: University of Delaware: 36.

Singh, L. (2003). *Variability and constancy in infants' formation of lexical categories.* Brown University, Providence, RH.

Sinnott, J. M., & Aslin, R. N. (1985). Frequency and intensity discrimination in human infants and adults. *Journal of the Acoustical Society of America, 78*, 1986-1992.

Siqueland, E. R., & Delucua, C. A. (1969). Visual reinforcement of nonnutritive sucking in human infants. *Science, 165*(3898), 1144-1146.

Soderstrom, M., Seidl, A., Nelson, D. G. K., & Jusczyk, P. W. (2003). The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language, 49*, 249-267.

Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In E. E. David & P. B. Denes (Eds.), *Human communication: A unified view* (pp. 51-66). New York: McGraw-Hill.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics, 17*, 3-45.

Stevens, K. N. (1993). Inferring articulatory movements from acoustic data. *Journal of the Acoustical Society of America, 93*, 2416.

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America, 64*(5), 1358-1368.

Strange, W. (1987). Information for vowels in formant transitions. *Journal of Memory and Language, 26*(5), 550-557.

Strange, W. (1989). Dynamic specification of coarticulated vowels spoken in sentence context. *Journal of the Acoustical Society of America, 85*(5), 2135-2153.

Strange, W., Verbrugge, R. R., Shankweiler, D. P., & Edman, T. R. (1976). Consonant environment specifies vowel identity. *Journal of the Acoustical Society of America, 60*(1), 213-224.

Streeter, L. A. (1976). Language perception of 2-month-old infants shows effects of both innate mechanisms and experience. *Nature, 259*, 39-41.

Sun, X. (2002). The determination, analysis, and synthesis of fundamental frequency (Doctoral dissertation, Northwestern University, 2002). *Dissertation Abstracts International B, 63*(11), 5195.

Sundberg, J. (1979). Maximum speed of pitch changes in singers and untrained subjects. *Journal of Phonetics, 7*, 71-79.

Sussman, H. M., McCaffrey, H. A., & Matthews, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America, 90*(3), 1309-1325.

Swoboda, P. J., Morse, P. A., & Leavitt, L. A. (1976). Continuous vowel discrimination in normal and at risk infants. *Child development, 47*, 459-465.

Thorpe, L. A. (1986). *Infants categorise rising and falling pitch*. Paper presented at the International Conference of Infant Studies.

Trehub, S. E. (1973). Infants' sensitivity to vowel and tonal contrasts. *Developmental Psychology, 9*(1), 91-93.

Trehub, S. E. (1976). The discrimination of foreign speech contrasts by infants and adults. *Child Development, 47*(2), 466-472.

Trehub, S. E., Bull, D., & Thorpe, L. A. (1984). Infants' perception of melodies: The role of melodic contour. *Child Development, 55*(3), 821-830.

Trehub, S. E., Thorpe, L. A., & Morrongiello, B. A. (1987). Organizational processes in infants' perception of auditory patterns. *Child Development, 58*(3), 741-749.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, A. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences, 104*(33), 13273-13278.

van de Weijer, J. (1998). *Language input for word discovery.*

Varfis, A. (1993). On the use of two traditional statistical techniques to improve the readability of Kohonen Maps. *Proceedings NATO ASI Workshop Statistics Neural Networks.*

Verbrugge, R. R., Strange, W., Shankweiler, D. P., & Edman, T. R. (1976). What information enables a listener to map a talker's vowel space? *Journal of the Acoustical Society of America, 60*, 198.

Vihman, M. M. (1986). Ontogeny of phonetic gestures: Speech production. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor-theory of speech perception* (pp. 69-84). Hillsdale: Erlbaum.

Vihman, M. M. (1993). The construction of a phonological system. In B. deBoysson-Bardies, S. deSchonen, P. W. Jusczyk, P. MacNeilage & J. Morton (Eds.), *Developmental Neurocognition: Speech and face processing in the first year of Life* (pp. 411-419). Dordrecht: Kluwer.

Vihman, M. M., & Miller, R. (1988). Words and babble at the threshold of language acquisition. In M. D. Smith & J. L. Locke (Eds.), *The emergent lexicon* (pp. 151-183). New York: Academic.

Vouloumanos, A., & Werker, J. F. (2004). Tuned to the signal: the privileged status of speech for young infants. *Developmental Science 7*(3), 270-276.

Wang, W. S.-Y. (1967). Phonological features of Tone. *International Journal of American Linguistics, 33*(2), 93-105.

Wang, W. S.-Y., & Li, K.-P. (1967). Tone 3 in Pekinese. *Journal of Speech and Hearing Research, 10*, 629-636.

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development, 1*(2), 197-234.

Werker, J. F., & Lalonde, C. E. (1988). Cross-language speech perception: Initial capabilities and developmental change. *Developmental Psychology, 24*(5), 672-683.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7*(1), 49-63.

Wessinger, C. M., Buonocore, M. H., Kussmaul, C. L., & Mangun, G. R. (1997). Tonotopy in human auditory cortex examined with functional magnetic resonance imaging. *Human brain mapping, 5*, 18-25.

Whalen, D. H., & Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica, 49*, 25-47.

Xu, C. X., Xu, Y., & Sun, X. (2003). Effects of consonant aspiration on mandarin tones. *Journal of the International Phonetic Association, 33*, 165-181.

Xu, Y. (1994). Production and perception of coarticulated tones. *Journal of the Acoustical Society of America, 95*(4), 2240-2253.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics 25*, 61-83.

Xu, Y. (1998). Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica, 14*, 350-357.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics, 27*, 55-105.

Xu, Y. (2001). Sources of tonal variations in connected speech. *Journal of Chinese Linguistics, monograph series, 17*, 1-31.

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication, 46*, 220-251.

Xu, Y., & Liu, F. (2006). Tonal alignment, syllable structure and coarticulation: Toward an integrated model. *Italian Journal of Linguistics, 18*, 125-159.

Xu, Y., & Liu, F. (2007). Determining the temporal interval of segments with the help of F0 contours. *Journal of Phonetics, 35*, 398-420.

Xu, Y., & Sun, X. (2002). Maximum speed of pitch change and how it may relate to speech. *Journal of the Acoustical Society of America, 111*, 1399-1413.

Xu, Y., & Wang, Q. E. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication 33*, 319-337.

Xu, Y., & Xu, C. X. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics, 33*, 159-197.

Yip, M. (1989). Contour tones. *Phonology, 6*, 149-174.

Yip, M. (2002). *Tone*. Cambridge: Cambridge University Press.

Zemlin, W. R. (1988). *Speech and hearing sciences: Anatomy and physiology*. Englewood Cliffs, NJ: Prentice-Hall.