Longueur de confusion sur la plage vocalique

Ben Kaehler, John Smith and Joe Wolfe

School of Physics
University of New South Wales, Sydney, Australie
Tél: (61) (2) 93854954 - Fax (61) (2) 93856060
J.Wolfe@unsw.edu.au - http://www.phys.unsw.edu.au/~jw/index.html

ABSTRACT

Subjects were asked to identify monosyllabic synthesized words. The formants of the synthesized vowels were known, so the choices of a subject produces a perceptual map of his/her language. We compare such maps with similar maps for the vowels produced during speech. The chance of identifying a sound as a particular vowel decreases with its normalized displacement from the mean position of that vowel in the vocal plane. A plot of probability of identification as a function of separation in this plane defines a characteristic resolution/confusion 'distance' for a particular language. We report results of an experiment to determine such characteristic distances using synthesized speech and an automated testing routine.

1. INTRODUCTION

La plage vocalique est continue mais sa division en phonèmes est quantique. La reconnaissance des voyelles est un des exemples les plus longtemps étudiés du processus de catégorisation perceptive sur un axe continu.

Le nombre de voyelles varie considérablement selon les langues. Un nombre plus grand donne, en principe, la capacité de transmettre l'information à une vitesse élevée, mais à condition que la taille de la discrétisation dans l'espace continu dépasse la sensibilité ou la résolution du récepteur. Nous avons proposé une façon de définir une résolution moyenne dans un cas simple [Dow97], et calculé des valeurs à partir d'un ensemble de phonèmes français. Les mesures de la résolution sont statistiques et donc fonction de la façon d'échantillonner l'espace vocalique. La synthèse des sons permet un échantillonnage de tous les paramètres, sous contrôle explicite de l'expérimentateur. Nous présentons ici des mesures de la catégorisation perceptive de voyelles synthétiques.

Nous ne considérons ici que des voyelles à l'intérieur de mots monosyllabiques. Ceux-ci sont des cas simples, parce que, dans la parole normale, la reconnaissance des voyelles dépend fortement du contexte lexical et grammatical. Maints exemples existent, néanmoins, où la résolution d'une seule voyelle dans un mot isolé est nécessaire pour la compréhension, en particulier quand il s'agit de noms propres.

1.1 Mots prononcés ou synthétiques ?

Dans une étude précédente [Dow97; Dow00], un jury de douze francophones ont écouté trois répétitions de 250 voyelles prononcées par les sujets dont les fréquences de résonance du conduit vocal étaient mesurées. Les membres du jury indiquèrent à quel mot français le son qu'ils venaient d'entendre ressemblait le plus.

Cette expérience avait les avantages suivants : (i) le son était naturel et (ii) les résonances était déterminées par un algorithme qui n'avait pas besoin du jugement humain (on n'avait pas besoin de repérage formantique). Son désavantage était le manque de contrôle sur l'échantillonnage de la plage vocalique : le groupement des sons autour des voyelles naturelles du français limitait l'expérience et influençait le choix vers ces régions. De plus, l'existence de régions vides sur la plage vocalique empêche l'investigation de la perception des sons dans ces régions par une expérience dont les stimulis sont des voyelles naturelles.

Pour cette raison nous faisons une deuxième expérience, où les membres de plusieurs jurys écoutent et classent une série de sons synthétisés avec une gamme de formants connus qui échantillonnent la plage en densité uniforme. Pour cette expérience, nous sommes en train d'étudier trois langues différentes (l'anglais, l'espagnol et le français) mais jusqu'à présent nous n'avons des résultats statistiquement significatifs que pour l'anglais.

Pour chaque langue, nous analysons le classement des voyelles afin de déterminer un déplacement caractéristique ou longueur de confusion sur la plage vocalique de perception dans les conditions de l'expérience.

1.2 Formants et résonances

Dans une première approximation, les voyelles des langues occidentales sont classées et caractérisées par les deux ou trois premiers formants (F1, F2, F3), même si la durée et le ton sont parfois importants. Nous réservons le mot 'formant' pour décrire un pic dans l'enveloppe spectrale du son d'une voyelle. Dans l'étude précédente, nous avons mesuré les résonances acoustiques (R1, R2) du conduit vocal qui produisait la voyelle. On s'attend à ce que les fréquences des formants soient approximativement égales à celles des résonances, mais des différences non négligeables peuvent être introduites par l'interaction source-conduit de la source glottale et par l'impédance de radiation [Fan73].

1.3 Plans de production et de perception

Dans les espaces (F1, F2, F3) ou (R1, R2, R3), on peut classer les points en voyelles selon l'intention du locuteur ou selon la perception de l'auditeur [Fan73, Lan77, Dow97]. Quand un sujet prononce une voyelle, on peut mesurer ses formants ou les résonances du conduit vocal pendant la production. Un ensemble de mesures constitue ce que nous appelons le plan de production des voyelles de cette langue (et de ce sujet). Pour créer ce que nous appelons le plan de perception ou plan perceptif, un sujet identifie, comme voyelle de sa langue, un son dont on connaît les formants ou un son produit par un conduit vocal dont on connaît les résonances.

D'habitude, la compréhension d'une langue parlée, et surtout des voyelles isolées, est le résultat d'une forte ressemblance entre le plan de production et celui de perception, mais la compréhension des accents divers et des cas spéciaux tels que la parole sous hélium montre que la ressemblance peut être relative plutôt qu'absolue. Nous rapportons ici quelques différences quantitatives et qualitatives entre ces deux plans.

1.4 Distances sur la plage

Que veut dire déplacement sur la plage vocalique ? La plage de variation de R2 ou F2 est plus grande que celle de R1 ou F1. Pour cette raison, nous avons défini [Dow97] un déplacement sans dimension entre les points *a* et *b* sur la plage en deux dimensions comme suit:

$$\mathrm{d} \equiv \sqrt{\left(\frac{R1_a - R1_b}{\sigma_1}\right)^2 + \left(\frac{R2_a - R2_b}{\sigma_2}\right)^2} \qquad (1),$$

où σ_1 est l'écart type mesuré pour tous les résonances R1 de la langue, σ_2 est celui des R2. Dans l'expérience rapportée ici, les formants remplacent les résonances dans l'équation (1).

1.5 Résolution et longueur de confusion

Si on déplace une voyelle de sa place moyenne sur la plage vocalique, *quel sera le déplacement caractéristique moyen à partir duquel les auditeurs commencent à la confondre avec une autre ?* Pour chaque voyelle, on peut trouver ($\overline{F1}, \overline{F2}$) ou ($\overline{R1}, \overline{R2}$), le point du plan (F1,F2) ou (R1,R2) qui a les valeurs moyennes de tous les sons reconnus comme cette voyelle. Ce point est appelé la "place moyenne". Plus on est loin de ce point, plus grande est la chance de reconnaître le son comme une autre voyelle, ou comme un son qui n'est pas une voyelle. Nous avons trouvé [Dow97,Dow00] que la probabilité de reconnaissance d'un son (R1,R2) comme une voyelle ($\overline{R1}, \overline{R2}$) diminue de façon exponentielle avec la distance entre ces points. Le calcul de la longueur de confusion est compliqué parce que les mesures statistiques d'identification sont fonction de l'échantillonnage de la plage vocalique.

2. METHODOLOGIE

2.5 Mots synthétiques

Les voyelles sont placées dans des mots monosyllabiques, choisis pour chaque langue afin de minimiser le nombre de mots sans sens. En anglais, la forme 'h<V>d' ne donne qu'un seul mot sans sens en anglais[1]. Pour les voyelles <V>, F1 et F2 furent altérés par pas de 5% sur un quadrilatère ((300,700) (700,1000) (1450,700) (1950,300) (valeurs en Hz)) du plan (F2,F1) et F3 a pris quatre valeurs entre 2.1 et 3.1 kHz. L'enveloppe spectrale a été calculée explicitement et les voyelles ont été synthétisées par somme de sinus. Les fréquences fondamentales moyennes sont respectivement 132 et 250 Hz pour les voix "masculines" et "féminines", avec une réduction légère et cubique pendant le mot. Les valeurs (en %) du *flutter, creak, jitter* et *shimmer* sont (0.5, 0.5, 1.5, 1.5) et (0, 0.2, 0, 0) pour les deux. La largeur de bande des formants a été déterminée par la régression polynomiale décrite par [Haw95], et la taille par régression linéaire sur les fréquences centrales des formants. Le /h/ est du bruit blanc, caractérisé dans le domaine spectral pour des échantillons tous les 10 Hz et passé par les formants de la voyelle suivante. Les locus pour le /d/ sont 1749 Hz et 2000 Hz et un délai de 50 ms précède une plosive de 30 ms.

2.6 Classement des sons synthétiques

Le système de classement est automatisé. Les sujets répondent aux questions posées par un ordinateur qui synthétise les 'mots' en ordre aléatoire pour chaque auditeur. Chaque mot est produit trois fois à la suite, et un sous-ensemble d'approximativement 100-130 mots décrivant l'espace paramétrique (F0,F1,F2,F3) est présenté dans un ordre aléatoire pour chaque sujet. Pour l'expérience sur l'anglais, 113 volontaires, étudiants de l'Université de Nouvelle Galles du Sud à Sydney, ont participé. Ils cliquent sur des 'boutons' nommés *had, hard, head, herd, heed, hid, hoard, hod, hood, hud, who'd* et *none of these*. Une autre série de boutons leur permet de classer le son sur cinq niveaux entre « *quite unnatural* » et « *quite natural* ». Pour les autres langues, nous sommes toujours en train de faire des mesures.

3. RESULTATS ET DISCUSSION

Les auditeurs anglophones ont donné aux mots synthétisés avec une voix masculine (c'est à dire de faible F0) une valeur moyenne de 4.8 sur la gamme *quite unnatural* (1) à *quite natural* (5). Les commentaires de certains membres du jury après les séances de mesure indiquent qu'ils ne

[1] 'heed' [i], 'hid' [ɪ], 'head' [ɛ], 'had' [æ], 'hard' [ɑ], 'hod' [ɒ], 'hoard' [ɔ], 'hood' [ʊ], 'who'd' [u], 'hud' [ʌ] et 'herd' [ɜ]. Malgré l'acronyme HUD (Head-Up Display), 'hud' n'est pas encore un mot anglais.

s'étaient pas rendus compte que les mots avaient été synthétisés.

3.1 Plan de perception

On a calculé pour chaque voyelle ($\overline{F2}, \overline{F1}$), les moyennes des valeurs de chaque son identifié comme cette voyelle. Les positions sur le plan perceptif des voyelles anglaises (Fig. 1) ressemblent globalement aux positions sur les plans de production en (F1,F2) [Ber67] et en (R1,R2) [Epp97]. La similarité entre /i/ et /ɪ/ n'est pas surprenante: en production, ces voyelles sont largement distinguées par longueur (/ɪ/ est courte). Le plan de perception des voyelles "masculines" est déplacé vers l'origine par rapport au plan de voyelles "féminines" (c'est à dire de F0 élevée non présentées). Ce déplacement rappelle le déplacement bien connu pour le plan de production.

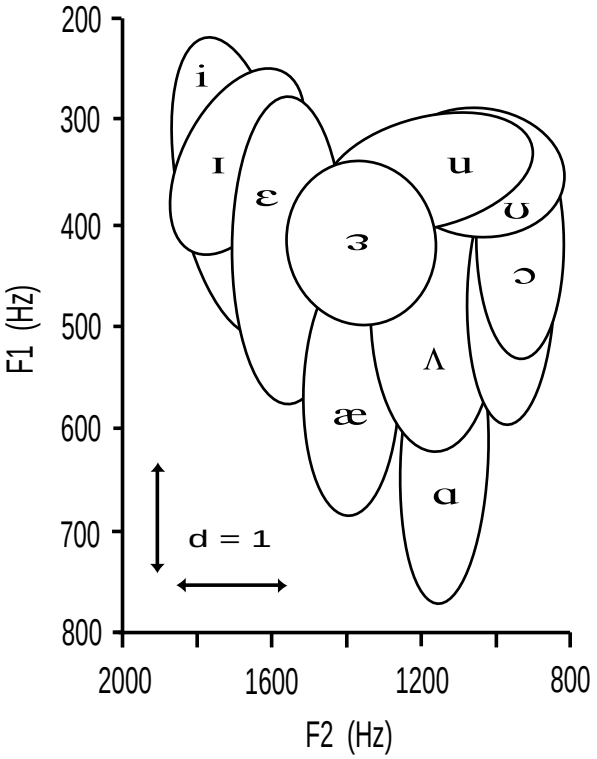

Figure 1 Plan perceptif des voyelles non nasalisées 'masculines', synthétisées pour échantillonner le quadrilatère vocalique en densité uniforme. Dans cette expérience, les auditeurs sont anglophones. Les axes sont proportionnés de telle sorte que les barres, qui indiquent les écarts types (σ_1 et σ_2) pour tous les R1 et R2, soient de longueurs égales. Ces barres sont donc de longueur d = 1, selon la définition (1). La pente du grand axe est le coefficient de régression linéaire et les axes des ellipses indiquent l'écart type pour chaque voyelle dans cette direction et la direction perpendiculaire.

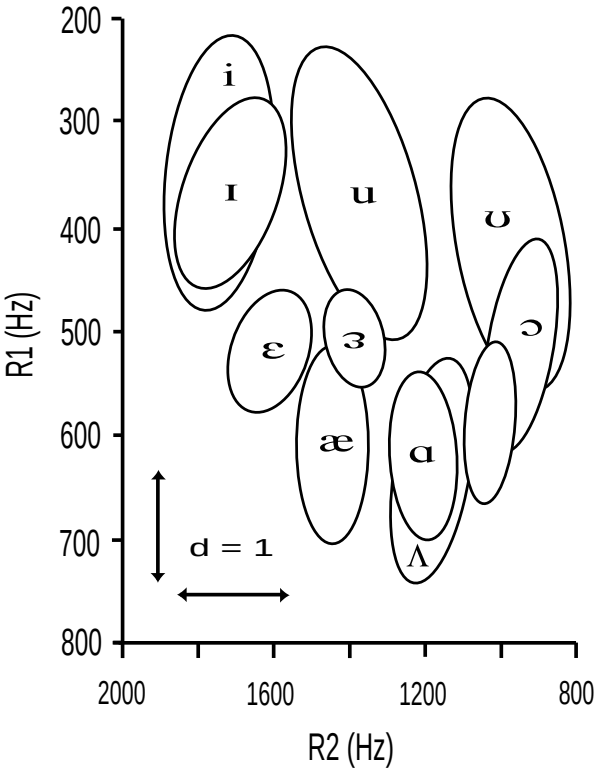

Figure 2 Le plan de production pour les voyelles de l'anglais australien, calculé à partir des données de [Epp97]. Mesures des résonances (R1,R2) des conduits vocaux de 33 jeunes australiens masculins lorsqu'ils prononcent les voyelles indiquées.

Dans l'expérience que nous rapportons ici, le quadrilatère vocalique est échantillonné partout, et le recouvrement de la plage de perception montrent que, même si l'anglais ne se sert pas de certaines régions de la plage, les anglophones reconnaissent des sons dans ces régions comme des voyelles. 12% des choix était '*none of these*', mais la distribution de tels choix couvre le plan entier.

3.2 Comparaison des plans de production et de perception

Pour ces sons synthétiques, la distribution de points pour une voyelle perçue est typiquement plus grande que sa distribution pour la même voyelle produite (voir Figures 1 et 2.). (Cette observation est aussi vraie pour le français, et nous avons mis ces résultats sur le site internet [Dow00].) La forme des distributions est différente aussi: dans le plan de perception les distributions sont plus larges (c'est à dire que le rapport $\sigma(F2)/\sigma(F1)$ pour une voyelle est typiquement plus grand que le même rapport dans le plan de production). Les positions relatives sur les deux plans sont approximativement semblables mais quelques différences sont évidentes. Sur le plan de production, [ʌ] (*hud*) et [ɑ] (*hard*) ont presque la même moyenne: ces deux voyelles sont distinguées en production largement par la durée ([ʌ] est courte). Pour

[ʌ] perçue, F1 est inférieure à sa valeur dans le plan de production, pour [ɑ] perçue, F1 est supérieure à sa valeur dans le plan de production.. Sur le plan de production, il y a une région (entre [u] et [ʊ]) dont les locuteurs anglophones ne se servent que rarement (Figure 1). Si F1 est faible, les sons dans cette région vide sont souvent perçus comme [u] ou parfois [ʊ]. Pour F1 plus important, ils sont perçus comme [ʌ].

Cette région du plan de production (F2 ~ 1200 Hz, F1 < 700 Hz) est vide en français aussi où elle fait partie du triangle nasal [Lan77;Dow00]. Dans notre étude précédente, pourtant, l'emploi de voyelles prononcées comme stimulus nous a empêché de mesurer cette région du plan de perception [Dow97].

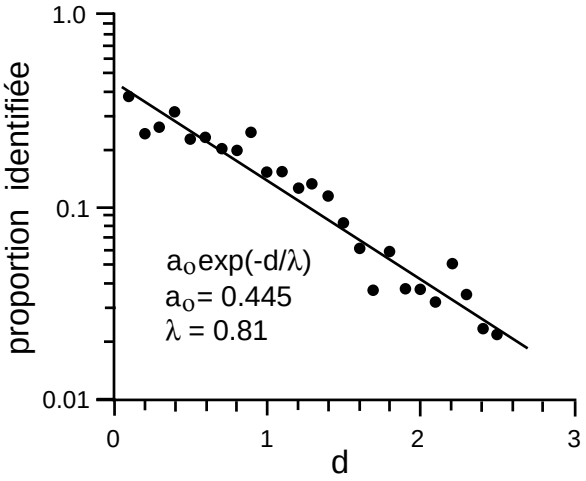

Figure 3 La proportion d'identifications en fonction de *d* sur le quadrilatère vocalique échantillonné en densité uniforme par des sons synthétiques.

3.3 Longueur de confusion

Pour chaque voyelle de place moyenne $(\overline{F2}, \overline{F1})$ le déplacement de chaque son de formants (F2,F1) a été calculé, et le nombre d'identifications a été compté pour chaque anneau $j.\delta \le d \le (j+1).\sigma$, $\delta = 0.05$. La Figure 3 montre la proportion identifiée dans chaque anneau en fonction de son rayon moyen. La figure montre aussi une fonction déduite des données: une exponentielle simple de longueur λ.

Pour l'anglais sous ces conditions $\lambda = 0.81$, c'est à dire que sa longueur de confusion est du même ordre mais est légèrement plus grande que l'écart type de tous les formants (en unités sans dimension, $\sigma(R1) \equiv 1 \equiv \sigma(R2)$). La Figure 1 et cette observation montrent que le plan de perception n'est pas divisé avec une grande précision. Le degré de recouvrement entre régions de voyelles contiguës est important. Hors de tout contexte, donc, on attend une confusion des voyelles voisines. La longueur de confusion est aussi du même ordre que FO. (λ est sans dimension. Sa taille en Hz est fonction de l'orientation. Dans la direction R1, $\lambda = 0.6$ F0. Dans la direction R2, $\lambda = 1.7$ F0.) Ce résultat n'est pas surprenant non plus : dans le domaine fréquentiel, l'enveloppe spectrale est échantillonnée par pas de F0, et on s'attend à ce que le limite de la résolution perceptive des formants soit de cet ordre, en l'absence d'autres informations.

3.4 Étude en cours

Il serait intéressant de pouvoir comparer des mesures des longueurs de confusion de langues ayant des nombres différents de voyelles. Pour cette raison, nous faisons actuellement des mesures automatisées en (F0,F1,F2,F3) pour le français (16 voyelles) et l'espagnol (5 voyelles), sous un protocole standardisé sur la forme décrite ici pour l'anglais. Cette étude nous donnera aussi (pour le français) la comparaison des méthodes des stimulus synthétiques et naturels.

Remerciements.

Nous remercions l'Australian Research Council qui a subventionné, en partie, cette recherche. Nous remercions aussi tous nos volontaires.

BIBLIOGRAPHIE

[Ber67] Bernard, J.R.L. (1967), " Length and identification of Australian vowels ", Australasian Universities Modern Language Assoc. Vol 27, pp. 100-120.

[Dow97] Dowd, A., Smith, J.R. and Wolfe, J. (1997). "Learning to pronounce vowel sounds in a foreign language using acoustic measurements of the vocal tract as feedback in real time." Language and Speech, Vol 41, pp. 1-20.

[Dow00] Dowd, A., Smith, J.R. and Wolfe, J. (1900). "French vowels" http://www.phys.unsw.edu.au/~jw/french.html

[Epp97] Epps, J., Dowd, A., Smith, J.R. and Wolfe, J. (1997) "Real time measurements of the vocal tract resonances during speech", Eurospeech'97, pp. 721-724. www.phys.unsw.edu.au/~jw/Eurospeech.html

[Epp97] Epps, J., Smith, J., et Wolfe, J. (1997) "A novel instrument to measure acoustic resonances of the vocal tract during speech", Measurement Science and Technology, Vol. 8, pp. 1112-1121.

[Fan73] Fant, G. (1973), " Speech Sounds and Features ", MIT, Cambridge, Mass.

[Haw95] Hawks, J.W. et Miller, J.D. (1995). "A formant bandwidth estimation procedure for vowel synthesis." J. Acoust. Soc. Am., 97, pp. 1343-1344.

[Lan77] Landercy, A. et Renard, R (1977). "Éléments de Phonétique." Didier, Bruxelles.