# Contexte et nature des réalisations phonétiques en parole conversationnelle

*Christine Meunier*
LPL, UMR 7309, 5 av. Pasteur – 13604 Aix-en-Provence
Christine.Meunier@lpl-aix.fr

RESUME

Depuis une dizaine d'années, les recherches en phonétique se sont tournées avec intérêt vers la description des grands corpus de parole naturelle, non lue. Ce nouveau terrain d'investigation ouvre de nombreuses perspectives mais pose également de nouvelles questions aux phonéticiens. Ce papier évalue, dans un premier temps, le contexte lexical et phonologique dans lequel les réalisations phonétiques sont produites, contexte très différent de celui des corpus construits. Ensuite, nous abordons la question de l'annotation, déterminante pour les analyses phonétiques. Enfin, nous évoquons quelques cas spécifiques de réduction phonétique qui offrent de nouvelles perspectives pour nos interprétations concernant la production de la parole.

ABSTRACT

**Context and nature of phonetic realizations in conversational speech**

Since a decade, research in phonetics has turned with interest to the description of large corpora of casual speech. This new field of research opens up many opportunities but asks also new questions for phoneticians. Firstly, this paper evaluates the lexical and phonological context in which phonetic realizations are produced. This context is noticeably different from lexical context in constructed corpora. Next, we address the question of phonetic annotation which is critical for phonetic analyses. Finally, we discuss some specific cases of phonetic reduction which offer new perspectives for our interpretations of speech production.

MOTS-CLES : parole spontanée, grands corpus, données lexicales, annotation phonétique, alignement automatique, réduction phonétique.

KEYWORDS : spontaneous speech, large corpora, lexical data, phonetic annotation, phonetic reduction.

## 1 Introduction

Depuis une dizaine d'années, les recherches en phonétique se sont tournées avec intérêt vers la description de types de parole naturelle, non lue. Nous faisons ici une distinction entre les corpus construits a priori par l'expérimentateur (lecture de sons, syllabes ou mots produits isolément ou dans des phrases porteuses, textes, etc.) et les corpus non construits par l'expérimentateur mais exploités a posteriori (parole spontanée, interviews, récits, conversations, etc.). L'analyse de cette deuxième catégorie de corpus implique la prise en compte de contextes variés. Notamment, les dimensions linguistiques telle que l'usage de la syntaxe à l'oral, la structure du discours ou encore l'influence des caractéristiques pragmatiques de la parole en situation naturelle sont autant de facteurs susceptibles d'influencer la production de la parole. Cette influence

peut, à certains égards, modifier nos connaissances sur la réalisation des sons en contexte. Notre intérêt se porte ici sur les corpus non construits dans l'objectif de mieux cerner, à la fois, le contexte mais aussi la nature des réalisations phonétiques. Cette description prend la forme de deux parties: 1/ un inventaire des caractéristiques lexicales et phonologiques est dressé de façon à comprendre en quoi ces caractéristiques sont distinctes de celles présentes dans des corpus construits; 2/ les spécificités phonétiques des corpus non construits sont abordées en évoquant l'impact de l'annotation automatique des corpus de parole sur les pratiques des phonéticiens.

## 2 Caractéristiques linguistiques de la parole conversationnelle

Nos descriptions se basent sur un style de parole spontanée et relâchée. Il s'agit de conversations entre des locuteurs qui se connaissent. Les productions phonétiques de ce style de parole peuvent être très éloignées des réalisations canoniques habituellement observées en parole lue[1]. Le corpus utilisé ici -*Corpus of Interactional Data* (CID, Bertrand et al, 2008)- est un enregistrement audio-vidéo de dialogues spontanés entre des locuteurs français natifs (8h, 16 locuteurs). Une Transcription Orthographique Enrichie (TOE, Bertrand et al, 2008) a été réalisée et corrigée manuellement. A partir de cette TOE, un convertisseur graphème-phonème suivi d'un aligneur permettent d'obtenir une suite phonétique alignée sur le signal de parole.

### 2.1 Contexte lexical des productions phonétiques

La distribution des formes lexicales présentes dans le CID est semblable à celle que l'on trouve dans l'ensemble des corpus oraux spontanés et des grands corpus textuels. On y retrouve les caractéristiques de la loi de Zipf avec des mots dont la fréquence est inversement proportionnelle à leur rang dans le corpus. Les formes lexicales n'apparaissant qu'une seule fois dans le corpus sont au nombre de 3259 tandis que la forme la plus fréquence apparaît 3130 fois, ce qui est comparable à d'autres corpus de parole spontanée (Torreira, 2010). En revanche, la spécificité des mots fréquents des corpus conversationnels repose sur la nature de ces mots ("ouais", "je", "tu", etc.) que l'on ne retrouve pas avec cette fréquence dans des corpus radiophoniques ou des textes journalistiques. Les 8 heures du CID contiennent 6611 formes lexicales différentes tandis que la totalité des occurrences est de 102457. La moitié des occurrences du corpus totalise seulement 39 formes différentes.

| Forme | Est | c' | ouais | et | de | tu | pas | je | ça | le |
|---|---|---|---|---|---|---|---|---|---|---|
| Nbre | 3130 | 3018 | 2916 | 2679 | 2033 | 2027 | 1895 | 1893 | 1817 | 1655 |
| % | 3,05 | 2,95 | 2,85 | 2,61 | 1,98 | 1,98 | 1,85 | 1,85 | 1,77 | 1,62 |

TABLE 1 – Occurrences des 10 formes lexicales les plus fréquentes

On remarquera que les mots les plus fréquents (table 1) sont courts et la plupart sont des mots fonction (déterminants, conjonction, etc.). Plusieurs travaux ont pu montrer que ces caractéristiques lexicales ont une influence sur les productions phonétiques (Johnson,

[1] Voir à ce propos le *Special Issue on Speech Reduction* (*Journal of Phonetics,* vol. 39, n°3) dans lequel de nombreux articles décrivent ces phénomènes en parole spontanée.

2004; Meunier & Espesser, 2011). Sur les 39 formes les plus fréquentes, seulement 3 sont bisyllabiques ("était", "enfin", "avait") et aucune n'est un nom. 57% des mots du corpus sont des monosyllabes. Les quatre compositions syllabiques les plus fréquentes totalisent plus de 50% du corpus et regroupent des mono- et des bisyllabes (table 2).

| | monosyllabes | | bisyllabes | |
|---|---|---|---|---|
| Forme | CV | V | CV.CV | V.CV |
| % | 22 | 17,5 | 7 | 6 |

TABLE 2 – Décomposition syllabique des mots les plus fréquents du corpus.

Ces caractéristiques rendent l'exploitation statistique des données très délicate. En effet, plusieurs facteurs sont en interaction et il est difficile de les exploiter séparément. Comparer le mot de contenu fréquents et rares revient à comparer très peu de mots répétés de très nombreuses fois à une grande quantité de mots répétés très peu de fois.

## 2.2 Contexte phonologique des productions phonétiques

### 2.2.1 Structures syllabiques

La décomposition syllabique du corpus montre un total de 139751 syllabes[2]. La fréquence syllabique est, là encore, conforme à ce que l'on peut trouver dans des corpus de textes journalistiques ou les bases de données lexicales (Goldman et al., 1996), la structure CV étant de loin la plus fréquente et représentant, à elle seule, plus de la moitié des syllabes produites. Les six structures syllabiques les plus fréquentes représentent 99% des syllabes du corpus, ce qui rend les autres structures très marginales (table 3).

| Forme | CV | V | CVC | CCV | CCVC | VC |
|---|---|---|---|---|---|---|
| % | 60,5 | 13 | 11,5 | 10,5 | 2 | 1,5 |

TABLE 3 – Structures syllabiques les plus fréquentes.

### 2.2.2 Les phonèmes du corpus

Les phonèmes[3] sont issus de la transcription du corpus pour laquelle les experts avaient la possibilité d'indiquer les élisions (grâce à la TOE, Bertrand et al., 2008). Les 272166 phonèmes (53% de consonnes) sont donc ceux qui ont été perçus et transcrits par les experts. Pour le CID, les voyelles à timbre variable ont été regroupées[4]. La fréquence des phonèmes produits dans le CID est sensiblement comparable à celle que l'on peut trouver dans différentes bases de données (ici *Lexique*[5]). On retrouve donc parmi les phonèmes les plus fréquents les voyelles e, A, @ et les consonnes r, s, t, l (figure 1). La voyelle e est surreprésentée dans le CID en raison de la fréquence du mot "ouais" dans ce corpus alors que ce mot est évidemment absent des corpus utilisés par *Lexique*. En revanche, R est

[2] La syllabation du corpus a été effectuée à l'aide du syllabeur développé au LPL par B. Bigi (Bigi et al., 2010). Cette syllabation est indépendante des frontières lexicales.

[3] Transcrits en code SAMPA: http://www.phon.ucl.ac.uk/home/sampa/french.htm

[4] e et E sont codés e; o et O sont codés o; a et A sont codés A; @ (schwa), 2 et 9 sont codés @). En vue d'une comparaison, les mêmes regroupements ont été effectués pour les fréquences de *Lexique*.

[5] *Lexique*, site réalisé par Boris New & Christophe Pallier (http://www.lexique.org)

sous-représenté dans le CID. Cela peut-être du au fait que cette consonne fait partie des élisions les plus fréquentes (voir plus loin). On notera également la forte représentation de w, ce qui, là encore, peut-être du à la fréquence du mot "ouais".

La majorité des phonèmes sont réalisés dans un nombre de mots très restreints. Par exemple, 50% des réalisations de la voyelle A se trouvent dans seulement 13 mots différents ("pas", "ça", "a", "la", etc.) et 20% des réalisations de la voyelle y se trouvent dans le pronom "tu" (très fréquent en parole conversationnel). Là encore, ces mots sont essentiellement des mots fonction monosyllabiques. Le support lexical de ce type de corpus est ainsi très éloigné de celui utilisé dans des corpus construits.

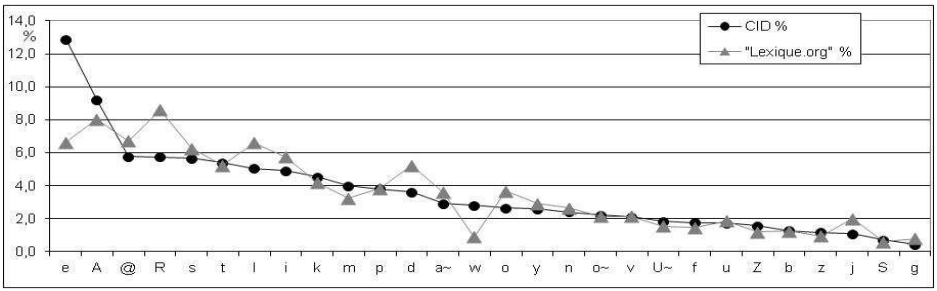


FIGURE 1 – Fréquences des phonèmes du CID comparées à celles extraites de *Lexique.*

### 2.2.3 Réalisations particulières identifiées

Les experts transcripteurs du CID avaient la possibilité de coder les réalisations particulières (variantes perçues) ainsi que les élisions identifiées. 2948 réalisations particulières ont été codées. 20% de ces réalisations concernent les cas spécifiques du pronom "je" suivi d'une consonne non voisée ("s" la plupart du temps). Il s'agit donc pour l'essentiel des séquences "je sais" (produit Se) ou "je suis" (produit SYi). Les deux phonèmes sont ici fusionnés (lieu articulation du premier phonème Z et voisement de la deuxième consonne s, le tout donnant un seul phonème S). Cette forme, très fréquente, est particulièrement bien identifiée par les auditeurs et peut être considérée comme une forme figée, assez prédictible. Les autres réalisations codées sont en lien avec l'accent des locuteurs. Notamment, de nombreux @ ont été ajoutés comme réalisation particulière là où le transcripteur s'attendait à son absence. De même le pronom "tu" est souvent transcrit "ti" ce qui est une spécificité des locuteurs du sud-est.

### 2.2.4 Élisions identifiées

10925 élisions ont été codées. Les dix élisions les plus fréquentes (table 4) représentent 99% des élisions codées dans le corpus. @ est clairement le phonème le plus souvent identifié comme manquant, ce qui n'est pas surprenant puisque ce symbole comprend le schwa dont la présence ou l'absence sont souvent bien identifiées par les auditeurs.

| phonème | @ | l | y | R | a~ | v | e | i | u | d |
|---|---|---|---|---|---|---|---|---|---|---|
| % | 35,8 | 19,1 | 8,4 | 8,6 | 5 | 3,1 | 3,1 | 2,1 | 13 | 1,1 |

TABLE 4 – Les dix élisions les plus fréquemment codées par les experts

# 3 Phonétique des corpus de parole naturelle

Cette deuxième partie est consacrée à une approche phonétique du corpus. Dans un premier temps, la méthodologie concernant l'annotation phonétique est abordée car elle est centrale dans l'exploitation des résultats. Nous verrons en quoi les différentes approches sont autant de sources d'information nouvelles pour l'exploitation des données phonétiques. Nous aborderons enfin les cas spécifiques des phénomènes de réduction dans ce type de parole. Ces phénomènes pourraient nous apporter un éclairage nouveau sur l'interaction entre des contraintes physiologiques et linguistiques.

## 3.1 L'annotation phonétique

Pour être analysés, les corpus de parole ont besoin d'être annotés. L'annotation phonétique sur des grands corpus est difficilement envisageable manuellement tant elle est coûteuse en temps. Ainsi, les processus automatiques tels que l'alignement basé sur les transcriptions des experts fournissent une annotation indispensable à l'exploitation des grands corpus. Ces annotations ont parfois besoin d'être corrigées par un expert selon le type d'analyse envisagé (Fougeron et al., 2010). Ainsi, la plupart du temps, l'annotation phonétique est réalisée en plusieurs phases: transcription, alignement automatique puis, selon les besoins des analyses, corrections manuelles par des experts. Il ne s'agit donc pas de choisir entre annotation manuelle et automatique mais plutôt de les utiliser de façon complémentaire. Ces différentes phases sont autant d'étapes permettant de nouvelles perspectives pour les analyses phonétiques.

La **transcription** revêt une importance considérable pour la phonétisation (Bigi et al., 2012): 1/ elle permet de minimiser les erreurs de l'aligneur (le codage des réalisations particulières et des élisions évitent des annotations erronées); 2/ elle est une source d'information considérable concernant les variations perçues par les auditeurs. Ce deuxième point nous permet de distinguer les variations phonologiques ou stéréotypées (perçues par les locuteurs) des variations non perçues et souvent non prédictibles (voir plus loin 3.2).

**L'alignement** automatique, et plus précisément les erreurs qu'il produit, fournit de précieuses informations concernant la localisation des zones déviantes (Fredouille & Pouchoulin, 2011). Il est ainsi possible d'utiliser les résultats de l'alignement pour identifier les séquences de forte réduction en localisant, par exemple, les suites de segments ayant la durée minimale affectée par l'aligneur.

La **correction** de l'alignement est désormais souvent la seule occasion dont dispose le phonéticien pour expertiser les réalisations phonétiques et ainsi identifier les phénomènes spécifiques de la parole spontanée. En effet, le risque est grand d'utiliser uniquement les annotations automatiques et, ainsi, de passer à côté des caractéristiques phonétiques de la parole en situation naturelle telle que le phénomène de **réduction**.

## 3.2 Réductions, altérations et variations

Sous le terme de *réduction* on entend aussi bien les élisions que les altérations des segments phonétiques. On distinguera trois types de réduction qui n'impliquent pas les mêmes conséquences aussi bien pour les auditeurs que pour l'interprétation linguistique:

- Les réductions *phonologiques* telles que la chute du schwa; ces formes de réduction sont intégrées dans les modèles phonologiques; elles sont souvent prédictibles et reproductibles; leur niveau de dépendance est phonologique.
- Les réductions *stéréotypées* ou *figées*, telles que celles que l'on observe couramment en parole naturelle sur des séquences identiques ("je ne sais pas" devient Sepa); ces formes sont souvent prédictibles et reproductibles; leur niveau de dépendance peut être lexical, dialectal ou individuel.
- Les réductions *opaques*; ce terme est utilisé volontairement car nous avons peu de connaissance sur ces phénomènes; elles ne sont, en général, pas perçues par les transcripteurs et ne sont donc pas rendues visibles dans l'alignement automatique (figure 2); ces formes sont difficilement prédictibles ou reproductibles d'un point de vue phonologique ou lexical car il est probable que le niveau de dépendance se situe en amont (prosodie, discours, etc.).

On notera que, pour les réductions *phonologiques* ou *stéréotypées*, les transcripteurs sont à même de coder la réalisation particulière ou l'élision car elles sont perceptibles. En revanche, le codage de la troisième catégorie est beaucoup plus aléatoire. Ces réductions sont d'autant plus difficiles à coder qu'elles ne répondent pas à notre codage "discret" des élisions. Nous considérons souvent que les réductions se manifestent par l'absence d'un phonème. Or, très souvent, les réductions *opaques* relèvent d'une fusion ou coalescence entre plusieurs segments qui rend impossible l'identification des segments préservés ou omis (figure 2).

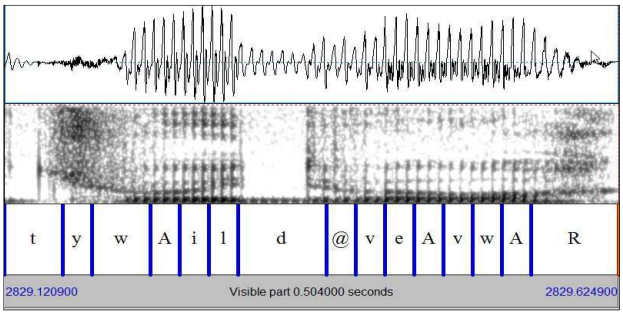


FIGURE 2 – Séquence transcrite "tu (v)ois il devait avoir". Seul le v est codé comme manquant. Annotation issue de l'aligneur.

Une première hypothèse consiste à considérer qu'il s'agit d'hypo-articulation répondant à des contraintes physiologiques et dont la conséquence est une sous-spécification du niveau phonétique lorsque la redondance de l'information linguistique le permet. Dans cette hypothèse, ces productions ne contribuent pas à l'information. La compréhension du message serait alors garantie par des informations descendantes permettant de pallier la sous-spécification phonétique (Warren & Obusek, 1971). L'observation des réductions présentes dans le CID nous amènent à soutenir une autre hypothèse: certains de ces phénomènes seraient régis par des contraintes physiologiques mais répondraient également à des contraintes du système linguistique. Dans plusieurs cas, nous avons pu noter que le processus de réduction tendait à préserver des caractéristiques phonétiques porteuses d'information. Par exemple, dans la figure 3, le A est transcrit mais n'est pas réalisé, la production réelle est donc sdveet. Habituellement, dans ce contexte, on devrait

trouver une assimilation de voisement entre s et d, ce qui n'est pas le cas. Notre hypothèse est que la perception correcte de la séquence est préservée par cette absence d'assimilation "témoin" de la présence sous-jacente du A.

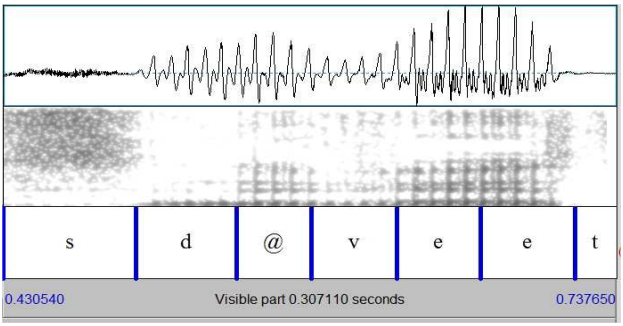


FIGURE 3 – Séquence transcrite "ça devait ê(tr)". La segmentation est proposée par un expert humain.

De même, dans la séquence "une assistante" (figure 4), le transcripteur n'a pas noté d'élision alors que le signal semble indiquer une suite assa~t. L'observation du signal devrait nous amener à considérer que les segments ist ont été omis, mais l'écoute de la séquence ne va pas dans ce sens. Il est probable qu'ici des indices spectraux et temporels permettent à l'auditeur de percevoir la version canonique et non la version réduite.

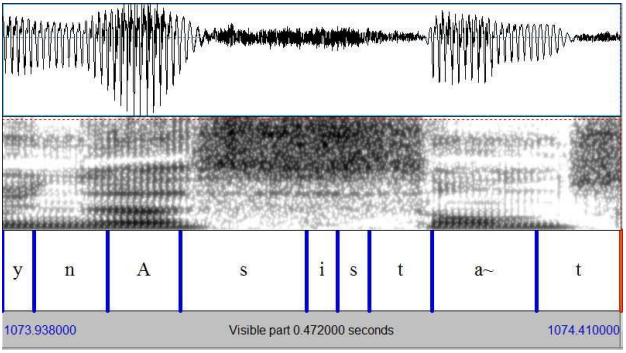


FIGURE 4 – Séquence transcrite "une assistante". Aucun phonème n'est codé comme manquant. Annotation issue de l'aligneur.

Notre hypothèse est que, dans de nombreux cas, les réductions *opaques*, contrairement aux réductions *phonologiques*, ne correspondent pas systématiquement à l'élision simple d'un segment mais sont caractérisées par des processus articulatoires complexes et variés tendant à préserver des indices pertinents qui rendent accessible l'information phonétique, et donc le message linguistique.

## 4 Conclusion

Nos connaissances physiologiques, physiques ou linguistiques des sons du langage se sont considérablement développées au cours du XXème siècle. Toutefois, les travaux portant sur des corpus de parole lue ont conduit à une vision assez figée des productions

sonores. Nous savons désormais que le contexte linguistique de la parole lue est très éloigné des situations de production non contrôlées. Sans remettre en question les résultats obtenus en parole lue, les travaux récents sur les productions phonétiques en parole spontanée questionnent le lien entre contraintes physiologiques et contraintes linguistiques. L'analyse des phénomènes de réduction (notamment *opaques*) est toutefois extrêmement complexe car ils sont non reproductibles (chaque séquence semble unique concernant le contexte phonétique et les unités lexicales impliquées), peu accessibles d'un point de vue perceptif et peu adaptés aux analyses acoustiques (certains gestes articulatoires pourraient avoir un rôle perceptif important sans laisser d'indices acoustiques interprétables). Il semble donc nécessaire d'envisager les analyses phonétiques au travers de méthodologies complémentaires telles que le traitement automatique de la parole, l'expertise phonétique, la prise en compte d'autres niveaux linguistiques et l'enregistrement de données articulatoires.

## Remerciements

Ce travail a été réalisé grâce au soutien financier du projet OTIM (Philippe Blache, LPL, ANR BLAN08-2_349062). Remerciements spéciaux à R. Espesser, B. Bigi et S. Rauzy.

## Références

BERTRAND, R., BLACHE, P., ESPESSER, R., FERRE, G., MEUNIER, C., PRIEGO-VALVERDE, B., RAUZY, S. (2008). Le CID - Corpus of Interactional Data - Annotation et exploitation multimodale de parole conversationnelle. *in Traitement Automatique des Langues*, 49, 105-134.

BIGI, B., PERI, P., BERTRAND, R. (2012). Influence de la transcription sur la phonétisation automatique de corpus oraux. *Actes des XXIXèmes journées d'Etudes sur la Parole*, Grenoble, Juin 2012.

BIGI, B., MEUNIER, C., NESTERENKO, I., BERTRAND, R. (2010). Syllable boundaries automatic detection in spontaneous speech. *In proceedings LREC*, malte, mai 2010, 3285-3292.

FOUGERON, C., AUDIBERT, N., FREDOUILLE, C. MEUNIER, C. GENDROT, C., PANSERI, O. (2010). Comparaison d'analyses phonétiques de parole dysarthrique basées sur un alignement manuel et un alignement automatique. *Actes des XXVIII Journées d'Etude sur la Parole*, Mons, mai 2010, 365-368.

FREDOUILLE, C., POUCHOULIN, G. (2011). Automatic detection of abnormal zones in pathological speech. *Proceedings of ICPhS 2011*, Hong-Kong, 699-702.

GOLDMAN, J.PH., CONTENT, A., FRAUENFELDER, U.H. (1996). Comparaison des structures syllabiques en français et en anglais. *Actes des XXIèmes Journées d'Etudes sur la Parole*, Avignon.

JOHNSON, K., (2004). Massive reduction in conversational American English. In: Yoneyama, K., Maekawa, K. (Eds.), *Spontaneous Speech: Data and Analysis. Proc. 1st Session of the 10th Internat. Symposium*, Tokyo, Japan, 29–54.

MEUNIER, C., ESPESSER, R. (2011) "Vowel reduction in conversational speech in French: The role of lexical factors", Journal of Phonetics, Vol. 39, Issue 3, 271-278.

TORREIRA, F., ADDA-DECKER, M., & ERNESTUS, M. (2010). The Nijmegen corpus of casual French. *Speech Communication*, 52, 201-212.

Warren, R.M. & Obusek, C.J. (1971). Speech perception and phonemic restauration, *Perception & Psychophysics*, 9, 358-362