SÉGRÉGATION DE VOYELLES SIMULTANÉES: EFFETS DU NIVEAU RELATIF ET DE LA DIFFÉRENCE DE F_0

Alain de CHEVEIGNÉ

Laboratoire de Linguistique Formelle, CNRS/Université Paris 7, 2 place Jussieu, 75251, Paris

Tél.: +33 1 44273633, email: alain@linguist.jussieu.fr

ABSTRACT

Subjects were presented with stimuli consisting of single or "double" (concurrent) vowels. They had to decide whether each stimulus contained one or two vowels, and which vowels they were. In the case of double vowels, constituents had either the same fundamental frequency (F_0) or F_0s differing by 6%. They had either the same RMS level, or levels differing by 10 or 20 dB. The average number of vowels reported and the identification rate of each constituent were measured for each condition. When F_0 differed by 6%, subjects answered two vowels more often than at unison. Identification accuracy was also better than at unison when the target was at 0, -10 or -20 dB relative to the competing vowel. For stronger targets (10 or 20 dB), identification was almost perfect and therefore little affected by F_0 differences. These results are compatible with the hypothesis that segregation occurs according to a mechanism of *harmonic cancellation* rather than harmonic enhancement. They are incompatible with recent models of vowel segregation.

1. INTRODUCTION

Pour séparer perceptivement les sons, le système auditif utilise, entre autres indices, les différences de fréquence fondamentale (ΔF_0). Brokx et Nooteboom (1982) ont trouvé que l'identification de phrases concurrentes (naturelles ou synthétiques) était meilleure lorsqu'elles étaient prononcées sur des tons différents qu'à l'unisson. Le même avantage a été trouvé pour l'identification de paires de voyelles synthétiques stationnaires (Scheffers, 1983; Assmann & Summerfield 1990; Culling & Darwin 1993). De nombreux modèles et méthodes de séparation de sons harmoniques ont été proposés pour expliquer ce phénomène, ou pour le reproduire dans des systèmes de traitement de la parole (voir de Cheveigné 1993 pour une revue).

Parmi eux, le modèle de Meddis et Hewitt (1992) est accepté comme le plus plausible. Ce modèle comprend un banc de filtres dont chaque canal est traité par un modèle de cellule ciliée produisant une probabilité de décharge nerveuse. La fonction d'autocorrélation (ACF) de cette probabilité est calculée, et les ACF de tous les canaux sont additionnées pour former une ACF globale. Le maximum de l'ACF globale (dans une gamme de délais) sert à définir la *période dominante* de la réponse. Le modèle effectue alors une partition des canaux entre ceux qui sont dominés par cette période (c.a.d dont l'ACF y présente un pic), et les autres. La somme des ACF du premier groupe de canaux représente la "voyelle dominante", et la somme des ACF des canaux restants représente la "voyelle dominée". Après cette étape de ségrégation, chaque voyelle est identifiées par comparaison de la partie de l'ACF globale comprise entre 0 et 4,5 ms à des patrons de référence.

Le fonctionnement du modèle de Meddis et Hewitt dépend ainsi d'une partition de la population de canaux selon leur périodicité. Si tous étaient dominés par une seule période (par exemple si une voyelle était plus intense que l'autre), le modèle ne fonctionnerait pas et on ne constaterait alors aucun effet de ΔF_0. L'expérience décrite ici explore cette possibilité en faisant varier le niveau relatif entre voyelles. Un autre objectif était

de trouver un niveau relatif qui permette de s'affranchir des effets de plafond souvent constatés dans les expériences de "voyelles doubles": Lorsque leur niveau est égal, l'identification des deux voyelles est souvent quasi-parfaite, et donc insensible aux paramètres d'expérience. Le déséquilibre de niveau rend la tâche plus difficile pour une des voyelles, et améliore ainsi la sensibilité.

2. MÉTHODES

Des voyelles japonaises isolées ou concurrentes (somme de deux voyelles) furent présentées à des sujets japonais qui devaient juger à chaque fois s'ils entendaient une ou deux voyelles, et lesquelles. Les six sujets étaient informés que chaque stimulus comprenait une ou deux voyelles distinctes, mais ne recevaient aucun feedback.

Les cinq voyelles (/a/, /e/, /i/, /o/, /u/) furent synthétisées à une fréquence d'échantillonage de 16 kHz avec des F_0 de 125 et 132,5 Hz. Les voyelles doubles furent obtenues en formant toutes les combinaisons de F_0 (pour produire deux ΔF_0: 0 et 6%) et de voyelles distinctes (10 paires), et en additionnant les constituants avec des niveaux RMS relatifs de -20, -10, 0, 10 et 20 dB. Les voyelles doubles étaient ainsi au nombre de (10 paires) x (2 ΔF_0) x (2 ordres de F_0) x (5 niveaux relatifs) x (3 répétions) = 600 paires, auxquelles furent ajoutés 240 voyelles simples, pour un total de 840 stimuli par session. Les voyelles simples servaient à rendre l'ensemble conforme à la description faite aux sujets, et à repérer d'éventuels effets des paramètres de synthèse sur la qualité des voyelles. Après ajustement à un niveau RMS standard, les stimuli furent présentées aux sujets via casque, à un niveau sonore compris entre 63 et 70 dBA. Chaque sujet participa à cinq sessions.

Pour toutes les conditions, le nombre moyen de réponses par stimulus et le taux d'identificaton furent calculés. Chaque voyelle isolée fut jugée correctement identifiée si la réponse (une ou deux voyelles) comprenait le nom de cette voyelle. Chaque constituant de voyelle double fut jugée correctement identifiée si la réponse (une ou deux voyelles) comprenait le nom du constituant. Ces réponses furent classées selon la nature du constituant (phonème, F_0), la nature de la voyelle concurrente, et leur relation (ΔF_0, niveau relatif). Cette technique d'analyse en terme de taux d'identification des *constituants* diffère de celle utilisée dans les expériences classiques, qui mesurent généralement un taux d'identification de *paires* (deux voyelles correctes).

3. RÉSULTATS

Les résultats présentés ici sont moyennés sur les facteurs *sujet*, *session*, F_0, et *paire*. À l'unisson (Fig. 1, trait épais), les sujets tendent a faire une réponse double lorsque les voyelles sont de même niveau. Lorsque l'une domine l'autre, les réponses doubles sont plus rares, mais leur nombre reste appréciable même pour les voyelles isolées (à droite). À ΔF_0=6% (trait fin), les réponses doubles sont plus nombreuses quel que soit le niveau.

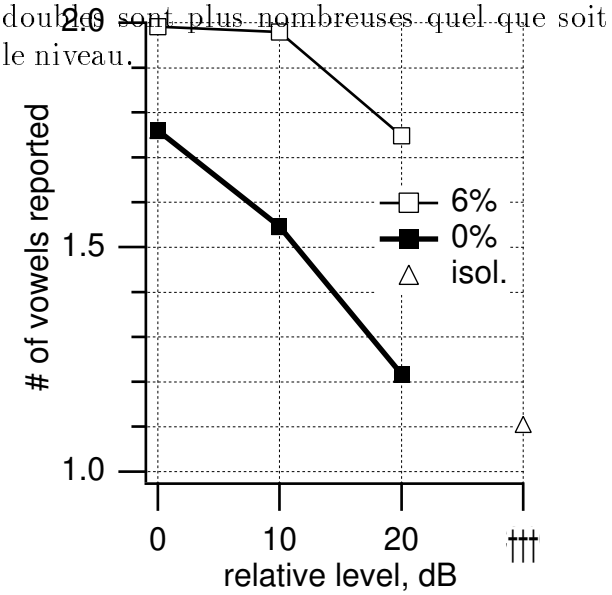

Fig. 1 *Nombre moyen de voyelles répondues par stimulus en fonction du niveau relatif, pour deux valeurs du ΔF_0. Triangle à droite: voyelles isolées.*

À l'unisson, l'identification d'une voyelle est d'autant meilleure qu'elle domine en niveau sa concurrente (Fig. 2, trait épais). À ΔF_0=6% (trait fin), le taux d'identification est plus élevé qu'à l'unisson, en particulier lorsque la cible est faible par rapport à la voyelle concurrente. L'effet de ΔF_0 est plus fort à -10 dB qu'aux autres niveaux. L'obtention d'effets expérimentaux relativement

forts est d'un intérêt pratique, puisqu'ils sont plus faciles à mettre en évidence de façon statistiquement robuste. Cet avantage serait cependant nul si la *variabilité* était également plus élevée.

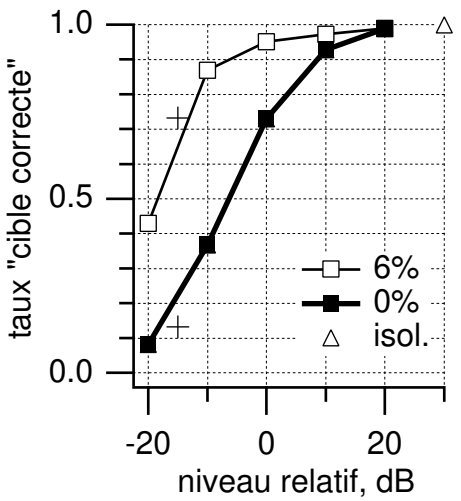

Fig. 2 *Taux d'identification d'une voyelle cible en fonction du niveau relatif à la voyelle concurrente, pour deux valeurs de ΔF_0. Croix: taux obtenus à -15 dB dans une autre expérience avec les mêmes sujets (de Cheveigné et al. 1996a). Triangle à droite: voyelles isolées.*

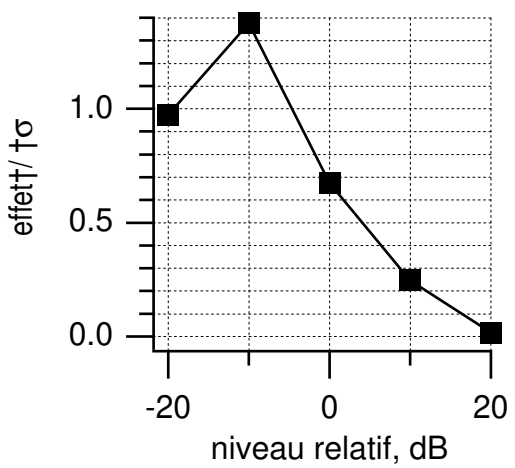

Fig. 3 *Rapport entre la taille de l'effet de ΔF_0 et sa déviation standard, en fonction du niveau relatif.*

La Fig. 3 montre qu'il n'en est rien: le rapport de la taille de l'effet (différence des taux d'identification à 6 et 0 %) à sa déviation standard est bien plus élevé à -10 dB qu'à d'autres valeurs du niveau relatif. À l'inverse un effet de ΔF_0sur une cible dominante sera difficile a mettre en evidence de façon statististiquement robuste.

4. DISCUSSION

Nos sujets pouvaient répondre une ou deux voyelles par stimulus (à la différence des expériences classiques citées en Introduction, où le sujet devait obligatoirement répondre deux voyelles). Le nombre moyen de réponses est une mesure intéressante, reflétant une perception de la "multiplicité" des sources. À l'unisson ce nombre est élevé lorsque les voyelles sont de même niveau (et donc que le stimulus ne ressemble à aucune voyelle simple). Il est également élevé à tous niveaux lorsqu'il y a un ΔF_0. Notre nouvelle procédure affecte aussi les taux d'identification, et tend à produire des effets de ΔF_0 plus marqués que dans les expériences classiques (de Cheveigné et al. 1996a).

L'identification dépend peu de ΔF_0 lorsque la cible est forte (10 ou 20 dB), du fait de l'effet de plafond. En revanche la dépendance est marquée lorsque la cible est faible (-10 ou -20 dB). Ce résultat suggère que la ségrégation s'opère selon un mécanisme d'*annulation harmonique*, selon lequel la structure harmonique (F_0) de la voyelle concurrente facilite son élimination (de Cheveigné 1993; de Cheveigné et al. 1995). En effet, le F_0 de la voyelle concurrente est facile à estimer lorsque la cible est faible. Cette hypothèse est confortée par d'autres résultats (Lea 1992; Summerfield & Culling 1992; de Cheveigné et al. 1996b). Une hypothèse rivale est que la structure harmonique de la cible faciliterait son identification (hypothèse de *renforcement harmonique*). Nous pouvons l'écarter dans les conditions où la cible est faible (-10 ou -20 dB), puisque son F_0 serait alors particulièrement difficile à estimer. Lorsque la cible est forte, l'effet de plafond nous empêche de conclure à un quelconque effet de renforcement harmonique puisque l'identification est déja parfaite sans le concours du ΔF_0. Jusqu'à présent, peu de résultats expérimentaux confortent l'hypothèse de renforcement harmonique, qui pourtant inspire nombre de méthodes d'élimination de voix parasites...

Dans l'introduction, nous avons suggéré que le modèle de ségrégation de voyelles de

Meddis et Hewitt (1992) pourrait ne pas fonctionner si une voyelle était trop dominante. Le modèle fut implémenté et appliqué à nos stimulis dans le cas ΔF_0=6%. Pour des niveaux relatifs modérés (0 ou 10 dB), la partition des canaux s'effectue bien comme prévu. En revanche à 20 dB de différence de niveau inter-voyelles, et pour certaines paires, *tous* les canaux sont dominés par la même périodicité, empêchant ainsi toute partition. Faute de partition, on ne devrait constater aucun effet bénéfique de ΔF_0sur l'identification. Pourtant nos résultats ont révélé un effet marqué pour ces mêmes paires, que le modèle de Meddis et Hewitt ne peut donc pas expliquer. Un modèle pouvant les expliquer est proposé dans de Cheveigné (1996).

5. CONCLUSION

L'identification d'une voyelle synthétique accompagnée d'une voyelle concurrente est meilleure lorsque leurs F_0 diffèrent de 6% plutôt que 0%. L'effet est le plus robuste lorsque le niveau de la voyelle cible est de -10 dB par rapport à la voyelle concurrente, et reste marqué à -20 dB, résultat que le modèle accepté comme le plus plausible, celui de Meddis et Hewitt (1992), ne peut pas prédire.

Le nombre moyen de voyelles répondues par stimulus est plus élévé lorsque les F_0 sont différents. La différence de F_0 joue le rôle d'un indicateur de *multiplicité* des sources.

6. REMERCIEMENTS

Ce travail a été conduit dans le cadre d'un accord de collaboration entre ATR Human Information Processing Laboratories, le CNRS, et l'Université Paris 7. Merci à ATR pour son hospitalité, et au CNRS pour l'autorisation d'absence. S. McAdams et C. Marin ont participé à la préparation des expériences, H. Kawahara, M. Tsuzaki et K. Aikawa ont aidé à leur élaboration, R. Kubo a assisté à leur conduite, et J. Culling a fourni les programmes de synthèse.

7. BIBLIOGRAPHIE

Assmann, P. F. and Summerfield, Q. (1990). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies." *J. Acoust. Soc. Am.*, 88, 680-697.

Brokx, J. P. L. and Nooteboom, S. G. (1982). "Intonation and the perceptual separation of simultaneous voices," *Journal of Phonetics* 10, 23-36.

Culling, J. F. and Darwin, C. J. (1993). "Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0," *J. Acoust. Soc. Am.*, 93, 3454-3467.

de Cheveigné, A. (1993). "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," *J. Acoust. Soc. Am.*, 93, 3271-3290.

de Cheveigné, A. (1996). "Concurrent vowel segregation III: a neural time-domain model of harmonic interference cancellation," *J. Acoust. Soc. Am.*, en préparation.

de Cheveigné, A., McAdams, S., Laroche, J. and Rosenberg, M. (1995). "Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement," *J. Acoust. Soc. Am.*, 97, 3736-3748.

de Cheveigné, A, Kawahara, H, Tsuzaki, M. and Aikawa, K. (1996a). "Concurrent vowel segregation I: effects of relative level and F0 difference," *J. Acoust. Soc. Am.*, en préparation.

de Cheveigné, A., McAdams, S., Marin, M. (1996b). "Concurrent vowel segregation II: effects of phase, harmonicity and task," *J. Acoust. Soc. Am.*, en préparation.

Lea, A. (1992). "*Auditory models of vowel perception*," unpublished doctoral dissertation, Nottingham.

Meddis, R. and Hewitt, M. J. (1992). "Modeling the identification of concurrent vowels with different fundamental frequencies," *J. Acoust. Soc. Am.*, 91, 233-245.

Scheffers, M. T. M. (1983). "*Sifting vowels*," unpublished doctoral thesis, Gröningen.

Summerfield, Q. and Culling, J. F. (1992). "Periodicity of maskers not targets determines ease of perceptual segregation using differences in fundamental frequency". 124th meeting of the ASA [*J. Acoust. Soc. Am.* 92, 2317 (A)].