Titre: Identification de voyelles simultanées harmoniques et inharmoniques.

(ancien titre: Identification of concurrent harmonic and inharmonic vowels)

Auteurs:

Alain de Cheveigné
Laboratoire de Linguistique Formelle (CNRS/Université Paris 7), case 7003, 2 place Jussieu, F-75251 Paris Cédex 05, France, tel. (33-1) 44273633, fax (33-1) 44277919, email: alain@linguist.jussieu.fr

Stephen McAdams
Laboratoire de Psychologie Expérimentale (CNRS), Université René Descartes, 28 rue Serpente, F-75006 Paris, France and IRCAM, 1 place Stravinsky, F-75004 Paris, France, tel (33-1) 40519848, fax 40517085, email: smc@ircam.fr

Jean Laroche et Muriel Rosenberg
Département Signal, Télécom Paris (ENST/CNRS), 46 rue Barrault, F-75634 Paris Cédex 13, France, tel. (33-1) 45817862, fax (33-1) 45887935, email laroche@enst.fr, rosenber@enst.fr.

résumé: It is known that the auditory system uses harmonicity cues to separate concurrent voiced sounds that differ in fundamental frequency (F0). What is less clear is whether the harmonicity of the target sound is used to enhance it within the mixture, or whether that of the background is used to cancel it and thus allow the target to emerge. An experiment was designed to answer this question. Subjects were presented with pairs of concurrent synthetic vowels, each of which was either harmonic or inharmonic. Results were scored according to the harmonicity of the vowel identified (the target), and that of the second vowel (the ground). For a given target, identification was best for harmonic grounds, except when target and ground were both harmonic and had the same F0. This is compatible with the cancellation hypothesis. On the other hand, identification for a given ground was worse when the target was harmonic. This is the opposite of the effect predicted by the enhancement hypothesis.

1. INTRODUCTION

Notre environnement produit des sons multiples qui se chevauchent dans les domaines temporel et spectral. Le système auditif organise les corrélats de cette information acoustique en des ensembles qu'il attribue à chaque source. Cela a sans doute contribué à la survie des ancêtres de notre espèce longtemps avant l'apparition du langage ou de la musique.

Les informations qui contribuent à la séparation des sources sonores sont aussi bien acoustiques que cognitives: disparité binaurale, différences de fréquence fondamentale (F0), modulation de fréquence, asynchronisme d'attaque, lexique, attention, contexte, etc. [1-5]. Parmi elles, la différence de F0 a reçu beaucoup d'attention. Des expériences psychoacoustiques ont montré que l'identification de voyelles synthétiques mélangées est meilleure lorsque leurs F0 diffèrent [6-13]. Divers modèles ont été proposés, ainsi que des méthodes et des algorithmes de séparation à usage de traitement de la parole (voir [14] pour une revue). Certains utilisent l'harmonicité de la voix à identifier pour la faire ressortir (*renforcement harmonique*), d'autres celle de l'autre voix pour l'éliminer (*annulation harmonique*), d'autres enfin les deux. Chaque stratégie a ses avantages et inconvénients [14,15,19].

La question se pose de savoir si l'une, l'autre ou les deux stratégies sont utilisées par le système auditif. Lea [16] a présenté à des sujets des paires de voyelles dont chacune était soit voisée, soit chuchotée, en leur demandant de les identifier. L'identification d'une voyelle était meilleure lorsque l'*autre voyelle* était voisée. Cela conforte l'hypothèse d'annulation. En revanche elle n'était pas meilleure lorsque la voyelle cible était elle-même voisée, ce qui infirme l'hypothèse du renforcement harmonique. Cependant avec une méthode un peu différente [17,18] Lea a abouti à un résultat plus nuancé.

L'expérience de Lea suppose que les voyelles voisées et chuchotées sont équivalentes en termes de qualité phonétique et de pouvoir masquant, sauf pour ce qui dépend de leur structure harmonique. Or il n'existe pas une façon univoque d'apparier le spectre discret d'une voyelle voisée au spectre continu d'une voyelle chuchotée. Lea s'est appuyé pour cela sur un modèle d'excitation au niveau de la membrane basilaire. On ne peut exclure qu'un autre modèle, ou d'autres paramètres, n'aboutissent à des résultats différents. Pour cette raison nous avons utilisé ici, à la place des voyelles chuchotées, des voyelles inharmoniques dont la structure et la densité spectrale sont proches de celle d'une voyelle harmonique.

2. METHODES

2.1. Stimuli et présentation

Les voyelles utilisées (/a/, /e/, /i/, /o/, /u/ du français) étaient représentées par des allophones pris parmi un

ensemble de dix pour chaque voyelle, extraits de la base de données de parole naturelle du GRECO par une procédure d'échantillonnage, analyse, re-synthèse et sélection par écoute. Les stimuli ont été créés par synthèse additive. Les fréquences des partiels d'une voyelle inharmonique ont été obtenues en modifiant chaque fréquence d'une série harmonique par une quantité aléatoire dans une plage qui est le plus petit de ± 3% ou de la moitié de la distance entre harmoniques adjacents. La F0 nominale d'une série inharmonique est, par définition, la F0 de la série harmonique dont elle est issue. Les amplitudes des partiels ont été calculées de façon à préserver l'enveloppe spectrale de l'allophone. Chaque partiel démarrait en phase sinus. Les stimuli débutaient et finissaient par des rampes cosinusoidales de 25 ms, leur durée totale totale rampe comprise étant de 200 ms.

Les voyelles ont été présentées par paires en évitant les paires de voyelles identiques (10 paires possibles). Elles pouvaient avoir la même F0 (125 Hz) ou une F0 différente (125 Hz ± 1/4 ton), et chacune pouvait être soit harmonique, soit inharmonique (8 conditions d'harmonicité et de ΔF0). Pour les voyelles inharmoniques, une série différente a été utilisée pour les différents allophones: au sein de chaque paire contenant deux voyelles inharmoniques, les séries de partiels étaient donc différentes.

On ne peut exclure que la qualité phonétique ou le pouvoir masquant d'une voyelle dépendent du choix de l'allophone, du choix du pattern inharmonique, de la F0, ou de l'ordre de présentation. Pour éviter tout biais systématique, les précautions suivantes ont été prises:
– chaque condition d'harmonicité et de ΔF0 a été représentée par le même ensemble d'allophones,
– pour les voyelles inharmoniques, le même pattern inharmonique servait aux différentes F0 d'un allophone,
– chaque condition a été doublée pour être représentée dans l'ordre de F0 haut-bas et bas-haut,
– l'appariement des allophones et l'ordre des stimuli ont été choisis au hasard pour chaque présentation et chaque sujet.

Trente sujets ont participé à l'expérience. Les 160 conditions ont été présentées trois fois dans un ordre aléatoire, via casque et à un niveau de 60 dBA environ. Après chaque présentation, le sujet devait répondre deux voyelles dans un ordre quelconque.

2.3. Dépouillement des réponses

Les réponses ont été dépouillées de la façon suivante. Pour chaque voyelle présente dans le stimulus, la réponse a été jugée correcte si le nom de cette voyelle figurait dans la paire répondue par le sujet. Cette réponse partielle a été classée selon l'harmonicité de la voyelle (la cible), celle de l'autre voyelle (le fond) et la différence de F0. La procédure a été répétée avec l'autre constituant du stimulus, en renversant les rôles de cible et de fond. Les conditions après dépouillement (et leur notation) sont schématisées dans la Figure 1.

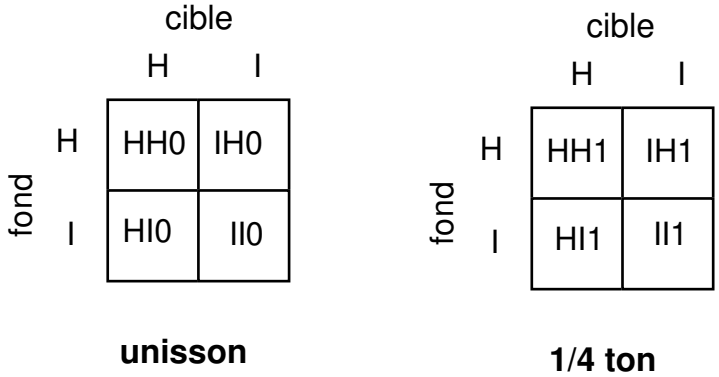

Fig. 1 Les huit conditions de ΔF0 et d'harmonicité de la cible et du fond.

Les résultats ont été soumis à une analyse de variance (ANOVA)avec mesures répétées. Les interactions entre facteurs étant significatives et importantes, nous présentons séparément chaque effet partiel.

3. RESULTATS

Les résultats sont présentés dans les figures 2 et 3 sous deux formats différents. Lorsque cible et fond sont harmoniques, l'identification de la cible est d'environ 6% meilleure lorsqu'il y a une différence de F0. Pour les paires contenant une voyelle inharmonique, il n'y a en revanche aucun effet significatif de ΔF0 (Fig. 2).

Lorsque la cible est inharmonique (Fig. 3, traits pleins), son identification est meilleure d'environ 3%

lorsque le fond est harmonique. Il en est de même pour une cible harmonique, à condition qu'il y ait une différence de F0. En revanche, à l'unisson l'identification est meilleure si le fond est inharmonique.

Quels que soient le ΔF0 ou la nature du fond, l'identification de la cible est meilleure lorsqu'elle est inharmonique (Fig. 3, traits pleins). L'avantage atteint 8% pour un fond harmonique et un ΔF0 nul, il est d'environ 3% pour les autres conditions.

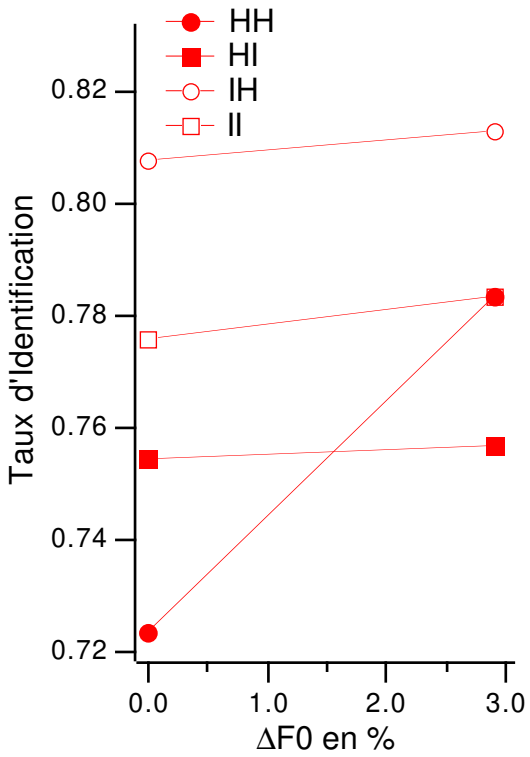

Fig. 2. Taux d'identification de la cible en fonction de ΔF0 pour les quatre conditions d'harmonicité cible/fond.

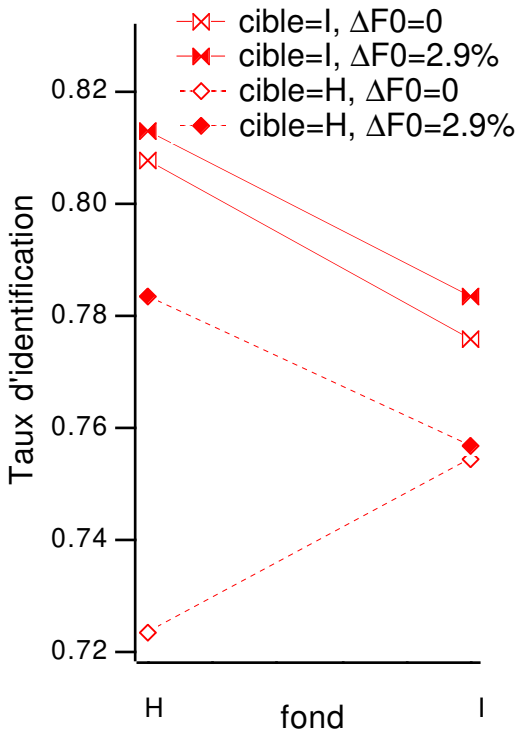

Fig. 3. Taux d'identification en fonction de la nature du fond pour une cible inharmonique (trait plein) ou harmonique (pointillé) et pour les deux conditions de ΔF0.

4. DISCUSSION

Dans la condition HH (Fig. 2), ΔF0 a un effet semblable à celui décrit dans les travaux cités en Introduction. En revanche dès que la cible ou le fond sont inharmoniques, ΔF0 n'a plus d'effet.

L'hypothèse d'annulation harmonique prédit une identification meilleure lorsque le fond est harmonique, sauf si la cible est elle aussi harmonique et la ΔF0 est nulle (puisque l'annulation du fond entraînerait alors celle de la cible). C'est bien ce que l'on constate (Fig. 3), à ceci près que, quand la ΔF0 est nulle et la cible harmonique, l'identification est carrément *moins bonne* lorsque le fond est harmonique. Une explication possible est que les voyelles de fond inharmoniques se comportent comme des voyelles quasi-harmoniques dont la F0 diffère légèrement de leur F0 nominale (elles ont en général une hauteur différente de celle d'une voyelle harmonique de même F0). Dans la condition HI0 une certaine annulation serait alors possible, alors qu'elle reste inopérante dans la condition HH0. Quoi qu'il en soit, nos résultats sont compatibles avec l'hypothèse d'annulation harmonique.

L'hypothèse de renforcement harmonique prédit que l'identification sera meilleure lorsque la cible est harmonique, sauf si le fond est lui aussi harmonique et la ΔF0 nulle. Si l'hypothèse est fausse, on s'attend à un effet nul. Or, nous constatons un effet *inverse*.: lorsque la cible est harmonique, elle est moins bien identifiée (Fig. 2 et 3). Ce résultat est inattendu. Une explication possible est que, dans notre tâche, le système auditif fait appel systématiquement au mécanisme d'annulation harmonique. Celui-ci se calerait sur la cible lorsqu'elle est harmonique, et compromettrait ainsi son identification. Quoi qu'il en soit, le manque de résultat net en faveur du renforcement harmonique est étonnant. Il va à l'encontre de l'idée classique selon laquelle le système auditif utilise l'harmonicité des composantes d'une cible pour les grouper [6], ou que le voisement facilite l'extraction de la parole d'un fond bruité. Il est en revanche à rapprocher d'un résultat présenté à ce congrès [19]: l'annulation est plus efficace que le renforcement pour éliminer une voix parasite dans une tâche de reconnaissance de la parole.

Les taux d'identification des constituants des paires IH et HI sont différents, et ils varient avec les conditions de façon différente. La ségrégation n'a donc pas le caractère symétrique qui, selon Bregman [5, p 669], est le propre des mécanismes "primitifs", par opposition à ceux qui emploient des "schémas". Elle ne semble pas non plus due au simple décalage de fréquence entre partiels des deux voyelles qu'introduit l'inharmonicité. De tels décalages on attendrait des effets égaux pour les deux voyelles.

En revanche, on ne peut exclure une explication en termes d'effets de phase. Les partiels des voyelles harmoniques sont en phase sinus, et ceux des voyelles inharmoniques peuvent s'interpréter comme s'écartant progressivement de cette phase. Si le pattern de phase sinus allait de pair avec un pouvoir masquant et une résistance au masquage plus faibles, les effets sur les taux d'identification seraient semblables à ceux que nous avons constatés. De tels effets de phase ont été constatés [20, 21, 22]. Cette question reste à explorer.

REMERCIEMENTS

Ce travail a reçu le soutien d'une action incitative du programme "Sciences de la Cognition" du Ministére de la Recherche et de l'Espace. Nous remercions Nina Fales pour son assistance lors des expériences.

BIBLIOGRAPHIE

[1] Cherry, E. C. (1953), "Some experiments on the recognition of speech with one, and with two ears", JASA 25, 975-979.
[2] Brokx, J. P. L. and S. G. Nooteboom (1982), "Intonation and the perceptual separation of simultaneous voices", Journal of Phonetics. 10, 23-36.
[3] Darwin, C. J. and J. F. Culling (1990), "Speech perception seen through the ear", Speech Comm. 9, 469-475.
[4] McAdams, S. (1989), "Segregation of concurrent sounds. I: Effects of frequency modulation coherence", JASA 86, 2148-2159.
[5] Bregman, A. S. (1990), Auditory scene analysis, MIT Press, Cambridge, Mass.
[6] Darwin, C. J. (1981), "Perceptual grouping of speech components differing in fundamental frequency and onset-time", QJEP 33A, 185-207.
[7] Assmann, P. F. and Q. Summerfield (1990), "Modeling the perception of concurrent vowels: vowels with different fundamental frequencies", JASA 88, 680-697.
[8] Culling, J. F. and C. J. Darwin (1993), "Perceptual separation of simultaneous vowels: within and across-formant grouping by F0.", JASA 93, 3454-3467.
[9] Scheffers, M. T. M. (1983), "Sifting vowels", Thesis, University of Gröningen.
[10] Summerfield, Q. and P. F. Assmann (1991), "Perception of concurrent vowels: effects of harmonic misalignment and pitch-period asynchrony", JASA 89, 1364-1377.
[11] Summerfield, Q. and J. F. Culling (1992), "Auditory segregation of competing voices: absence of effects of FM or AM coherence", Phil. Trans. R. Soc. Lond. B 336, 357-366.
[12] Summerfield, Q. (1992), "Roles of harmonicity and coherent frequency modulation in auditory grouping" in "The auditory processing of speech", edited by B. Schouten, Mouton de Gruyter, 157-165.
[13] Zwicker, U. T. (1984), "Auditory recognition of diotic and dichotic vowel pairs", Speech Comm. 3, 256-277.
[14] de Cheveigné, A. (1993), "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing.", JASA 93, 3271-3290.
[15] de Cheveigné, A. (1993), "Time-domain comb filtering for speech separation", ATR technical report TR-H-016.
[16] Lea, A. (1992), "Auditory models of vowel perception", Thesis, University of Nottingham.
[17] Lea, A. P. and Q. Summerfield (1992), "Monaural segregation of competing voices", Proc. ASJ committee on Hearing. H-92-31, 1-7.
[18] Lea, A. and M. Tsuzaki (1993), "Segregation of competing voices: perceptual experiments", Proc. ASJ spring meeting, 361-362.
[19] de Cheveigné, A., H. Kawahara, K. Aikawa and A. Lea (1994), "Speech separation for speech recognition", *this congress*.
[20] Darwin, C. J. and R. B. Gardner (1986), "Mistuning of a harmonic of a vowel: grouping and phase effects on vowel quality", JASA. 79, 838-845.
[21] Culling, J. F. and C. J. Darwin (1993), "Perceptual and computational separation of simultaneous vowels: cues arising from low frequency beating", submitted for publication.
[22] Assmann, P. F. and Q. Summerfield (1993), "Some effects of duration on the perception of concurrent vowels", submitted for publication.