Saillances perceptives du DEV et de la Transition Formantique dans l'identification de la parole accélérée

Caroline Jacquier et Fanny Meunier

Laboratoire Dynamique du Langage (UMR 5596 – Université Lumière Lyon 2)
14, avenue Berthelot – 69363 Lyon Cedex 07, France
Tél.: +33 (0)4 72 72 65 34 - Fax: +33 (0)4 72 72 65 90
Mél: *jacquier@isc.cnrs.fr, fanny.meunier@univ-lyon2.fr*

ABSTRACT

The cognitive use of the phonetic and acoustic features still needs to be specified for speech comprehension. Our study explored the temporal encoding of acoustic cues during natural speech perception by normal adults. We focused on two short attributes of speech: Voice Onset Time (VOT) and second formant transition. Normal hearing subjects had to identify disyllabic CVCV non-words that have been time-compressed on both acoustic cues simultaneously (Experiment 1) and separately on each cue (Experiments 2 and 3). First results showed a large inter-individual variance and demonstrated that the VOT is a more salient segmental cue than the formant transition. However, both acoustic cues are needed for a fine perception of speech.

1. INTRODUCTION

La perception de la parole est modulée par des facteurs environnementaux comme le bruit ou bien directement par des variations du locuteur lui-même, par exemple une voix cassée, qui vont altérer l'intégrité acoustique du signal. La plupart du temps, la parole et le message qu'elle transporte reste compréhensibles en dépit de ces dégradations. Certains mécanismes cognitifs semblent compenser et permettre la reconstruction du signal de parole altéré Cependant, il semblerait que ces capacités de perception et de compréhension de la parole dégradée soient propres à chacun [1]. Dans ce champ de recherche, des études ont souligné l'importance des indices acoustiques dans la perception de la parole [2-4]. En effet, le signal de parole est composé de nombreux indices acoustiques qui ont des degrés d'importance différents dans la perception et la bonne compréhension de la parole. Une étude précise de ces segments acoustiques est nécessaire afin d'identifier les indices acoustiques les plus pertinents dans l'intelligibilité de la parole. Il a été observé notamment que des enfants ayants des troubles du langage et de l'apprentissage montraient des difficultés à percevoir des segments brefs du signal de parole [5]. Cependant, la nature des indices acoustiques qui induisent ces déficits cognitifs n'est pas encore bien identifiée. Le but de notre étude est de spécifier et de connaître le rôle des indices acoustiques brefs et leur implication ou non dans la reconstruction cognitive de la parole dégradée par des sujets sains.

1.1. La reconstruction de la parole

La compréhension de la parole chez les sujets normo-entendants est une faculté cognitive très robuste qui résiste aux variabilités acoustiques intrinsèques du signal de parole. Dès 1970, l'expérience de Warren met en évidence cette capacité de restauration cognitive : lorsqu'un phonème est remplacé par un bruit à l'intérieur d'un mot, le sujet perçoit toujours le mot dans sa totalité [6]. Cependant, les capacités de reconstruction dépendent à la fois de la nature et du degré de distorsion appliqués au signal. Dans notre étude, nous nous sommes intéressées à la dimension temporelle du signal acoustique en accélérant certains segments, dans le but d'évaluer l'importance relative de deux indices acoustiques (le Délai d'Etablissement du Voisement, DEV et la Transition du Formant 2, TF2) dans la reconstruction cognitive de la parole dégradée.

1.2. Le délai d'établissement du voisement

Selon Lisker et Abramson [7], le DEV est défini comme l'intervalle de temps entre l'explosion de l'occlusive et le début du voisement. Le DEV peut-être négatif si le voisement débute avant la fin de l'explosion, nul si la synchronisation est parfaite et positif si le voisement commence un certain temps après la fin de l'explosion. Ce dernier cas correspond au phénomène de l'aspiration. Ainsi, les valeurs de DEV donnent des informations sur le degré de voisement des consonnes. De récentes études, utilisant un continuum de DEV, ont cherché à déterminer la valeur seuil du DEV pour laquelle la confusion entre deux consonnes apparaît [8]. Cette confusion rend compte de la notion de perception catégorielle qui est définie comme une frontière d'identification entre deux sons de catégories différentes [9].

1.3. La Transition Formantique

La TF correspond à un changement rapide de fréquence au moment de l'explosion de la consonne occlusive. Les changements rapides de fréquence sont primordiaux pour l'identification des segments acoustiques. La transition du second formant est un indice pour déterminer le lieu d'articulation des occlusives. Serniclaes et al. [10] ont utilisé un continuum [ba]-[da] dans lequel ils modifiaient la valeur de la fréquence initiale de la transition (second et troisième formants). Ils ont ainsi montré que les enfants

dyslexiques avaient une meilleure perception des différences intra-catégorielles que les enfants sans trouble du langage. Dans notre étude, nous nous intéressons aux effets de la compression temporelle d'indices acoustiques sur l'intelligibilité de la parole sur des sujets sans trouble du langage. Trois expériences ont été effectuées dans lesquelles nous avons accéléré soit les deux indices acoustiques ensemble (Expérience 1), soit les indices séparément (Expérience 2 : le DEV et Expérience 3 : la TF2).

2. MATERIEL ET METHODE

2.1. Expérience 1

Matériel

Les stimuli étaient composés de 64 non-mots bisyllabiques de forme CVCV et de 16 fillers de forme VCV. Quatre consonnes occlusives et deux voyelles ont été combinées pour construire chaque stimulus. Les quatre consonnes occlusives de la langue française sont différenciables par leur voisement (voisée : /b/, /d/ *vs* non voisée : /p/, /t/) et leur lieu d'articulation (labial : /b/, /p/ *vs* dental : /d/, /t/). Les deux voyelles antérieures /a/ et /i/ se différencient par leur degré d'aperture. Chaque consonne apparaît avec chacune des autres consonnes dans les deux syllabes et avec les deux voyelles ($4C_1 x 4C_2 x 2V_1 x 2V_2 = 64$ CVCV). Les stimuli ont été produits par un locuteur français et enregistrés dans une chambre sourde avec un microphone Sony ECM-MS907. Les fichiers sons étaient sauvegardés sous le format wav et échantillonnés à 22 kHz (stéréo, 16 bits). La durée de chaque indice acoustique (le DEV et la TF2) a été mesurée pour chacun des items (figure 1). Le DEV a été segmenté entre le début de l'explosion et le début du voisement. Alors que la TF2 a été délimitée à partir du changement brusque et rapide de la direction du F2 jusqu'à l'état stable de la voyelle. Pour les deux indices de chaque syllabe, la durée a été accélérée selon quatre conditions expérimentales de compression : une condition contrôle correspondant à la durée initiale, une condition 50% correspondant à 50% de la durée initiale, de même, une condition 25% et une condition 0%. La compression temporelle des indices acoustiques se fait avec le logiciel Praat (méthode Pitch-Synchronous Overlap Add: PSOLA). Avec le logiciel Praat, les parties segmentées du signal acoustique peuvent être accélérées alors que le reste du signal reste intact. Ainsi, chaque syllabe peut être accélérée indépendamment l'une de l'autre selon deux conditions expérimentales différentes.

Déroulement de l'expérience

Les participants étaient assis dans une pièce silencieuse face à un écran d'ordinateur. Les stimuli étaient émis en modalité auditive binaurale à l'aide d'un casque (Beyerdynamic DT 48, 200Ω) et présentés dans un ordre aléatoire différent pour tous les participants. Les participants étaient informés qu'un signal de parole, pas nécessairement un mot, allait être présenté dans le casque et qu'ils devaient taper sur le clavier tout ce qu'ils avaient entendu. Un entraînement leur était préalablement proposé.

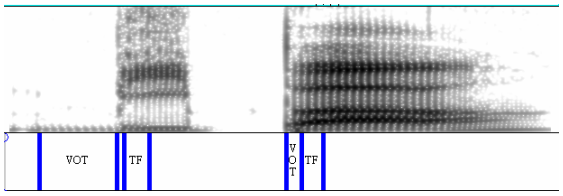

Figure 1 : Spectrogramme du non-mot [bipa]. (VOT=Voice Onset Time; TF=Transition Formantique)

Participants

Trente deux participants, âgés de 18 à 32 ans, de langue maternelle française, et n'ayant jamais connu aucun trouble auditif, du langage, ou neurologique, ont participé à cette expérience.

2.2. Expériences 2 et 3

Matériel et Déroulement des expériences

Dans l'Expérience 2, nous avons accéléré uniquement la durée du DEV alors que dans l'Expérience 3, c'est la durée de la TF2 que nous avons compressée. Ces compressions temporelles suivent les mêmes conditions expérimentales que dans l'Expérience 1. Les stimuli et le déroulement des expériences étaient les mêmes que pour l'Expérience 1.

Participants

Deux groupes de seize participants, âgés de 18 à 23 ans (Expérience 2) et âgés de 19 à 23 ans (Expérience 3), de langue maternelle française, et n'ayant jamais connu aucun trouble auditif, du langage, ou neurologique, ont participé à cette expérience. De plus, aucun n'avait participé à l'Expérience 1.

3. RESULTATS

3.1. Expérience 1 : Compression DEV et TF2

Nous avons calculé les pourcentages d'identification des participants pour les items, les consonnes et les voyelles. Globalement, les voyelles étaient mieux conservées et identifiées que les consonnes. De plus, selon la position (première ou deuxième syllabe), l'intelligibilité de la consonne était différemment modulée par la compression. En effet, nous avons observé que la première consonne était moins bien identifiée que la seconde (Table 1).
Une ANOVA à deux facteurs incluant la Position (première syllabe S1/seconde syllabe S2) et les 4 conditions de Compression a mis en évidence un effet significatif de la Position [$F(1, 31)= 28.97$, $p<.05$], un effet significatif de Compression [$F(3, 93)=652.32$, $p<.05$] et un effet d'interaction [$F(3, 93)=5.75$, $p<.05$] : les performances d'identification de la consonne dépendent de sa position à l'intérieur du non-mot. De plus, le taux d'identification a été sensible à toutes les conditions de compression. Une ANOVA à trois facteurs incluant le Voisement (voisé/non voisé), le Lieu d'articulation (labial/dental) et les conditions de

Compression a été également réalisée. Pour C1, nous avons observé un effet significatif du Voisement [F(1, 31)=544.50, p<.05], du Lieu d'articulation [F(1, 31)=9.62, p<.05] et de Compression [F(3, 93)=409.19, p<.05]. Il est à noter que les occlusives voisées étaient mieux identifiées que les occlusives non voisées aux conditions 50%, 25% et 0% et que les consonnes labiales étaient mieux identifiées que les dentales aux conditions 25% et 0%. L'analyse des erreurs des participants a montré que les occlusives non voisées étaient totalement perdues et jamais confondues. De la même manière, nous avons observé pour C2, un effet significatif du Voisement [F(1, 31)=22.19, p<.05], du Lieu d'articulation [F(1, 31)=8.03, p<.05] et de Compression [F(3, 93)=356.71, p<.05]. Contrairement à C1, les occlusives non voisées étaient mieux identifiées que les occlusives voisées aux conditions de compression 25% et 0%, alors que les consonnes labiales étaient mieux identifiées que les dentales seulement à 0%. Un autre résultat intéressant était la confusion du /b/ et du /d/ avec la consonne liquide /l/. Par ailleurs, nous avons observé une grande variabilité inter-individuelle des performances. Cette variabilité entre les 32 participants est plus importante pour la condition 25% pour les deux consonnes (SD = 0.13 et SD = 0.15). Par exemple, pour C1, les performances vont de 94% pour le meilleur participant à 25% d'identification pour le moins bon. Cette variabilité reflète probablement une différence dans les capacités à reconstruire la parole dégradée, alors même que les participants n'avaient rapporté aucun trouble ni du langage, ni auditif. Cependant, des différences de capacités auditives, sans être pathologiques, pourraient expliquer ces résultats.

3.2. Expérience 2 : Compression DEV

Globalement, la moyenne du taux d'identification était plus haute que dans l'Expérience 1. Les voyelles étaient toujours mieux identifiées que les consonnes. Et, on observait toujours l'effet de la position (table 1).

Une ANOVA à deux facteurs a montré un effet significatif de la Position [F(1, 15)=6.04, p<.05], un effet significatif de Compression [F(3,45)=125.26, p<.05] mais pas d'interaction [F(3,45)<1]. L'identification était significativement différente seulement pour la condition 0% (p<.05). Donc, le DEV doit être totalement supprimé pour induire des troubles de la perception. Dans l'ANOVA à trois facteurs, pour C1, nous avons mis en évidence un effet significatif du Voisement [F(1, 15)=39.68, p<.05], du Lieu d'articulation [F(1, 15)=6.90, p<.05] et de Compression [F(3, 45)=65.85, p<.05]. Comme dans l'Expérience 1, nous avons noté que les occlusives voisées étaient mieux identifiées que les non voisées à 50%, 25% et 0%. De plus, les labiales étaient mieux identifiées que les dentales à 25% et 0%. La nature des erreurs faites par les participants montrent que les occlusives non voisées sont la plupart du temps perdues. De la même manière, nous avons observé pour C2, un effet significatif du Voisement [F(1, 15)=10.53, p<.05], de Compression [F(3, 45)=58.07, p<.05] mais pas du Lieu d'articulation [F(1, 15)=2.50, n.s.]. Contrairement à C1, les occlusives non voisées étaient mieux identifiées que les occlusives voisées seulement à la condition 0%. La consonne /p/ était toujours la mieux identifiée alors que les consonnes /b/ et /d/ étaient confondues avec la consonne liquide /l/ à la condition 0%. Par ailleurs, une grande variabilité inter-individuelle était observée uniquement à 0% ce qui soulignerait la redondance des indices acoustiques.

Table 1 : Moyenne des taux d'identification (%) pour les trois expériences. Le signe (-) correspond à l'accélération et le signe (+) correspond à la durée initiale de l'indice.

	C1	C2	V1	V2	Items
Exp. 1 **DEV - / TF -**	70.4	77	98.9	99.1	55.4
Exp. 2 **DEV - / TF +**	87.3	90.1	99.8	99.8	78.9
Exp. 3 **DEV + / TF -**	93.6	97.8	100	99.9	91.2

3.3. Expérience 3 : Compression TF2

Dans l'Expérience 3, la moyenne du taux d'identification était plus haute que dans les deux premières. Un effet de plafonnement des performances semble apparaître dans cette expérience. Néanmoins, les voyelles étaient toujours mieux identifiées que les consonnes. Et, l'effet de la position était toujours démontré (Table 1).

Une ANOVA à deux facteurs a montré un effet significatif de la Position [F(1, 15)=36.54, p<.05], un effet significatif de Compression [F(3,45)=24.68, p<.05] et un effet d'interaction [F(3,45)=6.52, p<.05]. Nous avons montré également que le taux d'identification était significativement différent seulement pour la condition 0% (p<.05). Donc, la TF doit être totalement supprimée pour induire des difficultés de reconstruction cognitive. Dans l'ANOVA à trois facteurs, pour C1, nous avons mis en évidence un effet significatif du Lieu d'articulation [F(1, 15)=19.29, p<.05], de Compression [F(3, 45)=19.40, p<.05] mais pas du Voisement [F(1, 15)=1.67, n.s.]. De la même manière, nous avons observé pour C2, un effet significatif du Lieu d'articulation [F(1, 15)=12.45, p<.05], de Compression [F(3, 45)=11.52, p<.05] mais pas du Voisement [F(1, 15)=3.85, n.s.]. Dans les deux positions, les labiales étaient mieux identifiées que les dentales. Et nous avons souligné la fréquente confusion de la consonne /d/ avec la consonne /b/, ce qui met en évidence une erreur de lieu d'articulation quand la TF est modifiée. Par ailleurs, l'effet de plafonnement des performances observé expliquerait la faible variabilité inter-individuelle. En effet, la compression temporelle sur la transition formantique ne semble pas gêner les participants à percevoir les stimuli. La tâche était donc trop facile d'où l'effet « plafond » obtenu. Cette faible variabilité peut également refléter la redondance des indices acoustiques grâce à la compensation de la transition altérée par le DEV intact.

4. DISCUSSION

Dans cette étude, nous avons observé les effets de la compression temporelle d'indices acoustiques brefs sur

l'intelligibilité de la parole par des sujets normo-entendants. Nous avons accéléré soit les deux indices (le DEV et la TF2) en même temps (Expérience 1) soit les deux séparément (Expériences 2 et 3). En résumé, les voyelles étaient mieux identifiées que les consonnes et la consonne en attaque était moins bien perçue que la consonne intervocalique. De plus, les consonnes labiales étaient mieux rappelées que les consonnes dentales même si ces observations dépendent des indices acoustiques manipulés. Pour la première syllabe, les occlusives voisées sont mieux identifiées que les non voisées alors que l'effet inverse est observé pour la seconde syllabe. Cependant, la manipulation des indices séparément module ces effets : les performances restent très bonnes jusqu'à totale suppression de la TF2 dans l'Expérience 3. Ce résultat souligne la redondance des indices présents dans le signal de parole et qui sont impliqués dans la reconstruction cognitive de la parole dégradée. La grande variabilité inter-individuelle des performances observée pour la condition 25% dans l'Expérience 1 est également observée dans une moindre mesure pour la condition 0% dans les expériences 2 et 3. Dans les expériences 2 et 3, l'indice intact permet de compenser largement la dégradation. La redondance des informations acoustiques temporelles permet aux mécanismes de reconstruction cognitive d'être activés et de retrouver la bonne syllabe. La voyelle est toujours mieux identifiée que la consonne car la zone de l'état stable de la voyelle varie très peu par rapport à la forme acoustique de la consonne.

En ce qui concerne l'effet de position entre les deux syllabes, la différence de longueur classiquement décrite en Français - les consonnes intervocaliques (S2) sont plus longues que les consonnes d'attaque (S1) - pourrait expliquer pourquoi les participants font moins d'erreurs sur S2 que sur S1. Par ailleurs, l'information apportée par la transition finale de la voyelle dans S1 pourrait aider à reconstruire S2. Pour l'effet de voisement entre les deux syllabes, le contexte acoustique pourrait jouer un rôle important. Dans un contexte voisé (S2), les consonnes non voisées sont plus saillantes alors que dans un contexte silencieux (S1), les consonnes voisées émergent plus. La nature des erreurs montre cet effet, dans S1, les consonnes non voisées (/p/, /t/) sont la plupart du temps perdues mais pas confondues. Et en S2, les consonnes voisées (/b/, /d/) sont majoritairement confondues avec la liquide /l/. La confusion du /d/ correspond à une erreur de mode d'articulation mais pour le /b/, il y a en plus une erreur de lieu d'articulation. Des études en Anglais ont montré que les occlusives dentales sont souvent prononcées comme des liquides (e.g., /rider/). Au niveau articulatoire, quand une occlusive dentale est accélérée, la réduction de la constriction rappelle une friction de liquide. De manière générale, la meilleure identification des consonnes labiales est démontrée par la plus grande perte des /t/ que des /p/ en S1 et par une grande perte des /t/ et une confusion importante des /d./ en /l/ en S2. De même dans l'Expérience 3, cet effet est visible par une importante confusion du /d/ en /b/. Cette erreur de lieu d'articulation reflète l'effet de l'altération de TF2. Cependant, dans l'expérience 2, l'effet de lieu d'articulation sur S2 n'est pas observé à cause d'une meilleure identification du /t/. En résumé, les résultats nous permettent d'identifier l'importance de chaque indice dans l'identification de la parole. La TF2 intacte va améliorer uniquement l'identification du /d/ (Expérience 2) alors que le DEV original va améliorer l'identification de la majorité des consonnes (Expérience 3).

5. CONCLUSION

Les indices acoustiques ont des rôles spécifiques qui peuvent se compenser. Leur redondance est utilisée pour reconstruire le signal de parole dégradée. Cependant, le DEV semble tout de même plus robuste aux dégradations temporelles que la TF2.

BIBLIOGRAPHIE

[1] F. Meunier, T. Cenier, M. Barkat, and I. Magrin-Chagnolleau. Mesure d'intelligibilité de segments de parole à l'envers en français. In *proc. of XXIVèmes Journées d'Etude sur la Parole,* pages 117-120, 2002.

[2] W. Serniclaes. Etude expérimentale de la perception du trait de voisement des occlusives du Français. *Ph. D. Dissertation, Université Libre de Bruxelles, 1987.*

[3] R. D. Kent and K. L. Moll. Vocal-tract characteristics of the stop cognates. *Journal of Acoustical Society of America,* 46:1549-1555, 1969.

[4] L. Lisker and A. S. Abramson. Some effects of context on voice onset time in English stops. *Language and Speech,* 10:1-28, 1967.

[5] P. Tallal and M. Piercy. Developmental aphasia: rate of auditory processing and selective impairment of consonant perception. *Neuropsychologia,* 12:83-93, 1974.

[6] R. M. Warren. Perceptual restoration of missing speech sounds. *Science,* 167:392-393, 1970.

[7] L. Lisker and A. S. Abramson. A cross-language study of voicing in initial stops: acoustical measurements. *Word,* 20:384-422, 1964.

[8] B. McMurray, M. K. Tanenhaus and R. N. Aslin. Gradient effects of within-category phonetic variation on lexical access. *Cognition,* 86:B33-B42, 2002.

[9] A. M. Liberman, K. S. Harris, H. S. Hoffman and B. C. Griffith. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54:358-368, 1957.

[10] W. Serniclaes, L. Sprenger-Charolles, R. Carré and J. F. Demonet. Perceptual discrimination of speech sounds in developmental dyslexia. *Journal of Speech, Language, and Hearing Research,* 44:384-399, 2001.