Intelligibilité de la parole à plusieurs distances dans un bruit naturel

Julien Meyer, Laure Dentel, Fanny Meunier

10ème Congrès Français d'Acoustique
Lyon, 12-16 Avril 2010

Intelligibilité de la parole à plusieurs distances dans un bruit naturel

Julien Meyer[1], Laure Dentel[2], Fanny Meunier[3]

[1]Area de Linguistica, CCH, Museu Goeldi, Campus de Pesquisa, Av. Perimetral, 1901, Terra Firme, 66077-530, Belem, Brasil
[2]Groupe de recherche en COmmunication Sonore et PErception auditive de l'environnement (SCOPE), Boulevard Raspail, 75006 Paris, France
[3]Dynamique Du Langage, Institut des Sciences de l'Homme, 14 Avenue Berthelot, 69363 Lyon Cedex 07, France

jmeyer@museu-goeldi.br

Distant listening of speech is a common task performed daily by all human beings. In such an auditory situation, the speech signal is not only affected by the ambient noise but is also degraded during its in-air propagation between the emitter and the receptor. The present study takes a new path in speech perception in noise as it analyzes the phenomenon of intelligibility loss with increasing listener-to-speaker distance by considering the combined effect of amplitude attenuation and of a stable outdoor rural ambient noise (without wind or perceptible mechanical noise). Reference measurements and recordings were first performed during a Pilot Study in a milieu characterized by very low reverberation indexes and low acoustic pollution. These outdoor conditions of experimentation were used to build a realistic model for the design of a Laboratory Experiment aimed at observing the recognition performances of monosyllabic words for subjects with normal hearing thresholds. We simulated the outdoor amplitude attenuation due to propagation of the speech signal and masked it with the recorded interfering ambient noise. Several simulated distances of listening could be tested for each of the 36 French participants as different lists of 17 French words were played between the virtual distances of 11 to 33 m from the source (with a step of 2 meters). The played stimuli and the answers of the participants were analyzed at several levels: correct answers on words, intelligibility function, Speech Recognition Threshold (SRT) and recognition of the syllable structure. They were also observed at the level of phonemes: distribution of vowel recognition, or detailed consonant confusions and similarity judgments in function of both the distance and dB SNR levels.

1 Introduction

Ce travail s'interesse à l'intelligibilité de la parole à distances variables. Notre étude analyse la perte d'intelligibilité de la parole lors de l'augmentation de la distance entre deux interlocuteurs, dans un milieu ouvert avec un bruit de fond naturel, stable et calme. Une expérience de laboratoire a été menée sur la base d'une étude pilote de terrain. Une expérience de transmission acoustique en milieu extérieur a été conçue pour mesurer l'intelligibilité de mots français puis reproduite en laboratoire par souci d'un contrôle optimum des conditions de passage des 36 participants. Pour ce faire, nous avons simulé la distance par l'atténuation d'amplitude qu'elle engendre et partiellement masqué le signal de parole avec un bruit de fond environnemental enregistré sur le terrain.
Les études précédentes sur la perception de la parole qui ont testé l'écoute à distance ont concerné en grande majorité des conditions en milieu intérieur. Elles ont décrit les relations entre l'intelligibilité de la parole et de nombreuses variables acoustiques, comme les niveaux d'amplitude, les types de bruit masquant, les niveaux de bruit, et les conditions de réverbération de salles comme des salles de classes [1], des halls ou auditorium [2], ou même des tunnels [3]. De plus, l'étude de la reconnaissance de la parole à distance en intérieur a aussi été développée dans le cadre d'interfaces homme-machine en intégrant tout un ensemble de techniques de prise de son et d'algorithmes de traitement de la parole [4]..

En ce qui concerne l'influence des environnements naturels en extérieur sur la parole, la recherche s'est concentrée sur trois domaines : d'une part sur l'habilité des individus à ajuster tacitement les productions vocales pour compenser les pertes d'intensité dues à la propagation du son à distance [5]; d'autre part sur l'habilité humaine à estimer la distance du locuteur [6]. Enfin, sur le phénomène naturel d'adaptation acoustique du signal de parole en voix criée, parole sifflée ou même parole tambourinée pour accroître la distance de communication [7]. La présente étude étant issue d'une réflexion sur ce dernier thème, central dans le travail de recherche des deux premiers auteurs.

Dans la première partie, nous présentons notre méthodologie en soulignant que notre étude s'appuie sur un modèle expérimental en extérieur pour lequel la propagation du signal est simple (terrain plat, météo calme, effet de sol ou autres types de réverbération négligeables).
C'est le type de bruit de fond choisit et le masquage progressif avec la distance qui font l'originalité de notre approche. L'analyse de la reconnaissance et de la confusion des phonèmes, y est comparable aux nombreuses études sur la parole dans le bruit [8].

2 Méthodologie

2.1 Participants

Les 36 participants de l'expérience en laboratoire étaient des locuteurs natifs du francais, âgés de 18 à 30 ans; avec une audition normale testée par audiogramme. Quatre locuteurs natifs du francais, également avec une audition

normale testée par audiogramme, ont participé aux tests préliminaires réalisés lors de l'étude pilote en extérieur.

2.2 Stimuli

Au total 19 listes ont été construites et enregistrées dans un caisson insonorisé par un locuteur masculin entrainé à cette tâche et membre du Laboratoire DDL (avec un niveau moyen d'émission des mots de 77 dB(A)). Les listes de mots ont été composées de 17 mots qui ont été choisis en vertu de leur qualité de noms communs francais du vocabulaire courant, principalement des mots monosyllabiques, et quelques mots - moins de 5% - de structure syllabique CVV ou VVC. Toutes listes confondues, tous participants confondus et pour toutes les distances simulées, les proportions de chaque structure ont été les suivantes: 82,1% pour les CVC, 12,7% pour les CCV, 4,1% pour les CVV, 0,8% pour les VVC, et 0,3% pour les VCC.

De plus, toutes les listes ont été équilibrées en terme de:

- Fréquence d'occurrence de mot dans la langue francaise avec une moyenne par liste entre 3,79 et 3,91 d'après la méthode d'évaluation de New et al [9].
- Nombre de voisins phonologiques pour chaque mot, avec une moyenne comprise entre 19,59 et 20,1 par liste.
- Nombre de lettres par mot. La moyenne par liste étant comprise entre 4,5 et 4,6 lettres.
- Durée de prononciation de chaque mot, avec une moyenne par liste entre 547 et 553,4 ms.
- Alternance voyelle-consonne, avec environ le même nombre moyen d'alternances CVC.
- Genre des noms, avec le même nombre de noms masculins et féminins dans chaque liste.

Toutes les pistes audio originales ont été calibrées avec la même root mean square energy level. A partir de ces pistes audio utilisées lors de l'étude pilote en extérieur, nous avons construit de nouveaux fichiers audios simulant à la fois l'atténuation en amplitude due à la distance (cf. 2.3.2) et l'effet masquant du bruit de fond (cf. 2.3.3 et 2.3.4.). Ces derniers fichiers audios ont servi à l'expérience en laboratoire.

2.3 Conception et déroulement

Nous avons d'abord mis au point le protocole d'expérimentation en milieu extérieur lors d'une étude pilote de terrain puis concus sur ce modèle une expérience de laboratoire.

L'objectif de l'étude pilote était de tester la faisabilité du test dans des conditions réalistes en extérieur. Ce faisant, il s'agissait aussi de définir les distances de début et de fin d'expérimentation ainsi que le pas de distance entre deux mesures successives. Il s'agissait également de définir la durée du test en laboratoire – que nous voulions inférieure à une heure- et de préparer ce test par l'enregistrement du bruit de fond et par des mesures de référence sur l'amplitude.

La conception d'une expérience en laboratoire comportait deux avantages essentiels: elle permettait de placer tous les participants dans des conditions rigoureusements égales, en particulier par rapport aux niveaux de bruit de fond. De plus, elle permettait de calculer de manière très précise les rapports signal sur bruit qui arrivent aux oreilles du récepteur, le bruit étant ajouté après la simulation d'atténuation d'amplitude. Cette section décrit des éléments clefs de la conception de ces deux étapes.

2.3.1 Etude pilote et enregistrements

Le terrain d'expérimentation de l'étude pilote était un milieu ouvert situé dans les champs aux alentours de la ville de Vilanova i la Geltru, en Catalogne espagnole. Ce milieu était caractérisé par une faible pollution acoustique et de très faibles indices de réverbération. Il s'agissait d'un chemin plat entouré de champs secs non cultivés et couverts d'une végétation estivale basse et éparse. Toutes l'expérience pilote a été réalisée dans des conditions météorologiques et sonores calmes et stables, sans aucun bruit de fond mécanique audible. La session a été interrompue lorsqu'un bruit animal ou mécanique imprévu a perturbé nos mesures (avions, tracteurs, motos, oiseaux ou insectes). Mais un des critères du choix du terrain d'expérimentation était précisément le calme par rapport à ce type de perturbations acoustiques. Toutes les enregistrements et les tests ont été faits le même jour d'août 2007 lors d'une seule session (mesures réalisées sur une station météo portable (Geos skywatch): vitesse du vent < 1 m/s, degré d'humidité entre 57% et 65% et température entre 26ºC et 28ºC. L'ensemble de ces précautions a permis d'enregistrer le bruit de fond interférent utilisé lors de l'expérience en laboratoire et de prendre des mesures pour le calibrer.

De plus, pour tester la faisabilité de l'expérience d'intelligibilité, nous avons choisit de diffuser des stimuli à leur niveau original de production avec un haut parleur à haut rendement entre 200 Hz et 10 kHz (TVM Medium ARM190-00/8) situé à 1 mètre au dessus du sol pour simuler une personne émetrice assise sur une chaise. Les stimulis étaient joués avec un ordinateur portable Fujitsu Siemens connecté à un amplificateur (amplificateur Magnat Xcite 301) relié aux hauts parleurs. Le niveau d'émission a été calibré à l'aide d'un signal sinusoidal de référence de 1 kHz (mesuré à 1 mètre de la source avec un sonomètre BK 2240). Plusieurs distances d'écoute ont ont pu être testées pour chacun des 4 participants de l'étude pilote. Nous leur avons joué différentes listes de mots tous les deux mètres entre 11 et 33 mètres de distance réèlle source-récepteur.

2.3.2 Simulation de la distance

Pour l'expérience en laboratoire, nous avons choisit de simuler la propagation extérieure en reproduisant l'atténuation d'amplitude théorique due à la propagation sphérique du signal (en appliquant la loi en carré inverse). L'effet de sol et les effets atmosphériques n'ont pas été simulés car nous avons observé suite à l'étude pilote qu'ils ne jouaient pas un rôle primordial pour les objectifs de notre analyse sur le terrain d'expérimentation choisit.

2.3.3 Bruit de fond et rapport signal sur bruit

Le choix du bruit de fond et le réglage de son niveau par rapport au signal de parole sont les principaux facteurs affectant les performances d'intelligibilité. La mesure et le contrôle de ces deux aspects sont donc essentiels pour notre étude

Choix du bruit de fond :

Le bruit choisit pour l'expérience de laboratoire est représentatif du type de bruit de fond observé en milieu rural isolé, de jour et par temps calme. Les précautions d'enregistrement détaillées en 2.3.1 ont permit de capter un bruit de fond relativement stable (un écart-type de 1,2

dB(A)) dans ce milieu calme (moyenne de 49,3 dB(A)). D'une manière générale, il est notable que les caractéristiques des bruits de fond environnementaux que l'on peut observer en milieu rural isolés sont extrêmement variables. Ils dépendent de la situation géographique, du terrain, de la végétation, des circonstances météorologiques, du microclimat que le signal doit traverser et des bruits biologiques ou hydroliques comme des cris d'animaux, l'écoulement des rivières ou le flux et reflux de la mer. Cependant, le type de bruit que notre étude propose d'utiliser est typique de la base sur laquelle ces perturbations naturelles viennent s'ajouter. C'est aussi sur ce fond environnemental que s'ajoutent toutes sortes de 'bruits modernes' comme ceux produits par les véhicules à propulsion mécanique, et qui constituent la pollution sonore de notre époque industrielle.

Acoustiquement, un tel bruit naturel est caractérisé par de hauts niveaux d'énergie dans les basses fréquences du spectre de la voix (en particulier en dessous de 250 Hz), des niveaux d'énergie intermédiaires de 250 à 2000 Hz et des niveaux plus bas au dessus de 2 kHz (figure 1). Il est donc très différent des bruits artificiels comme les bruits 'speech shaped' ou les bruits blancs utilisés dans la plupart des expériences contemporaines de reconnaissance de la parole dans le bruit qui s'intéressent surtout à des conditions de communications rencontrées à l'intérieur de batiments ou en ville.

L'interférence de ce type de bruit avec la parole humaine n'a pas été étudiée de manière approfondie malgré son intérêt pour comprendre les conditions naturelles dans lesquelles le langage humain s'est développé au cours de son histoire.

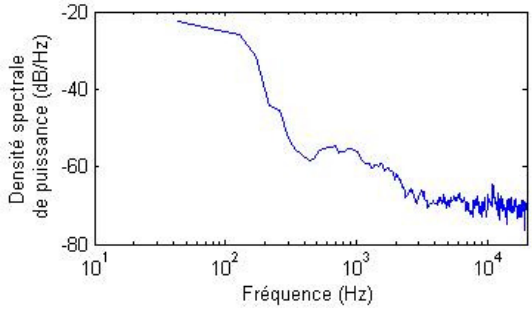

Figure 1: Spectre long-terme du bruit interférent.

Rapport Signal sur Bruit (RSB):

Pour analyser l'influence des niveaux de dB RSB sur la reconnaissance des mots, nous avons d'abord mesuré les niveaux de puissance sonore de chaque mot à chaque distance et déduit de ceux ci les niveaux de puissance sonore du spectre fréquentiel à long terme du bruit. De là, nous avons calculé les valeurs moyennes de dB RSB pour tous les mots joués à chaque distance (et leurs écart-types σ ; cf partie gauche Tableau 1). D'une part, nous avons alors observé que la distribution de ces valeurs RSB n'est pas linéaire avec la distance, ce qui est un effet de la loi d'atténuation de l'amplitude avec la distance. D'autre part, que les classes de catégories de distance se chevauchent suivant les valeurs dB RSB, avec un écart type de 1,88 dB RSB pour chaque distance. Ce second aspect reflète la variabilité des niveaux d'énergie sonore (sound power levels) inhérente à la production de parole du locuteur.

Afin d'évaluer plus précisément la relation entre l'intelligibilité de la parole et les valeurs de dB RSB, nous avons crée une autre classification des valeurs de RSB des mots. Nous avons créé 12 catégories de mots avec des valeurs de RSB proches, toutes listes confondues, toutes distances confondues et tous participants confondus (partie droite du Tableau 1). Ces classes ne se chevauchaient pas, ce qui permet plus d'analyses statistiques, en particulier le calcul du SRT (Speech Recognition Threshold).

Classes par distance	RSB (dB)	σ		Classes par RSB proches	RSB (dB)	σ
11 m	-8,7	1,88		1	-5,3	0,5
13 m	-10,2			2	-7,2	0,5
15 m	-11,4			3	-8,9	0,53
17 m	-12,5			4	-10,7	0,53
19 m	-13,5			5	-12,5	0,52
21 m	-14,3			6	-14,3	0,51
23 m	-15,1			7	-16,1	0,52
25 m	-15,8			8	-17,8	0,52
27 m	-16,5			9	-19,6	0,51
29 m	-17,1			10	-21,4	0,49
31 m	-17,7			11	-23,2	0,48
33 m	-18,2			12	-24,9	0.71

Tableau 1: Niveaux de RSB dB correspondant à deux différentes manières de grouper les mots: par distance d'écoute (gauche) ou par classes de mots ayant des rapports signal sur bruit proches (droite).

2.3.4 Séquences des expériences

Chaque participant, assis devant un ordinateur dans la salle expérimentale du Laboratoire Dynamique du Langage (CNRS, University de Lyon 2), avait pour tâche de suivre les instructions de l'expérience contrôlée depuis une interface dédiée. Tous les ordinateurs et les casques audios (Beyerdynamic DT48) utilisés étaient identiques avec des cartes sons identiques. Ce matériel a été calibré afin de reproduire les niveaux de production de référence. Pour chaque participant, le test commencait par une courte phase d'entrainement de 5 mots assurant la compréhension de la tâche. Puis, une première liste était jouée correspondant à une distance de 11 mètres depuis la source. Le participant avait la simple tâche de d'écouter chaque séquence sonore et d'essayer de comprendre le mot cible. Il était informé du fait que les sons joués étaient des mots. Il avait la permission de ne pas répondre s'il estimait n'avoir rien entendu ou s'il n'avait rien identifié dans les séquences entendues. Dans le cas de l'expérience pilote en extérieur, les participants répétaient les mots ou les sons entendus juste après leur écoute. Dans le cas de l'expérience de laboratoire les participants les tapaient au clavier de l'ordinateur dans un formulaire de l'interface et les validaient pour passer au mot suivant. Les participants ne recevaient aucune information sur leurs performances avant la fin du test. Lors de l'expérience pilote en extérieur, les listes et les réponses étaient enregistrées simultanément : au niveau du participant et à un mètre au dessus du sol (enregistreur: M-audio Audiotrack; microphone stereo AT 822). En laboratoire et en extérieur, ces processus étaient répétés tous les deux mètres jusqu'à 33 mètres, avec des listes distinctes pour chaque distance et chaque locuteur.

3 Résultats

Un programme spécifique fonctionnant sous Matlab a été développé pour analyser les réponses des sujets, les synthétiser et les représenter graphiquement en fonction de plusieurs paramètres, comme par exemple la composition phonétique des mots testés ou le RSB des mots perçus.

3.1 Performances générales de reconnaissance des mots

Les performances générales de reconnaissance des mots ont atteint 54,6% de réponses correctes pour tous les participants et toutes les distances. Ces performances baissent en moyenne avec la distance de 77,8 % à 11 mètres jusqu'à 35,9 % de réponses correctes à 33 mètres, avec une variabilité inter-individuelle importante a chaque distance (Figure 2). Il y a une correlation quasi linéaire très forte entre la distance et la perte d'intelligibilité (R^2=0,95). De plus, afin de mesurer la fonction d'intelligibilité variant en fonction des RSB sur les mots, nous avons utilisé la classification en 12 categories de mots avec des RSB proches (Tableau 1). Grâce à ces catégories nous avons pu déduire les valeurs de SRT (Speech Recognition Threshold) et la pente correspondante de la fonction de reconnaissance, pour toutes les listes et tous les participants. La valeur du seuil SRT est de -15,24 dB RSB. La pente est de 3,76 %/dB. Pour garantir leur validité audiologiques, ces valeurs ont été calculées à partir de plus de 450 mots dans chaque catégorie de niveau de RSB qui encadrent le point correspondant à 50% de réponses correctes [10].

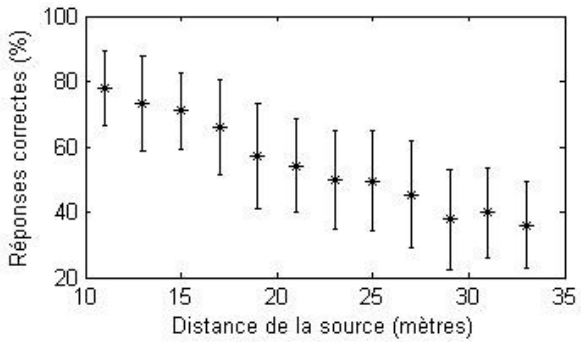

Figure 2: Performances de reconnaissance des mots pour 36 participants à 12 distances (moyenne de réponses correctes avec écarts types sur les participants)

3.2 Reconnaissance de la structure syllabique des mots, insertion et suppression de phonèmes

Les performances générales de reconnaissance de la structure syllabique des mots (CVC, CCV, CVV, VVC, or VCC) atteint 76,5%, toutes distances confondues. De plus, parmi les 2 types de structures les plus fréquentes qui représentent 94,8% des mots, CVC a été reconnues dans 80,2% et CCV dans 55,7%. L'évolution de la degradation des performances de reconnaissance de ces deux types de structures avec la distance confirme que seul CVC est une structure significativement resistante (figure 3). En outre, il apparaît que les erreurs sur ces structures avaient trois type d'origines principales : soit l'absence de réponse sur le mot, soit la suppression d'un phonème, soit l'insertion d'un phonème. L'absence de réponse concernait 20,6% de toutes les erreurs sur toutes les structure, 23,6% des erreurs sur les mots CVC, et 12,9% des erreurs sur les mots CCV. De plus, les erreurs avec des insertions de phonèmes atteignaient 36,8% de toutes les structures mal reconnues et respectivement 34,7% pour les CVC et 37,9% pour les CCV. Les erreurs avec les suppressions ont atteint 73,6% de toutes les structures non reconnues (resp. 73,3% et 78,2% pour CVC et CCV). Nous avons aussi mesuré la progression des erreurs avec la distance. Par exemple, les erreurs implicant des suppressions augmentent quasi linéairement de 11m (8,7%) à 33m (27,5%).

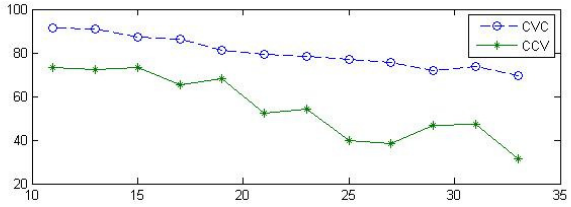

Figure 3: Reconnaissance des mots en CCV ou CVC (% de réponses justes en fonction de la distance)

3.3 Reconnaissance et confusion des phonèmes

3.3.1 Voyelles

Si l'on prend en compte toutes les voyelles à toutes les distances, les performances générales de reconnaissance des voyelles a été très élevé (91,5%). Parmi les 8,5% d'erreurs, une grande majorité est due à l'absence de réponse (85%) et seulement quelques unes confusion avec une autre voyelle (15%). De plus, il y a une grande variabilité de reconnaissance entre les voyelles (figures 4 et 5). Les voyelles /a, ɑ̃, ɛ/ ont été les mieux reconnues avec respectivement 97%, 96% and 96% de réponses correctes. SUivies des voyelles /ɶ̃, y, ɔ, i/ qui ont été reconnues à des performances supérieures à 90%, avec les meilleures performances pour /ɶ̃/ (93%). D'autre part, les voyelles /õ, u, œ/ ont un taux d'identification compris entre 88 et 89%. Finalement, /e/, /o/ et /ø/ ont montré des performances de reconnaissances avec respectivement 79%, 77% et 72,3% de réponses justes. Les voyelles /ø/ et /œ/ n'avaient pas assez de cas à chaque distance pour apparaitre dans des analyses plus poussées. De plus, parmi toutes les possibilités de confusion entre voyelles, aucune n'a été significativement plus fréquente qu'une autre.

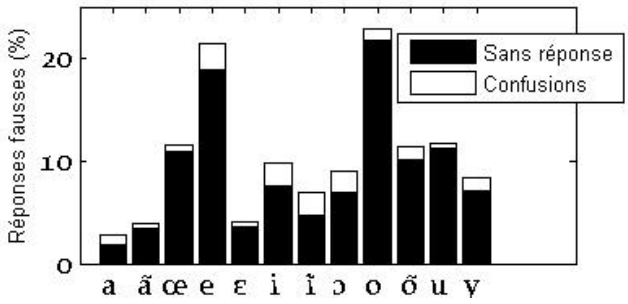

Figure 4: Distribution des erreurs pour chaque voyelle

La distribution des performances de reconnaissance des voyelles a aussi été mesurée en fonction des distances pour chaque participant (Figure 4). Ces performances sont restées supérieures à 90% jusqu'à 23 mètres et supérieures à 80% à n'importe qu'elle distance jusqu'à 31 mètres. Au fur et à mesure que la distance augmente, on observe en moyenne une plus grande variabilité des taux de reconnaissance de voyelles aux distances supérieures à 23 mètres (comme le montrent les valeurs d'écart type sur la Figure 5). Cet effet est en partie dû au fait que /ɔ/ et /i/ ont été significativement moins bien reconnues que les autres à des longues distances alors que leur comportements était cohérent avec le reste jusqu'à 23 mètres.

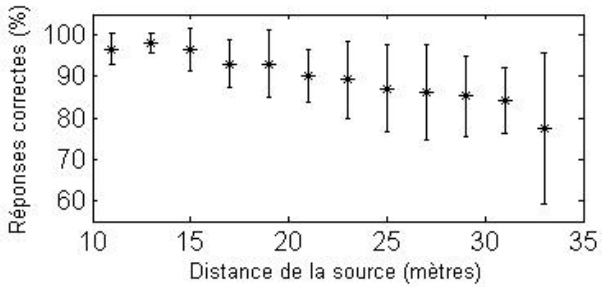

Figure 5: Performances de reconnaissance des voyelles, toutes voyelles confondues sauf /œ/ et /ø/; écart-types sur les participants.

3.3.2 Consonnes

Résultats généraux

Les performances de reconnaissance des consonnes ont été en moyenne de 70,3%, si l'on considère toutes les distances et toutes les consonnes. Nous avons observé une diminution des valeurs moyennes de réponses correctes de 88,1 % à 11 mètres jusqu'à 55,5% à 33 mètes (correspondant respectivement à des valeurs dB RSB de -8,7 et -18,7 (Tableau 1)). Nous avons aussi mesuré une forte variabilité entre consonnes. Parmi les consonnes les mieux reconnues il y a d'une part trois coronales fricatives: la fricative alvéolaire [s] (92,5% de réponses justes) et les fricatives post-alvéolaires [ʃ] et [ʒ] (95,9 % et 84,9% de réponses correctes). D'autre part, il y a deux semi-consonnes approximantes: la palatale approximante [j] (80 % de réponses correctes); et la post alvéolaire approximante [w], qui n'est présente que sous les formes [wa] ou [wɛ̃] précédées par une autre consonne (en contexte CCV), explaining its very high level of recognition (98,2%). Les consonnes liquides ont des taux de reconnaissance autour de 80%, avec [l] (81,4%) et [r] (76,9%). Parmi les nasales, [m] (73%) est la mieux identifiée, suivie de [n] (69,5%), et finalement de [ɲ] (64,6%) (en outre, [ɲ] n'est pas présente dans suffisament de cas pour figurer dans des analyses statistiques plus poussées). [t], [g] et [k] ont des taux de reconnaissance proches de 67%; alors que [b] et [z] étaient un peu moins bien reconnus (autour de 63%). Finalement, les performances de reconnaissance les plus basses étaient pour [f] (31%), suivies de [d] (49,4%), de [v] (53,2%) et de [p] (54,3%) (figure 6).

En outre, parmi les 29,7 % d'erreurs sur toutes les consonnes, 58,8% étaient dues à des confusions avec d'autres consonnes et 41,2% à des absence de réponse. Seules les liquides et les approximantes ont vu ces proportions inversée avec plus d'absence de réponse que de confusions. De plus, l'évolution générale des erreurs avec l'augmentation de la distance a montré que les confusions restent toujours plus fréquentes que l'absence de réponse, même si cette tendance est plus marquée à partir de 23 m (figure 7).

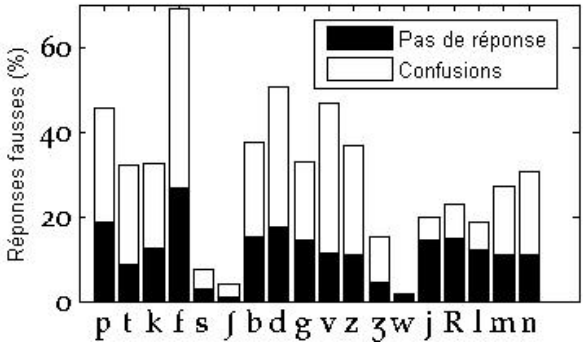

Figure 6: Distribution des erreurs pour chacune des consonnes

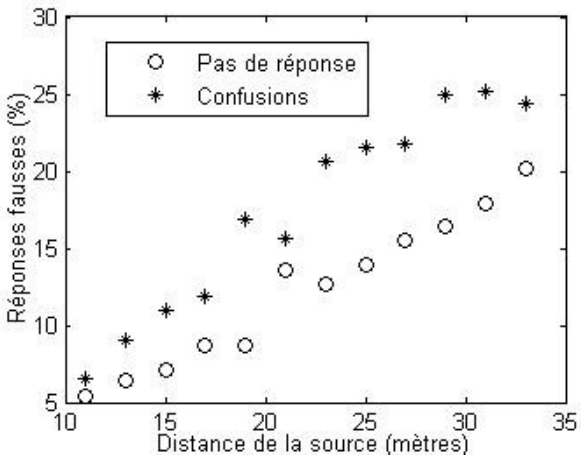

Figure 7: Proportion des absences de réponse vs. les confusions pour toutes les consonnes en fonction de la distance

Effet de la distance et du RSB sur la reconnaissance des consonnes

Pour comprendre l'évolution des taux de reconnaissance avec la distance, nous avons sélectionné les 17 consonnes les plus jouées [p, t, k, f, s, ʃ, b, d, g, v, z, ʒ, r, l, j, m, n] et calculé le pourcentage de réponse correcte de chacune d'entre elles à chacune des 12 distances du test (figure 8).

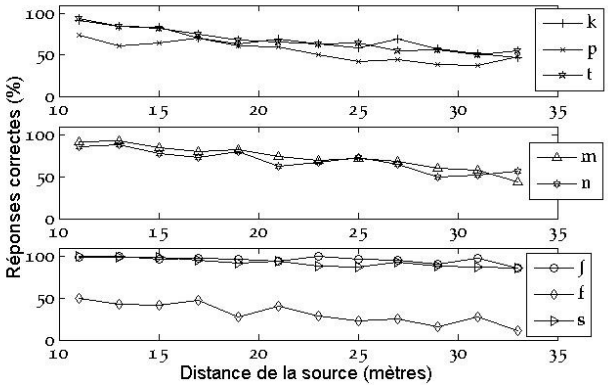

Figure 8 (partie 1): Distribution des taux de réponses justes en fonction de la distance pour chacune des 11 consonnes les plus fréquemment jouées.

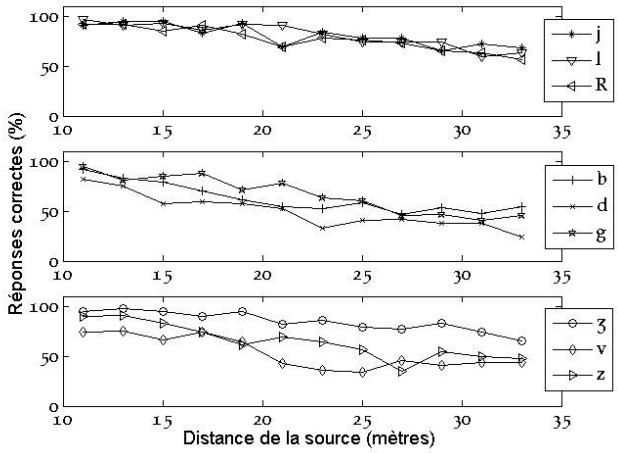

Figure 8 (partie 2): Distribution des taux de réponses justes en fonction de la distance pour chacune des 11 consonnes les plus fréquemment jouées.

Nous observons que [ʃ] et [s] sont les mieux reconnues à toutes les distances. Ce sont les seules consonnes à n'être presque pas affectées par l'atténuation et le bruit ambiant, vraisemblablement en raison de leur bande de fréquence étroite dans les fréquences élevées [11]. Les autres consonnes présentent des profils de perte d'intelligibilité avec la distance, et une chute des performances entre 20 et 40% suivant les consonnes. La mieux reconnue de ces dernières reste la fricative [ʒ] avec la chute de performance la moins élevée, puis vient l'approximante [j] et les liquides. Ces consonnes sont encore très bien reconnues jusqu'à 19 mètres. Ensuite, les consonnes [z, g, m, n, k, t] ont des profils similaires : elles sont encore très bien reconnues jusqu'à 15 mètres et terminent à 33 mètres à des taux voisins de 50% d'identification.

Confusions entre consonnes

Nous avons également mesuré les confusions entre consonnes parmi les 17 les plus jouées [p, t, k, f, s, ʃ, b, d, g, v, z, ʒ, r, l, j, m, n]. Le confugramme (figure 9) présente une vue générale de ces résultats. Dans la grande majorité des cas, les confusions impliquent des consonnes qui partagent au moins un trait phonétique lié au lieu d'articulation (labiales, coronales, dorsales) ou à la manière d'articulation (fricative, nasale, plosive, approximante). Les confusions liées au lieu d'articulation des consonnes concernent la majorité des confusions. Une analyse plus poussée de cet aspect sera réalisé dans une approche ultérieure de ces données.

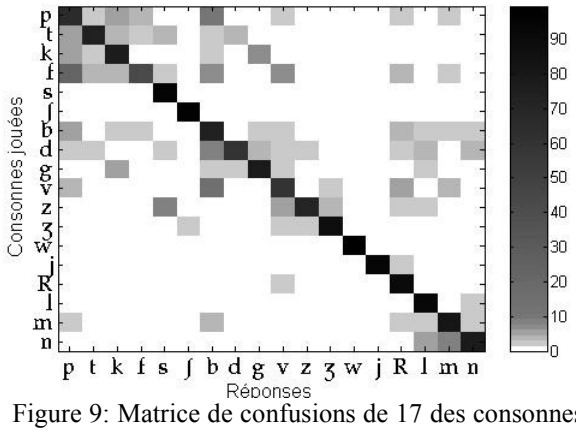

Figure 9: Matrice de confusions de 17 des consonnes du francais (niveaux de gris en % des confusions).

4 Conclusion

Cette étude originale porte sur l'interférence d'un bruit environnemental avec la reconnaissance de la parole humaine à distances variables. Nous avons présenté des résultats nouveaux permettant d'obtenir une analyse de la fonction d'intelligibilité (valeurs de seuil SRT et de pente), de l'évolution de la reconnaissance de la structure syllabique des mots ou de leurs phonèmes. Une publication complémentaire en fournira une interprétation complète.

Remerciements

Nous remercions Vincent Monatte pour son aide lors du passage des participants en laboratoire. Julien Meyer a été financé par la Fondation Fyssen et par ELDP, SOAS, Université de Londres. Laure Dentel a travaillé dans le cadre du Groupe de recherche en communication sonore et perception auditive de l'environnement (association de recherche SCOPE). Fanny Meunier a été financée par le CNRS et Speech In Noise (SPIN, ERC).

Références

[1] Bradley J. "Speech Intelligibility in Classroom", *J. Acoust. Soc. Am*. 81 (3), 846-854 (1986).

[2] Houtgast T., Steeneken H.J. "A review of the MTF concept in room acoustics and its use of estimating speech intelligibility in auditora", *J. Acoust. Soc. Am*. 77, 1069-1077 (1985).

[3] Imaizumi H., Kunimatsu S., Isei T. "Sound propagation and speech transmission in a branching tunnel", *J. Acoust. Soc. Am*. 108 (2), 632-642 (2000).

[4] Woelfel M., McDonough J. *Distant Speech Recognition*, Wiley Eds (2005).

[5] Zahorik P., Brungart D.S., Bronkhorst A.W. "Auditory distance perception in humans: a summary of past and present research", *Acta Acustica*, 91(1), 409-420 (2005).

[6] Zahorik P., Kelly, J.W. "Accurate vocal compensation for sound intensity loss with increasing distance in natural environments", *J. Acoust. Soc. Am*. 122 (5), EL143-EL150 (2007).

[7] Meyer J. "Typology and acoustic strategies of whistled languages: phonetic comparison and perceptual cues of whistled vowels", *J. Inter. Phonetic Assoc.* 38, 69-94 (2008).

[8] Summers V.W., Pisoni D.B., Bernacki, R.H., Pedlow R.I., Strokes M.A. "Effect of noise on speech production: acoustical and perceptual analysis", *J. Acoust. Soc. Am*. 84, 917-928 (1988)

[9] New, B., Pallier, C., Ferrand, L., Matos, R. "Une base de données lexicales du français contemporain sur internet : LEXIQUE". *L'Année Psychologique*, 101, 447-462 (2001).

[10] Gelfand, S.A. *Hearing: An introduction to psychological and physiological acoustics.* New York: Marcel Dekker (1998).

[11] Calliope, *La parole et son traitement automatique*. Masson, Paris. (1989).