La reconnaissance des sons consonantiques en cas de désynchronisation spectrale : avec et sans information spectrale fine

Marjolaine Ray, Olivier Crouzet
LLING – Laboratoire de Linguistique de Nantes – EA3827
Université de Nantes, Chemin de la Censive du Tertre, 44312 Nantes
www.lling.fr

RÉSUMÉ

Deux expériences ont été menées pour étudier l'identification de séquences VCV temporellement désynchronisées lorsqu'une partie de leurs informations spectrales était préalablement détruite : dans l'expérience 1, le spectre était intact (et contenait informations spectrales fines (TFS) et enveloppe d'amplitude), et dans l'expérience 2, seule l'enveloppe d'amplitude subsistait. Le but de cette étude était d'observer la baisse des scores d'intelligibilité en fonction du degré d'asynchronie et des indices acoustiques supprimés (i.e. la TFS). En accord avec les précédentes recherches concernant l'identification de phrases, on observe que l'identification des consonnes lors de tâches forcées de catégorisation engendre des scores de performance relativement hauts malgré des degrés de désynchronisation très élevés, bien que les scores de reconnaissance par consonnes soient très hétérogènes. Les causes possibles des divergences entre nos données et les résultats antérieurs sont envisagées.

ABSTRACT

Recognition of desynchronized consonantics sounds with and without fine spectral structure

This paper reports two experiments in which the identification of desynchronized VCV sequences was investigated with either both fine spectral structure (TFS) and envelope information or envelope information alone. These experiments compare the decrease in intelligibility scores with respect to the degree of desynchronization applied between spectral channels. The data are first analysed in terms of global performance intelligibility, then intelligibility of individual consonant sounds are investigated. Confirming previous data obtained in sentence identification tasks, it is shown that consonant identification in a forced choice categorisation task occurs with relatively high levels of performance even for strong levels of desynchronization, but that performance is highly variable depending on individual consonants. Various explanations for differences between our results and preceding work are discussed.

MOTS-CLÉS : Enveloppe d'amplitude, désynchronisation temporelle, structure spectrale fine, parole *vocodée*.

KEYWORDS: Temporal envelope, temporal desynchronization, fine spectral structure, vocoded speech.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 33–40,
Grenoble, 4 au 8 juin 2012.

1 Introduction

La perception des sons de parole se fait par le biais d'une analyse spectrale. Les filtres du système auditif analysent le signal sonore en fines bandes de fréquences, aussi a-t-on longtemps considéré que la perception et l'identification de la parole dépendaient largement des variations rapides d'énergie spectrale, c'est à dire de la structure spectrale fine. Néanmoins, certaines études ont démontré par la suite [8, 10] qu'une autre forme d'information tenait un rôle prépondérant dans la reconnaissance de la parole : l'information temporelle, une information basée sur les modulations d'amplitude relativement lentes, en fonction de l'axe temporel. Ces modulations d'amplitude se sont avérées être une source d'information particulièrement résistante en cas de dégradations des sons de parole. D'autre part, de nombreuses recherches [8, 3, 2, 6] ont montré que les indices spectro-temporels contenus dans l'enveloppe d'amplitude uniquement constituaient une source d'information suffisante pour obtenir une intelligibilité élevée.

Les travaux de Van Tasell, Soli, Kirby & Widin [10] ont montré que l'enveloppe d'amplitude globale d'un signal de parole contenait des informations non-négligeables. Après avoir extrait l'enveloppe d'un signal de parole grâce à un filtre passe-bas, selon trois fréquences de coupure différentes (20, 200 et 2000 Hz), Van Tasell et al. [10] ont appliqué ces enveloppes sur du bruit rose. Ces enveloppes provenaient donc du signal global et non de bandes de fréquence individuelles. Les trois types de stimuli ainsi générés donnaient lieu à des scores de reconnaissance assez faibles, mais significativement plus élevés que si les réponses avaient été aléatoires. Van Tasell et al. [10] en déduisent que l'enveloppe d'amplitude à elle seule transmet de l'information, et calculent les pourcentages moyens d'information transmise par les signaux (22% d'information transmise à 20 Hz de fréquence de coupure, 29% d'information à 200 Hz, 35% d'information à 2000 Hz). L'enveloppe temporelle véhicule donc de nombreuses informations même lorsque toutes les informations d'ordre fréquentiel ont disparu.

Shannon, Zeng, Kamath, Wygonski & Ekelid [8] divisent des signaux de parole – préalablement traités par un *vocoder* [1]– en 1, 2, 3 ou 4 bandes de fréquences. Ils observent qu'il suffit de trois bandes spectrales pour obtenir un score de reconnaissance proche de 90 %, et soulignent qu'aucune structure formantique n'est présente et que les transitions formantiques sont tout à fait brouillées. Le contenu spectral est donc très réduit, mais les indices temporels de l'enveloppe d'amplitude sont suffisants pour obtenir une reconnaissance de 90 %.

La dégradation des qualités acoustiques de l'envelopppe d'amplitude a permis d'étudier son rôle dans la perception des sons de parole. Arai & Greenberg [3] ont utilisé pour ceci la désynchronisation temporelle par bandes spectrales. Leur procédure consistait, tout d'abord, à diviser le spectre du signal de parole en 19 bandes de fréquence, puis ensuite, à désynchroniser temporellement ces bandes en les replaçant aléatoirement sur l'axe temporel.

La désynchronisation temporelle a un effet spécifique sur l'enveloppe d'amplitude globale du signal : en effet, tandis que toutes les informations spectro-temporelles sont conservées au sein de chaque bande de fréquence, l'enveloppe d'amplitude *globale* du signal, elle, subit une dispersion du fait de la désynchronisation temporelle de chacune de ces bandes. La dispersion et l'étalement réduisent alors la profondeur globale des modulations d'amplitude du signal. Arai & Greenberg [3] montrent que les auditeurs résistent bien à cette désynchronisation, même

1. Le *vocoder* utilisé par Shannon et al. [8] extrayait les modulations d'amplitude grâce à une Transformée de Hilbert, sélectionne la courbe positive de la modulation (*h*alf-wave rectification), puis la coupe grâce à un filtre passe-bas (à la fréquence de coupure désirée), et enfin utilise ces données pour moduler un échantillon de bruit.

lorsqu'elle atteint des seuils élevés : lorsque l'asynchronie maximale atteint 200 ms, le score de reconnaissance est encore de 50 %. En mesurant la baisse de profondeur d'amplitude des modulations, ils soulignent qu'elle semble être corrélée avec la baisse de l'intelligibilité.

Plus récemment, Fu & Galvin [2], ont poursuivi le travail de Arai & Greenberg [3] en étudiant la perception de signaux temporellement désynchronisés pour deux types de sons : des sons *vocodés* et des sons dont le spectre était intact. La désynchronisation était appliquée (1) aux signaux intacts et (2) aux mêmes signaux passés à travers un *vocoder* (de 4 ou 16 canaux). Fu & Galvin [2] observent que la perte de détails spectro-temporels ne détériore pas gravement l'intelligibilité tant que la désynchronisation est minime, mais que lorsque le niveau d'asynchronie augmente, la structure spectrale fine pourrait être source d'informations vitales pour la résistance aux dégradations. Par ailleurs, Fu & Galvin [2] relèvent aussi l'importance de la résolution spectrale pour l'intelligibilité lorsque les informations spectrales fines sont détruites, ce qui se traduit par une résistance plus élevée avec 16 bandes qu'avec 4 bandes.

Ces deux études [3, 2] ont eu recours à des tâches d'identification de mots dans des phrases. Dans les deux expériences présentées ici, nous cherchons à affiner notre compréhension de ces mécanismes en étudiant spécifiquement l'identification des segments consonantiques dans ces conditions dégradées, pour tenter d'analyser le rôle des différents types d'indices acoustiques en cas de désynchronisation temporelle pour les sons consonantiques spécifiquement. Cette approche nous permet non seulement d'étudier les performances globales d'identification et de les comparer aux résultats obtenus avec des phrases, mais aussi d'analyser plus spécifiquement le comportement de chaque consonne en fonction des dégradations appliquées.

2 Procédure

2.1 Protocole

Quinze sujets normo-entendants agés de 18 à 24 ans ont participé à l'expérience. Ils écoutaient des séquences VCV[2] au casque tandis que l'ensemble des séquences, transcrites orthographiquement, étaient présentées à l'écran. Les participants devaient cliquer sur l'item qu'ils pensaient avoir reconnu. La sélection engendrait alors la lecture aléatoire d'une autre item. Une séquence d'entraînement avait lieu avant les expériences, où les auditeurs entendaient chaque item (non-modifié) une fois.

2.2 Materiel

Les séquences VCV contenaient deux fois la même voyelle : [a], tandis que les consonnes pouvaient appartenir à l'ensemble du système consonantique français {p, t, k, b, d, g, f, s, ʃ, v, z, ʒ, m, n, ɲ, ʁ, l, w, j}. Ces séquences ont été préalablement enregistrées par une locutrice féminine. Les caractéristiques acoustiques des signaux ont ensuite été manipulées au sein de l'environnement de traitement du signal Octave [1]. Chaque séquence passait à travers un banc de filtres passe-bandes composé de 19 bandes (une condition identique à l'expérience de Arai & Greenberg [3], et proche des signaux de 16 bandes de Fu & Galvin [2]) . La largeur des

2. Voyelle - Consonne - Voyelle

bandes de fréquence et le calcul de leurs fréquences centrales était généré sur une échelle ERB en fonction du nombre de canaux en utilisant une implémentation de Slaney [9]. La réponse des filtres était basée sur un ordre 100 pour limiter les phénomènes de redondance à l'intérieur de chaque bande.

Dans l'expérience 1, tous les indices acoustiques du spectre étaient conservés au sein de chaque bande de fréquence. Les canaux générés par le filtre décrit précédemment étaient temporellement désynchronisés suivant la procédure décrite par Arai & Greenberg (1998) : 19 valeurs de décalage étaient générées linéairement de 0 à Dmax (le décalage maximal). Chacune de ces 19 valeurs de décalage était attribuée aléatoirement à l'une des 19 bandes de fréquence. Pour éviter la formation de poches à haute corrélation temporelle, le délai entre deux canaux adjacents était contrôlé pour être toujours supérieur à 1/4 de Dmax. Le degré global de désynchronisation dépendait de la valeur de Dmax : plus Dmax augmentait, plus les écarts entre les 19 valeurs générées linénairement augmentaient. Dans ces expériences, Dmax variait entre 0 et 240 ms : la première valeur non-nulle de Dmax était 60 ms, puis elle augmentait par pas de 20 ms. Au final, 11 degrés de désynchronisation ont été utilisés (0, 60, 80, 100, 120, 140, 160, 180, 200, 220, 240 ms). Une fois les bandes de fréquence désynchronisées entre elles ; elles étaient recombinées et passées à travers un filtre passe-bas éliminant toutes les fréquences supérieures à la fréquence de Nyquist.

Dans l'expérience 2, la structure spectrale fine (TFS) était supprimée et seules les variations lentes d'énergie étaient maintenues. Chacune des bandes de fréquence du banc de filtre passait à travers un *vocoder* qui multipliait aléatoirement chaque échantillon du signal par 1 ou -1 selon la procédure de Schroeder [7] [3]. Cette procédure était appliquée à chaque canal séparément, avant de refiltrer individuellement les bandes pour éviter tout artefact de fréquence. Les signaux résultants sont équivalents à la combinaison de 19 bandes de bruit dont les modulations d'amplitude sont similaires à celles des signaux naturels d'origine. Les séquences VCV ainsi produites étaient ensuite désynchronisées selon la même procédure que les signaux de l'expérience 1.

3 Résultats

Nous avons procédé à deux types d'analyses de données : le relevé des scores de reconnaissance globaux, et l'étude des scores de reconnaissance individuels par consonne.

3.1 Scores de reconnaissance : analyse globale

Les pourcentages d'identification correcte des deux expériences sont présentés dans la figure 1. Le pourcentage d'identification est représenté en fonction des niveaux moyens de désynchronisation (Dmax/2) entre les 19 bandes de fréquence. La ligne horizontale située à 5,26 % indique le pourcentage théorique qu'atteindraient les scores dans le cas de réponses aléatoires ($1/n \times 100$, n le nombre de réponses possibles, soit $n = 19$). L'intelligibilité chute à 55 % et 42 % de reconnaissance correcte pour les signaux intacts et les signaux vocodés respectivement lorsque le

3. La procédure de Schroeder [7] est différente de celle appliquée par Fu & Galvin [2] durant leur expérience. En effet, Fu & Galvin [2] utilisaient des modulations extraites grâce à une Transformée de Hilbert et qui permettent de contrôler quelle gamme des fréquences de modulation d'amplitude est conservée dans les signaux vocodés.

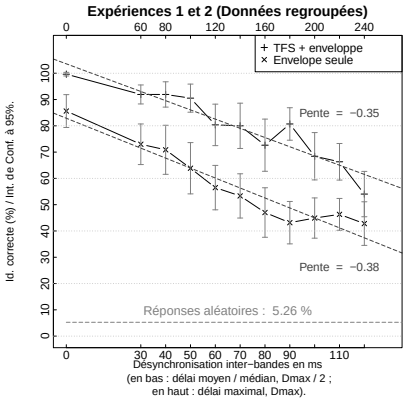
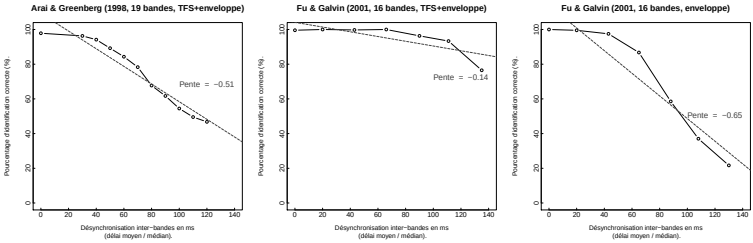


FIGURE 1: Scores de reconnaissance correcte pour les expériences 1 & 2 en fonction du degré de désynchronisation. Chaque pourcentage est issu de n observations, où $n = 19 \times 15$, 19 items et 15 sujets.



(a) Arai & Greenberg : signaux intacts (TFS + enveloppe)

(b) Fu & Galvin : signaux intacts (TFS + enveloppe)

(c) Fu & Galvin : signaux vocodés (enveloppe)

FIGURE 2: Scores de reconnaissance des études précédentes

degré d'asynchronie est le plus élevé. Dans tous les cas, les scores restent supérieurs au taux de réponses aléatoires, même aux degrés de désynchronisation les plus élevés.

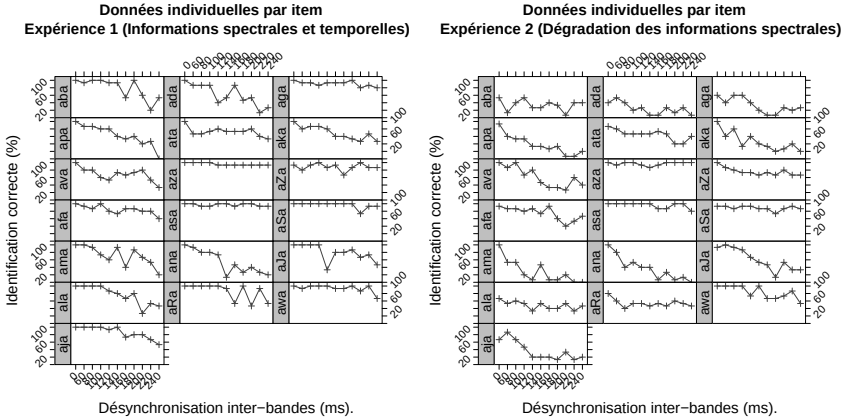
Nous pouvons étudier ces données selon une analyse linéaire, pour permettre une comparaison avec les résultats antérieurs. Les points de la régression linéaire pour nos résultats ont été calculés en fonction de la moyenne du pourcentage de réussite pour chaque individu (soit 15 points, chacun obtenus par un pourcentage calculé sur 19 résultats).

Les pentes des scores de performance de Arai & Greenberg [3] et de Fu & Galvin [2] sont assez

différentes de celles obtenues pour nos expériences (respectivement -0.51 pour [3] et, en ce qui concerne [2], -0,14 pour les signaux intacts et -0,65 pour les signaux vocodés, contre -0,35 et -0,38) [4]. Contrairement aux courbes des expériences 1 & 2 qui suivent des pentes semblables même si les scores de reconnaissance sont plus faibles pour les signaux *vocodés*, les performances de Fu & Galvin décrivent une pente beaucoup plus aïgue pour les signaux vocodés que pour les signaux intacts.

3.2 Scores de reconnaissance : analyse individuelle

En analysant l'intelligibilité de chaque son consonantique en fonction de l'aggravation de la désynchronisation (Figure 3), on observe qu'ils ne sont pas tous affectés de la même manière par les dégradations acoustiques. Globalement, la reconnaissance des fricatives est très peu



(a) Expérience 1

(b) Expérience 2

FIGURE 3: Expérience 1 et 2 : Reconnaissance des consonnes en fonction du degré de désynchronisation (description phonétique Sampa). (a) : enveloppe + TFS, (b) : enveloppe seule.
Les pourcentages de reconnaissance correcte représentés dans les Figures 3a et 3b sont calculés sur 15 observations pour chaque degré d'asynchronie.

affectée par la désynchronisation temporelle, surtout celle des fricatives non-voisées ; tandis que l'intelligibilité des occlusives baisse très vite, surtout celle des labiales. Ce comportement typique des occlusives et des fricatives se répète pour les deux expériences, avec et sans informations spectrales fines.

4. Les études de référence [2, 3] sont basées sur des tests d'intelligibilité de phrases dans lesquelles le contexte peut contribuer à la reconnaissance des mots, tandis que nous avons utilisé des logatomes. La comparaison reste néanmoins intéressante puisque l'identification phonétique intervient dans les deux cas

Leurs caractéristiques acoustiques et phonétiques [5] pourraient expliquer qu'ils résistent différemment aux phénomènes de désynchronisation et de dégradations spectrales, et donnent lieu à des scores de performances hétérogènes. Ces disparités entre les performances semblent pouvoir être associées à un certain nombre de traits distinctifs [4] des sons consonantiques. La meilleure résistance des fricatives en cas de désynchronisation temporelle pourrait s'expliquer par la longueur de leur bruit de friction, qui ne sera jamais totalement dispersé par l'asynchronie. Les fricatives [s, z, ʃ, ʒ] particulièrement génèrent une très forte intensité à des fréquences plus élevées [5] (le découpage logarithmique du filtre passe-bande ménage des bandes plus larges dans les hautes fréquences, dont le contenu spectral est moins touché par la désynchronisation). A l'inverse, la brièveté des occlusives pourrait les rendre très sensibles à la désynchronisation temporelle : les transitions formantiques, qui impliquent une variation spectro-temporelle très brève, chevauchent plusieurs bandes de fréquence et sont vite dispersées par l'asynchronie.

4 Discussion

En ce qui concerne les écarts entre nos résultats et ceux de Fu & Galvin (Figures 2b et 2c) : plusieurs sources peuvent être envisagées. La technique de désynchronisation utilisée lors de l'expérience de Fu & Galvin n'appariait pas les valeurs de décalage aléatoirement, mais selon une procédure séquentielle organisant les valeurs de décalage du canal le plus bas au plus élevé. Par ailleurs, la procédure que nous avons utilisé pour détruire la TFS dans les signaux est différente de celle utilisée par Fu & Galvin (cf. Materiel).

Nous n'envisageons pas que les scores de nos expériences, et ceux des études prédédentes [3, 2] chutent forcément de manière linéaire. De fait, les pentes des droites de régression ne sont peut-être pas le meilleur reflet des données obtenues. En effet, les tâches d'identification (oui/non) produisent des variables catégorielles, que les régressions linéaires ne peuvent pas représenter. Nous envisageons donc de calculer une régression logistique mixte, qui reflétera mieux nos variables catégorielles. Pour l'instant, la régression linéaire reste un outil de comparaison intéressant avec les précédentes études [3, 2], dont les données brutes ne sont pas disponibles.

Même si nous avons principalement utilisé des indices spectraux (brièveté des énergies, transitions formantiques) pour expliquer les résultats individuels par sons consonantiques, les écarts de résistance en fonction des traits phonétiques pourraient aussi s'expliquer par un affaiblissement variable des modulations d'amplitude selon les sons consonantiques. Shannon et al. [8] soulignent dans leurs travaux la faible importance des données spectrales dans l'identification des mots, puisque toutes les informations de type formantique ont disparu à l'intérieur de leurs items, et que le score de reconnaissance est malgré tout de 90 %. Ils font donc l'hypothèse que l'enveloppe temporelle est un vecteur d'information principal. La forme semblable de nos deux courbes de performances pourrait valider cette hypothèse, puisqu'en comparant la baisse de l'intelligibilité pour les deux expériences, avec et sans information spectrale fine, on observe que, bien que les scores soient plus bas pour l'expérience 2, la pente de la fonction n'est pas plus aïgue (- 0,35 pour l'expérience 1 vs. - 0,38 pour l'expérience 2). La disparition des informations spectrales fines provoque une baisse des performances, mais la chute de l'intelligibilité n'est pas plus rapide que lors de l'expérience 1.

Nous pouvons faire l'hypothèse, selon les observations d'Arai & Greenberg [3], que la forme des deux courbes corresponde à la baisse de la profondeur des modulations d'amplitude communes,

ce qui expliquerait qu'elles soient si semblables. Déterminer ceci nécessitera de calculer, lors de futures analyses, la profondeur des modulations d'amplitude pour savoir si sa diminution est corrélée avec la chute de l'intelligibilité.

Remerciements

Cette étude a été réalisée grâce au soutien du Conseil Régional des Pays de la Loire (convention n. 939 92 6513).

Références

[1] John W. Eaton. *GNU Octave Manual*. Network Theory Limited, 2002.

[2] Qian-Jie Fu and John J. Galvin. Recognition of spectrally asynchronous speech by normal-hearing listeners and Nucleus-22 cochlear implant users. *Journal of the Acoustical Society of America*, 109(3) :1166–1172, 2001.

[3] Steven Greenberg and Takayuki Arai. Speech intelligibility in the presence of cross-channel spectral asynchrony. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 933–936, 1998.

[4] Roman Jakobson, C. Gunnar M. Fant, and Morris Halle. Preliminaries to Speech Analysis. The distinctive features and their correlates. Technical Report 13, Acoustics Laboratory, Massachussetts Institute of Technology, 1952.

[5] Raymond D. Kent and Charles Read. *The Acoustic Analysis of Speech*. Singular Publishing, San Diego : CA, 1992.

[6] Mariken ter Keurs, Joost M. Festen, and Reinier Plomp. Effect of spectral envelope smearing on speech reception. I. *Journal of the Acoustical Society of America*, 91(5) :2872–2880, 1992.

[7] Schroeder, M. R. Vocoders : Analysis and Synthesis of speech. *Proc. IEEE*, 54 :720 – 734, 1966.

[8] Robert V. Shannon, Fan-Gang Zeng, Vivek Kamath, John Wygonski, and Michael Ekelid. Speech Recognition with Primarily Temporal Cues. *Science*, 270 :303–304, 1995.

[9] Malcolm Slaney. Auditory Toolbox. Apple Technical Report 45, Apple Computer Inc., Advanced Technology Group, 1993.

[10] Dianne J. Van Tasell, Sigfrid D. Soli, Virginia M. Kirby, and Gregory P. Widin. Speech waveform envelope cues for consonant recognition. *Journal of the Acoustical Society of America*, 82(4) :1152–1161, 1987.