

Constitution de corpus de parole semi-spontanée en environnement bruyant : intérêts et applications d'une telle méthodologie

Maëva Garnier¹, Danièle Dubois¹, Nathalie Henrich²

¹Laboratoire d'Acoustique Musicale. CNRS UMR 7604, UMPC, Ministère de la culture. 11 rue de Lourmel, 75015 Paris.

²Institut de la communication parlée, CNRS UMR 5009, INPG, Univ. Stendhal, 46 av. F. Viallet, 38031 Grenoble.

Courriel : garnier@lam.jussieu.fr

ABSTRACT

Speech adaptation in noisy environment is not only a physiological reflex but also induced by the search for intelligibility enhancement. Therefore, it seems necessary to study Lombard speech produced in situation of interactive communication. We will present two different game-oriented tasks conceived for the constitution of acoustic and articulatory corpuses of quasi-spontaneous speech produced in noise. Their principle, advantages and drawbacks will be discussed.

1. INTRODUCTION

La parole en environnement bruyant est d'un grand intérêt, tant du point de vue phonétique, phonologique ou de la reconnaissance automatique de la parole. Pour parvenir à s'entendre et à maintenir un bon niveau d'intelligibilité, les locuteurs ont en effet besoin d'adapter leur parole aux niveaux acoustique, phonétique, articuloire, ou prosodique. Ce phénomène est bien connu sous le nom d'effet Lombard [Jun93, Sum88]. Les modifications de la parole provoquées par l'immersion d'un locuteur dans un niveau de bruit moyen accroissent son intelligibilité [Dre57] tandis qu'elles diminuent -pour l'instant- les performances de systèmes de reconnaissance automatique, basés sur des modèles de parole produite en situation calme [Jun93]. La caractérisation de la parole Lombard a donc un intérêt pour mieux comprendre la notion d'intelligibilité mais aussi pour accroître la performance des systèmes de communication homme-machine utilisés dans des lieux publics bruyants.

Etudier la parole Lombard produite en tâche de lecture pose plusieurs problèmes. Tout d'abord, il est bien connu que les caractéristiques rythmiques et intonatives de la parole lue diffèrent de la parole spontanée, même en situation calme [Lie85, Aye94]. Les systèmes de communication homme-machine étant rarement confrontés à de la parole lue, il peut donc déjà être discutable de construire des modèles de reconnaissance sur ce type d'énonciation. Mais le problème fondamental réside surtout dans le fait que la parole Lombard est une adaptation motivée par la recherche d'intelligibilité. Les tâches classiques de lecture mises en œuvre en laboratoire n'intègrent pas la présence d'un interlocuteur ni la non connaissance préalable du discours produit par le locuteur. Cela n'est en effet pas toujours nécessaire et l'analyse de parole lue peut être considérée comme valide pour l'étude de nombreuses problématiques. Par contre, cette approximation doit être remise en cause pour l'étude de la parole Lombard, produite justement dans le but de

compenser la dégradation de l'intelligibilité et non pas seulement en réponse réflexe à l'atténuation du retour auditif.

L'idée d'étudier la parole totalement spontanée serait idéale, mais malheureusement assez utopique. Elle a très vite été abandonnée au profit de la parole lue, pour satisfaire à une démarche scientifique imposant la paramétrisation d'un phénomène et le contrôle des conditions expérimentales, nécessaires à la répétabilité des mesures et à la comparaison de plusieurs conditions ou de plusieurs individus.

Nous pensons personnellement que les conditions expérimentales doivent être contrôlées au maximum, mais non au point de passer à côté de l'objet d'étude. Aucune solution n'existant pour maximiser à la fois la spontanéité du discours et le contrôle du matériel linguistique, l'étude de la parole Lombard nécessite un compromis adapté au niveau de détail des questions que l'on souhaite étudier. Nous considérons ce compromis comme un choix méthodologique lucide et sérieux d'un point de vue épistémologique. Un tel compromis est réalisable avec une tâche semi-spontanée permettant de contrôler une partie du matériel linguistique tout en laissant de l'initiative et de la spontanéité aux locuteurs.

De telles tâches ont commencé à être développées pour l'étude de l'interaction entre les individus (collaborations, négociation, etc.). Ces tâches de description de cartes ("map tasks") [And91], de route [Lev83], ou de jeu de tangram [Cla86] ont ensuite été utilisées pour l'étude intonative de la parole, complétées par d'autres jeux [Sch 01, Pen]. Des dispositifs de "magicien d'Oz" initialement dédiés à l'étude de la communication homme-machine ont également ensuite été utilisés pour l'étude de la parole émotionnelle non actée [Aub05].

Nous nous sommes inspirées de ces tâches interactives pour concevoir deux nouveaux jeux plus adaptés aux problématiques que nous souhaitons examiner concernant la parole Lombard. Nous présenterons et discuterons les avantages et inconvénients de ces deux jeux interactifs correspondant à des niveaux de compromis différents entre le contrôle des conditions expérimentales et la spontanéité du discours. Le premier jeu, assez libre, a été conçu pour l'étude globale au niveau des énoncés et des mots, des stratégies individuelles d'adaptation acoustique et glottique en environnement bruyant. L'étude de caractéristiques articuloires et phonétiques plus précises nécessitant le contrôle de la structure syntaxique des énoncés et du contexte segmental, a motivé la conception du deuxième jeu plus contrôlé.

2. UN PREMIER JEU INTERACTIF ASSEZ LIBRE

2.1. Contraintes

Pour pouvoir effectuer des comparaisons entre plusieurs conditions de bruit et plusieurs locuteurs, nous avons choisi de retenir un certain nombre de mots cibles du langage courant, représentant l'ensemble des caractéristiques du français. La plupart des études anglophones sur la parole Lombard utilisent des listes de mots à deux syllabes accentuées de façon équivalente ("spondées") [Hir52], du vocabulaire aéronautique ou des listes de chiffres [Sum88, Jun93], ou encore des commandes vocales à destination d'interfaces homme-machines [Kim05]. Le vocabulaire de Miller et Nicely [Mil55] ou les phrases phonétiquement équilibrées de Harvard [Har69] sont également largement utilisées dans les expériences de production ou de perception de la parole dans le bruit. En français plus particulièrement, la base DB_bruit est constituée de logatomes CVC, de listes de nombres et de phrases [Zei94]. Les tests d'audiométrie et d'intelligibilité se basent souvent sur les listes de Fournier, constituées de mots mono- et bisyllabiques et de phrases phonétiquement équilibrées [Port59].

Dans le but de comparer nos résultats à ces études antérieures, nous avons retenu 16 spondées de deux syllabes, de structure CVCV, tirés ou inspirés de la liste de Fournier, et représentant la grande majorité des phonèmes du français (*bijou, chausson, cochon, dauphin, fusil, guenon, gitans, lagon, mairie, navet, panda, requin, sommet, toupie, vallée, zébu*). Nous avons ensuite cherché à construire une tâche de jeu interactif nécessitant l'usage de ces mots cibles au cours d'un dialogue non attendu à l'avance par l'interlocuteur, de façon à ce que le locuteur recherche réellement à être intelligible dans le cadre d'un vocabulaire limité. Les niveaux de bruit étant parfois très forts (jusqu'à 85dbSPL), et l'ensemble des conditions étudiées étant assez nombreuses, nous avons cherché à éviter que l'expérimentateur ne soit obligé de donner la réplique à chaque locuteur et de subir le bruit pour chaque enregistrement. C'est pourquoi nous avons imaginé une tâche où les locuteurs interagissent par binômes (ici 5 binômes ont été enregistrés), ce qui permet d'acquérir deux corpus en parallèle. Pour les mêmes raisons, nous avons minimisé autant que possible la durée du jeu afin de ne pas exposer de façon inutile les locuteurs au bruit.

2.2. Principe

Le jeu inclut deux partenaires : un "meneur" et un "suiveur". Le meneur dispose d'une carte sur laquelle sont disposées des images correspondant aux mots-cibles. Un chemin est tracé sur cette carte et relie la moitié de ces items (tracé en pointillé sur la figure 1). L'autre moitié des items est libre. Le "suiveur" dispose d'une liste d'associations mettant en relation les items déjà reliés avec chacun des items libres (cf. figure 1). Le but du jeu consiste à ce que les deux interlocuteurs échangent les informations complémentaires dont ils disposent pour parvenir à reconstituer, dans l'ordre, un deuxième chemin reliant les huit items libres (tracé en gras sur la figure 1).

Le "meneur" doit décrire librement au "suiveur" le chemin tracé sur sa carte (représenté en pointillés sur la figure 1). Il peut dire par exemple: "*Je pars du requin*", puis "*Je vais ensuite vers le cochon*", etc. Pour chaque nouvelle étape du chemin citée par le "meneur", le "suiveur" doit lui répondre en lui indiquant quel est l'item libre correspondant sur la liste d'association (cf. figure 1). Il va donc dire par exemple "*Le requin est associé au chausson*", puis "*le cochon va avec la vallée*", etc. Au fur et à mesure, le "meneur" va pouvoir tracer sur sa feuille le nouveau chemin (représenté en gras sur la figure 1) et le "suiveur" va pouvoir noter sur sa feuille l'ordre dans lequel les items libres s'organisent pour constituer le nouveau chemin (cf. figure 1). Lorsque le chemin de la carte du "meneur" a été parcouru jusqu'au bout, le nouveau chemin à trouver est donc normalement découvert par les deux partenaires. Le "meneur" doit alors le récapituler et s'assurer de la confirmation du "suiveur".

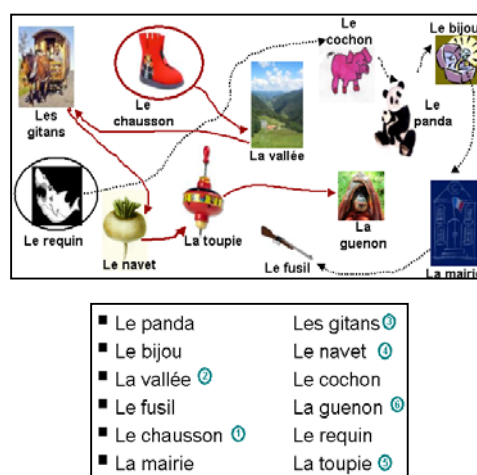


Figure 1 : Exemple pour 12 mots cibles de la carte du "meneur" et de la liste d'associations du "suiveur" à la fin du jeu. Le chemin en pointillé est fourni au "meneur" en début de jeu. Le chemin en gras est tracé au cours du jeu par le "meneur". La liste d'associations est fournie au suiveur en début de jeu. Les chiffres (de 1 à 6) sont notés au cours du jeu par le suiveur et indiquent l'ordre des items constituant le nouveau chemin à découvrir.

2.3. Avantages et intérêts d'un tel corpus

Si deux sujets sur dix ont montré des difficultés de compréhension de la tâche, celles-ci ont totalement disparu après quelques parties d'entraînement. Aucun locuteur n'a montré d'hésitation liée à une charge cognitive trop importante.

L'analyse de la parole semi-spontanée produite par les dix locuteurs a permis de répondre à différents questionnements concernant la parole Lombard [Gar06]. Nous ne présenterons pas ici ces résultats généraux mais plutôt l'analyse d'une petite partie du corpus. Celle-ci présente un grand intérêt pour la compréhension de l'adaptation en environnement bruyant, et apporte des arguments en faveur d'un tel corpus semi-spontané puisqu'elle n'aurait pu être enregistrée en tâche de lecture. Il s'agit des répétitions de mots suite à une

incompréhension manifeste ou explicite de l'interlocuteur. Ces énoncés en focalisation informative nous renseignent sur les représentations du locuteur concernant son intelligibilité, sur la façon dont il la réévalue, et sur la stratégie de surenchère ou de réadaptation qu'il met en œuvre en conséquent. Nous exposerons ici les résultats de la comparaison de la première occurrence des mots avec leur répétition.

Pour chacun des paramètres acoustiques analysés (répertoriés dans la Table 1), nous avons regroupé ensemble les répétitions pour lesquelles le paramètre augmente significativement, diminue significativement, ou ne varie pas significativement. Cela nous a permis d'établir un profil de chaque répétition dans un espace multiparamétrique. Une classification ascendante hiérarchique à partir de ces profils a dégagé 5 stratégies principales de répétition dans le bruit (cf. Table 1):

-la première stratégie est majoritairement adoptée (65% des cas), par des locuteurs des deux genres. Elle consiste à parler plus fort de façon globale. L'augmentation de l'intensité concerne aussi bien les voyelles que les consonnes, et plutôt la 1^{ère} syllabe (S1). Elle s'accompagne de l'augmentation conjointe de la F0 moyenne, de l'énergie dans la zone [2000-5000 Hz] par rapport à l'énergie de la zone [0-2000Hz] (que nous désignerons par la suite comme « timbrage » de la voix), de l'ambitus de F0 et de l'ambitus d'intensité.

-la deuxième stratégie est adoptée dans 16,2% des cas par des locuteurs des deux genres. Elle consiste à ralentir le débit de la parole, aussi bien au niveau des voyelles que des consonnes, et particulièrement sur la 1^{ère} syllabe (S1), prototypée phonétiquement (Nous avons défini ce paramètre de "prototypage" par un déplacement des deux premiers formants dans le sens d'un élargissement du triangle vocalique dans le plan F1/F2). On observe également une augmentation de l'ambitus de F0 et de la dynamique d'intensité, tandis que l'intensité des consonnes diminue. L'intensité et la F0 des 1^{ères} syllabes diminue, avec un prototypage de la 1^{ère} voyelle (V1).

-la troisième stratégie, adoptée dans 11,8% des cas par des locuteurs des deux genres, consiste à prototyper la première voyelle. Cette stratégie s'accompagne de la diminution de l'intensité, de la F0, de la durée des mots, de l'ambitus de F0 et de l'ambitus d'intensité.

-la quatrième stratégie est adoptée dans 4,4% des cas par 2 locutrices sur 10 uniquement. Elle consiste à renforcer le timbrage global sur toutes les syllabes, tout en diminuant l'intensité moyenne, la F0 moyenne, la

dynamique d'intensité ou la durée des mots, voyelles et consonnes. Par contre, la 1^{ère} syllabe est accentuée en intensité, F0 et en ambitus de F0.

-enfin **la cinquième stratégie**, assez marginale (1,5% des mots), peut être considérée comme inefficace, et correspond à une absence de changement entre la première occurrence du mot et sa répétition, exceptée une augmentation de l'intensité des consonnes. Cette stratégie s'accompagne d'une diminution du timbrage, en particulier sur la deuxième voyelle (V2) du CVCV.

Seuls 2 locuteurs sur 10 utilisent exclusivement une seule stratégie (la 1^{ère}) pour chaque répétition demandée. Dans certains cas, la première répétition n'est pas suffisante et est suivie d'une seconde, voire d'une troisième répétition. On observe que les locuteurs ont plutôt tendance à surenchérir sur leur première stratégie lors de la deuxième répétition, tandis qu'ils ont tendance à changer de stratégie à partir de la troisième répétition, témoignant d'une réévaluation de la situation et d'une réadaptation.

2.4. Inconvénients et limites

Une petite partie du corpus (1,5% des mots cibles) n'a pu être exploitée car les mots correspondant étaient bafouillés ou hésités. Cette proportion de données inexploitable nous semble tout à fait raisonnable. Cela implique toutefois la nécessité d'enregistrer un corpus semi-spontané au moins en double pour éviter les mots-cibles manquants. Cela peut par ailleurs être intéressant pour pallier la plus grande variabilité de la parole spontanée par une plus grande consistance du corpus.

Par ailleurs, nous devons mentionner que l'automatisation de l'analyse d'un corpus spontané se trouve légèrement complexifiée du fait que les phrases et l'ordre des mots ne sont pas toujours les mêmes. Cela nécessite la mise au point d'un système d'indexation pour tenir compte des différents cas, positions, ordres, etc.

Le protocole de ce premier jeu est approprié pour explorer la parole Lombard au niveau global des énoncés ou des mots-cibles, mais ne permet pas de réaliser des mesures rigoureuses de durée ou d'intonation, du fait que la position des mots-cibles dans les énoncés n'est pas contrôlée. De même, l'étude détaillée de certains segments phonétiques particuliers nécessiterait le contrôle du contexte vocalique ou consonantique de ces segments.

Table 1: Représentation des cinq stratégies différentes de répétition d'un mot dans le bruit. Chaque stratégie peut être associée à un profil d'évolution de différents descripteurs acoustiques. Les paramètres augmentant pour cette stratégie sont représentés en noir (+), ceux qui diminuent sont représentés en gris clair (-), ceux qui n'évoluent pas en gris foncé (0). Enfin, les paramètres dont l'évolution n'est pas stable au sein d'une même stratégie sont représentés en blanc.

	Intensité	F0	Durée	Timbrage	Ambitus F0	Dynamique d'intensité	Intensité S1	Intensité S2	F0 S1	F0 S2	Durée S1	Durée S2	Timbrage V1	Timbrage V2	Dynamique d'intensité S1	Dynamique d'intensité S2	Ambitus F0 S1	Ambitus F0 S2	Intensité Voyelles	Intensité Consonnes	Durée voyelles	Durée consonnes	Prototypage V1	Prototypage V2
Stratégie1	+	+		+	+	+	+		+	+							+		+	+				
Stratégie2		-	+		+	+	-		-		+				+	+		+		-	+	+	+	0
Stratégie3	-	-	-		-	-	-	0	-	0	-	0	-	-	-	0	-	-	-		-	0	+	
Stratégie4	-	-	-	+	+	-	+	-	+	-	-	-	+	+	-	-	+	0	-	0	-	-	0	
Stratégie5	0	0	0	-	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0	+	0	0	0	0

3. UN DEUXIÈME JEU PLUS CONTRÔLÉ

Ces considérations nous ont amenées à imaginer un deuxième jeu plus contrôlé, dans le but d'explorer en détail l'articulation de plusieurs voyelles et plusieurs consonnes bilabiales dans le bruit.

A partir d'un logatome CVCV choisi pour être assez neutre au niveau articulatoire : [lala], nous avons fait varier la 1^{ère} et la 2^{ème} voyelle pour explorer l'effet de l'ouverture de la mâchoire ([lela], [lila], [lale], [lali]), de l'arrondissement des lèvres ([lyla], [lula], [laly], [lalu]) et de la nasalisation ([lāla], [lalā]). De même nous avons fait varier la 1^{ère} et la 2^{ème} consonne pour explorer 3 types d'occlusives bilabiales : voisée ([bala], [laba]), sourde ([pala], [lapa]) et nasale ([mala], [lama]). Ces 17 logatomes n'appartiennent pas tous au lexique français, mais satisfont aux caractéristiques phonotaxiques du français. De la même façon que la parole lue diffère de la parole spontanée, l'énonciation de mots n'ayant aucun sens pour le locuteur diffère de la parole naturelle [Ste04]. Cependant, le sens d'un mot n'est pas défini par son appartenance à un dictionnaire mais par son contexte d'utilisation et son consensus entre les interlocuteurs. C'est pourquoi nous avons contourné ce problème en présentant ces logatomes aux locuteurs comme des noms propres de rivières dans le contexte de ce jeu. Pour éviter toute hésitation ou éventuel effet d'apprentissage de ces nouveaux mots, nous avons demandé aux locuteurs de se familiariser avec ces noms avant le jeu en les lisant plusieurs fois à voix haute.

Le jeu implique cette fois un seul locuteur qui doit communiquer des instructions à l'expérimentateur. Les 17 rivières sont représentées et disposées sur un tableau. Le but du jeu est de construire librement un chemin de flèches pour les relier. Le locuteur doit exprimer ses consignes à l'expérimentateur pour qu'il trace les flèches au tableau. Plusieurs contraintes doivent être respectées : (1) aucune rivière ne doit être laissée de côté, (2) une rivière ne peut pas participer au chemin par plus d'une flèche entrante et d'une flèche sortante, (3) le chemin doit être fermé. (4) Pour décrire chaque flèche du chemin, la structure des phrases est imposée : le locuteur doit suivre la forme : «La Riviere_n longe la Riviere_m» en respectant le sens de la flèche qui les joint (de Riviere_n vers Riviere_m). Ce protocole permet de disposer des mots-cibles en position initiale et finale des phrases, avec un contrôle du contexte segmental de la première consonne et de la dernière voyelle des mots-cibles. Le discours est toujours libre, non prévisible, et le locuteur a la nécessité d'être intelligible.

Les données acoustiques et articulatoires de ce corpus sont actuellement en cours d'analyse.

CONCLUSION

Nous avons présenté dans cet article les arguments théoriques en faveur de l'utilisation de tâches semi-spontanées pour l'étude de la parole Lombard. L'adaptation de la parole dans le bruit étant motivée par la recherche d'intelligibilité, le protocole doit donc réaliser un compromis entre le contrôle des énoncés et leur spontanéité. C'est ce que proposent les 2 jeux interactifs

présentés dans cet article, correspondant à deux niveaux de compromis adaptés à des niveaux de détail différents.

BIBLIOGRAPHIE

- [Aub05] Aubergé V., Rillard A., Audibert N. (2005), "De E-Wiz à E-Clone : méthodologie expérimentale pour la modélisation des émotions et affects authentiques", *Workshop sur les Agents Conversationnels Animés, Grenoble*.
- [Aye94] Ayers G.M. (1994), "Discourse functions of pitch range in spontaneous and read speech". *Working papers in linguistics, Columbus, Ohio*.
- [Bro83] Brown, G., A. Anderson, G. Yule, R. Shillcock. (1983). "Teaching Talk", *Cambridge University Press, UK*.
- [Cla86] Clark H.H., Wilkes-Gibbs D. (1986), "Referring as a collaborative process". *Cognition* 22(1), pp. 1-39.
- [Dre57] Dreher J.J., O'Neill J. (1957), "Effects of Ambient Noise on Speaker Intelligibility for Words and Phrases", *J. Acoust. Soc. Am.* 29, pp. 1320-23.
- [Gar06] Garnier M., Henrich N., Dubois D., Polack J.D. (2006), "Peut-on considérer l'effet Lombard comme un phénomène linéaire en fonction du niveau de bruit ? ", *Actes du 8ème CFA, Tours*.
- [Har69] "Harvard sentences. Appendix of: IEEE subcommittee on subjective measurements". (1969), *IEEE transactions on Audio and Electroacoustics* 17, pp.227-246.
- [Hir52] Hirsh I.J., Davis H., Silverman S.R., Reynolds E.G. Eldert E., Benson R.W. (1952), "Development of materials for speech audiometry", *J Speech Hear Disord.* 17(3). Pp. 321-37.
- [Jun93] Junqua J. (1993), "The lombard reflex and its role on human listener and automatic speech recognizers", *J. Acoust. Soc. Am.* 93 (1), pp. 510-524.
- [Kim05] Kim S. (2005), "Durational Characteristics of Korean Lombard Speech", *Proc. Interspeech, Lisbonne*.
- [Lev83] Levelt W., Cutler A. (1983), "Prosodic marking in speech repair", *Journal of Semantics*.
- [Lie85] Lieberman P., Katz W., Jongman A., Zimmerman R., Miller M. (1985), "Measures of the sentence intonation of read and spontaneous speech in American English", *J. Acoust. Soc. Am.* 77, pp. 649-657.
- [Mil55] Miller G.A., Nicely P.E. (1955), "An Analysis of Perceptual Confusions Among Some English Consonants", *J. Acoust. Soc. Am.* 27, 338-352.
- [Pen] Peng S.H., Beckman M.E. "Annotation conventions and corpus design in the investigation of spontaneous speech prosody in Taiwanese". <http://www.ling.ohio-state.edu/>
- [Por59] Portmann M., Portmann C. (1959), *Précis d'audiométrie clinique, Masson*.
- [Sch01] Schafer A.J., Speer S.R., Warren P., White S.D. (2001), "Prosodic influences on the production and comprehension of syntactic ambiguity in a game-based conversation task", *Fourteenth Annual CUNY Conference on Human Sentence Processing, Philadelphia*.
- [Ste04] Stephenson L.S. (2004), "An electropalatographic and acoustic analysis of frequency effects in the lexicon", *ph.D. thesis, Mcquarie university, Sydney, Australia*.
- [Sum88] Van Summers W., Pisoni D.B., Bernacki R.H., Pedlow R.I., Stokes M.A. (1988), "Effects of noise on speech production: Acoustic and perceptual analyses", *J. Acoust. Soc. Am.* 84, pp. 917-928.
- [Zei94] Zeiliger J., Serignat J.F., Autresser D., Meunier C. (1994), "BD Bruit, une base de données de parole de locuteurs soumis à du bruit". *Actes des Xèmes JEP*, pp. 287-290.

