Comparaison de mesures perceptives et automatiques de l'intelligibilité

Application à de la parole simulant la presbyacousie

Lionel Fontan[*] — Cynthia Magnen[**] — Julien Tardieu[**] — Isabelle Ferrané[*] — Julien Pinquier[*] — Jérôme Farinas[*] — Pascal Gaillard[**] — Xavier Aumont[***]

[*] *Université de Toulouse ; UT3 ; Institut de Recherche en Informatique de Toulouse – {fontan, ferrane, pinquier, jfarinas}@irit.fr*

[**] *Université de Toulouse ; UT2J ; OCTOGONE ; EA4156 – {magnen, tardieu}@univ-tlse2.fr*

[***] *Société Archean Technologies ; Montauban, France – xaumont@archean.fr*

RÉSUMÉ. *Cet article présente une étude comparative entre mesures perceptives et mesures automatiques de l'intelligibilité de la parole sur de la parole dégradée par une simulation de la presbyacousie. L'objectif est de répondre à la question : peut-on se rapprocher d'une mesure perceptive humaine en utilisant un système de reconnaissance automatique de la parole ? Pour ce faire, un corpus de parole dégradée a été spécifiquement constitué puis utilisé pour des tests perceptifs et enfin soumis à un traitement automatique. De fortes corrélations entre les performances humaines et les scores de reconnaissance automatique sont observées.*

ABSTRACT. *This study aims at comparing perceptive and automatic measures of speech intelligibility in the case of speech signals simulating the effects of age-related hearing loss (presbycusis). A new corpus especially designed for studying speech intelligibility and perception and the comparison of human speech recognition scores with Automatic Speech Recognition measures are presented. one of the long-term goals of this work being to provide automatic tools facilitating the tuning of hearing aids. By achieving strong correlations between human performances and ASR recognition scores, this work provides validation data for the use of an ASR system as a way to modelize speech intelligibility in the case of presbycusis.*

MOTS-CLÉS : *simulation de la presbyacousie, mesure de l'intelligibilité de la parole, reconnaissance automatique de la parole.*

KEYWORDS: *presbycusis simulation, speech intelligibility metric, automatic speech recognition.*

1. Introduction

La presbyacousie est une atteinte cochléaire bilatérale et relativement symétrique liée au vieillissement naturel du système auditif. Le phénomène survient chez des sujets âgés de plus de 50 ans et se caractérise par une diminution progressive et irréversible de l'acuité auditive.

Les effets néfastes de la presbyacousie sur la perception de la parole ont pour conséquence de réduire sensiblement l'aptitude des auditeurs à communiquer, en particulier dans des environnements bruités (Moore, 2007). Pour mesurer les effets spécifiques de la presbyacousie sur la réception de la parole, des tests vocaux sont utilisés en complément des audiogrammes tonaux : les tests d'intelligibilité de la parole – cf. l'ouvrage du Collège National d'Audioprothèse (2007) pour la description des tests couramment utilisés dans le domaine. Les épreuves d'intelligibilité consistent habituellement à faire écouter des listes de mots ou de phrases aux auditeurs et à leur demander de répéter ce qu'ils ont entendu (ANSI S3.5, 2007). Des scores de précision (pourcentages de reconnaissance) sont ensuite calculés en comparant les réponses des auditeurs avec les listes de stimuli cibles. Ces tests souffrent trois inconvénients majeurs :

1) ils sont fastidieux, car ils nécessitent des conditions de diffusion sonore suffisamment contrôlées, la présence d'un ou plusieurs jurys d'écoute pour valider ou invalider chaque réponse des auditeurs, et ils requièrent *in fine* le traitement (souvent manuel) des réponses pour la production de scores ;

2) le matériel linguistique utilisé se limite généralement à des listes de mots présentés en dehors de tout contexte. Plusieurs études récentes suggèrent que les scores d'intelligibilité ainsi obtenus ne sont que très peu corrélés avec les performances de sujets écoutant des phrases dans un contexte déterminé (Fontan *et al.*, accepté ; Hustad, 2008). Constitués de listes de mots (ou pseudo-mots), les tests d'intelligibilité présentent donc une validité externe limitée ;

3) enfin, pour le cas particulier de la presbyacousie, l'interprétation des résultats issus d'épreuves de réception de la parole peut être complexe dans la mesure où les sujets sont âgés et peuvent donc présenter des troubles cognitifs associés entravant également leur compréhension.

Pour pallier ces différents problèmes, le présent travail vise à créer un dispositif électronique et informatique permettant de fournir automatiquement des mesures allant de l'intelligibilité de mots à la compréhension de phrases, *via* l'utilisation d'un moteur de reconnaissance automatique de la parole[1]. Si des approches comparables ont été conduites concernant l'intelligibilité de locuteurs souffrant de troubles pathologiques de production de la parole (Maier *et al.*, 2009 ; Schuster *et al.*, 2006), il s'agit

1. Ce dispositif a fait l'objet du dépôt de brevet européen n°2136359 – *Procédé et appareil de mesure de l'intelligibilité d'un dispositif de diffusion sonore* (Aumont et Wilhem-Jaureguiberry, 2009).

à notre connaissance de la première étude portant sur la prédiction de l'intelligibilité dans le cadre d'une pathologie de l'audition.

Afin de mener à bien cet objectif, des mesures de référence ont été recueillies auprès de sujets ayant passé plusieurs tests d'intelligibilité et de compréhension de la parole (Fontan *et al.*, 2014). L'étude que nous présentons dans cet article constitue le premier volet de ce programme de recherche et concerne la prédiction des scores d'intelligibilité obtenus par les sujets dans une épreuve très classique du domaine de l'audioprothèse : la répétition de mots issus des listes de Fournier (1951) – cf. annexe A.

La figure 1 décrit les différentes phases de la constitution du corpus et des traitements effectués. Dans une première phase, un corpus d'enregistrements vocaux a été constitué à partir de la lecture, par trois locuteurs, des mots présentés en annexe A. Une simulation de la presbyacousie (décrite dans la section 2) a ensuite été appliquée sur ces enregistrements. Le protocole suivi pour la création de ce corpus est détaillé dans la section 3. Dans une deuxième phase, des mesures de référence ont été obtenues grâce à des tests d'intelligibilité (répétition des mots) réalisés auprès d'un groupe de 30 sujets sains : la section 4 décrit le protocole suivi pour la collecte des réponses des sujets. Dans une troisième phase, les mêmes stimuli ont été soumis à un système automatique de reconnaissance de la parole. Ce système ainsi que les adaptations qui ont été réalisées sont décrits dans la section 5. Enfin, une analyse comparative entre les résultats issus des mesures subjectives et ceux issus des mesures automatiques (section 6) est réalisée en vue d'étudier la possibilité de prédire les scores de référence (section 7).

2. Simulation des effets de la presbyacousie sur la perception de la parole

Du point de vue fonctionnel, la presbyacousie est une surdité de perception caractérisée par une perte des cellules sensorielles débutant à l'entrée de la cochlée, lésant dans un premier temps les cellules ciliées externes (CCE). Dès lors, la sensibilité et la sélectivité fréquentielles s'amenuisent, de même que la capacité à séparer le signal du bruit ambiant (Bouccara *et al.*, 2005). Du point de vue audiométrique, l'atteinte se traduit par une chute sélective dans les fréquences aiguës (les seuils étant relativement conservés jusqu'à 1 kHz), ce qui provoque des difficultés de compréhension de la parole, en particulier en milieu bruyant. Par ailleurs, le rôle de compression de l'intensité sonore que jouent les CCE est également largement atténué. En résulte une distorsion de la sensation d'intensité : les sujets deviennent progressivement intolérants aux sons forts produits dans leur entourage.

Afin de simuler les pertes auditives liées à la presbyacousie, deux grandes approches sont généralement utilisées (Ariöz, 2012). La première consiste à masquer la parole dans du bruit pour réduire l'audibilité du signal. Cette approche présente l'inconvénient majeur de ne pas être applicable pour la simulation d'atteintes sévères : le bruit de masquage devrait en ce cas être diffusé à des niveaux intolérables pour

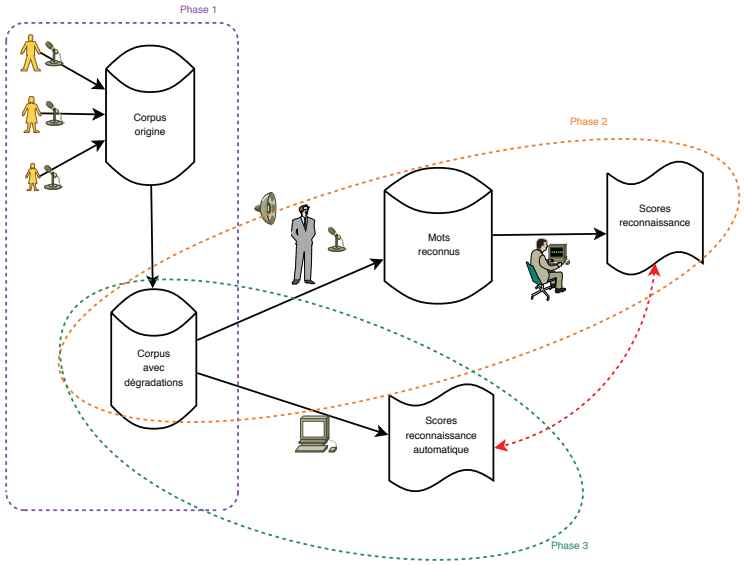

Figure 1. *Diagramme des différentes phases de production du corpus et des analyses effectuées*

l'oreille humaine (Moore, 2007). De plus, le masquage par le bruit ne semble pas conduire uniquement aux effets attendus sur l'audibilité du signal de parole ; il parait également provoquer chez les auditeurs des effets sur la perception du niveau sonore (Humes et Roberts, 1990). La seconde approche, qui correspond à celle utilisée dans cette étude, vise à effectuer des traitements informatiques sur le signal de parole pour simuler les effets produits par le système auditif déficient. Dans cette perspective les trois effets caractéristiques de la presbyacousie sont reproduits (Moore, 2007) :

– **l'augmentation des seuils d'audibilité**. L'intensité du signal de parole est plus ou moins réduite en fonction des bandes de fréquences : plus les fréquences sont élevées plus l'intensité du signal est abaissée ;

– **la réduction de la sélectivité fréquentielle**. Pour simuler cet effet, un lissage du spectre fréquentiel est conduit, en suivant généralement la procédure décrite par Baer et Moore (1993) ;

– **le recrutement de sonie**. Ce phénomène est engendré par une réduction de la dynamique d'intensité chez les sujets presbyacousiques ; en conséquence, ces derniers perçoivent les variations d'intensité de manière exagérée. Pour simuler cet effet, l'enveloppe du signal est multipliée par elle-même (Moore et Glasberg, 1993).

Les différents traitements des signaux de parole sont effectués de manière plus ou moins importante en fonction de la sévérité du cas de presbyacousie dont on cherche à simuler les effets.

3. Constitution d'un corpus simulant les effets de la presbyacousie

Pour cette étude nous avons choisi de recourir à des listes de mots communéme utilisées par les médecins ORL et les audioprothésistes lors de tests d'audiogramm vocaux : les listes de mots dissyllabiques de Fournier (1951). Pour des raisons de pr ticabilité des tests subjectifs, nous avons circonscrit notre corpus à un sous-ensemb de 6 listes de 10 mots, plus une liste qui a été utilisée pour l'entraînement des auditeu au test.

Les listes de mots ont été prononcées par trois locuteurs francophones afin d'o tenir différents types de voix : un homme (46 ans), une femme (47 ans) et u enfant (12 ans). Les trois locuteurs ont été enregistrés à l'aide du logiciel Reap (`http://www.reaper.fm`) dans la cabine audiométrique PETRA (`http://petra univ-tlse2.fr`) avec un microphone omnidirectionnel Sennheiser MD46 et u console de mixage TASCAM DM-3200. Le niveau des enregistrements a ensuite é égalisé en sonie par trois auditeurs sur une interface permettant d'accorder le nive de chaque fichier audio avec celui d'un fichier de référence. Le gain moyen défini p les trois auditeurs a ensuite été appliqué sur les fichiers audio. Ainsi, nous disposo d'un ensemble de 210 enregistrements dont 180 seront utilisés pour les tests et 30 po l'entraînement des auditeurs.

Afin de reproduire les effets de la presbyacousie sur ces enregistrements, no avons suivi la procédure décrite dans Nejime et Moore (1997), combinant les trait ments de signaux de parole pour simuler les effets de l'augmentation des seuils d'a dibilité, de la réduction de la sélectivité fréquentielle et du recrutement de sonie te que présentés dans la section précédente. Ces algorithmes dépendent de la sévérité trouble dont on cherche à simuler les effets, représentée en entrée du système par audiogramme tonal indiquant les pertes auditives en dB pour 15 fréquences allant 125 Hz à 16 kHz.

Les valeurs de 9 audiogrammes tonaux typiques des cas de presbyacousie pour d personnes ayant un âge théorique de 60 à 110 ans ont été calculées à partir des donné de l'étude épidémiologique de Cruickshanks *et al.* (1998) portant sur 3 753 sujets. L courbes de régression suivant les seuils d'audibilité pour chacun des âges théoriqu sont représentées dans la figure 2.

Un premier traitement des signaux de parole a consisté à appliquer des filtres cc respondant aux valeurs d'abaissement des seuils d'audibilité, et à simuler le recrut ment de sonie. Cette simulation a été réalisée en élevant l'enveloppe du signal à puissance 2, ce qui correspond à une atteinte moyenne. Nous avons choisi de ne p faire varier ce facteur pour limiter le nombre de nos variables ; de plus, il se trou que chez les sujets presbyacousiques des différences interindividuelles important peuvent apparaître quant à l'évolution de ce phénomène en fonction de l'aggravati des pertes auditives (Moore, 2007). Pour ce qui concerne la simulation de la perte sélectivité fréquentielle, dans l'algorithme de Nejime et Moore (1997) son applicati dépend directement de la sévérité du trouble dont on cherche à simuler les effets. Po cela un grade de sévérité est défini en fonction de la moyenne des pertes du sujet po

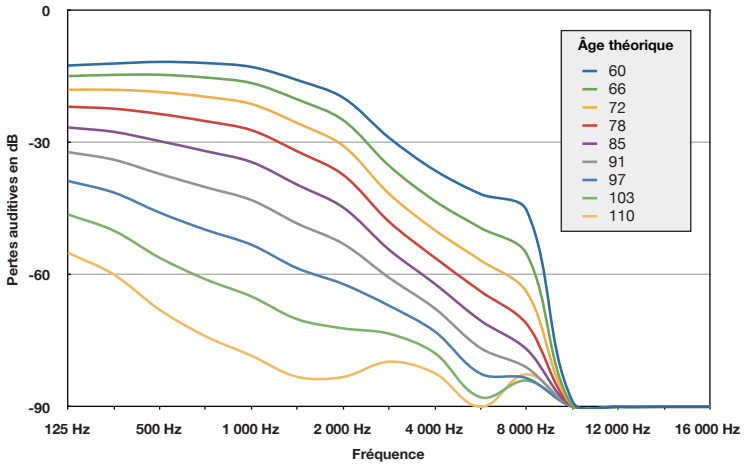

Figure 2. *Pertes typiques de la presbyacousie pour neuf âges théoriques compris entre 60 et 110 ans*

les fréquences comprises entre 2 kHz et 8 kHz ; en ce qui concerne les pertes que nous avons définies, les grades de sévérité correspondants sont indiqués dans la table 1. Plus les grades sont importants, plus le traitement simulant la perte de sélectivité fréquentielle (le lissage du spectre fréquentiel) est amplifié.

Fréquence / Âge	125 Hz	250 Hz	500 Hz	750 Hz	1000 Hz	1500 Hz	2000 Hz	3000 Hz	4000 Hz	6000 Hz	8000 Hz	Sévérité
60 ans	12	12	12	12	13	16	20	29	36	42	45	Légère
66,25 ans	15	15	15	15	16	20	25	35	43	49	55	Moyenne
72,5 ans	18	18	19	20	21	26	31	42	50	57	64	Moyenne
78,75 ans	22	22	24	25	27	32	37	48	56	64	71	Moyenne
85 ans	27	28	30	32	34	40	45	54	62	70	77	Sévère
91,25 ans	32	34	37	40	43	48	53	61	68	77	81	Sévère
97,5 ans	39	41	46	50	53	59	62	67	73	83	84	Sévère
103,75 ans	46	50	56	61	65	70	72	73	78	88	84	Sévère
110 ans	55	60	68	74	78	83	83	80	82	90	83	Sévère

Tableau 1. *Pertes auditives en dB pour des âges théoriques allant de 60 à 110 ans, et degrés de sévérité associés*

Ces 9 dégradations ont été appliquées aux 180 enregistrements initiaux, donnant ainsi un ensemble de 1 620 stimuli utilisables pour les tests d'intelligibilité. Une dizaine de stimuli ont été produits en complément pour servir à l'entraînement des auditeurs.

4. Mesures subjectives d'intelligibilité

La première partie de cette étude a consisté à réaliser des mesures subjectives d'intelligibilité de la parole auprès d'un panel de 30 auditeurs francophones natifs âgés de 18 à 30 ans (Fontan *et al.*, 2014) selon le protocole suivant :

– le niveau d'audition de chaque participant a été vérifié par un audiogramme tonal, avec comme critère d'exclusion une moyenne des pertes auditives supérieure à 15 dB entre 2 kHz et 8 kHz ;

– chaque auditeur a participé individuellement au test après une phase d'entraînement sur 10 stimuli (1 fichier non dégradé + 9 fichiers reproduisant les 9 niveaux de dégradation) issus du sous-corpus dédié ;

– le test proprement dit a consisté en l'écoute de 60 stimuli (6 non dégradés et 54 dégradés – soit 6 par condition de dégradation) diffusés aléatoirement, les auditeurs ayant pour consigne de répéter chaque mot qu'ils entendaient. Ils ont été par ailleurs encouragés à fournir une réponse même dans les cas où ils n'entendaient que partiellement voire pas du tout le mot diffusé. Cette méthode de mesure de l'intelligibilité, consistant à demander aux sujets de répéter des listes de mots puis à faire le compte de la proportion de mots correctement répétés, est tout à fait classique, notamment pour l'audioprothèse (cf. l'ouvrage du Collège National d'Audioprothèse, 2007, qui reprend de manière assez exhaustive les épreuves d'audition courantes dans le domaine).

Ces mesures subjectives ont servi de référence dans le reste de l'étude et seront comparées avec des mesures obtenues automatiquement en utilisant un système de reconnaissance de la parole dont les caractéristiques sont décrites dans la section suivante.

5. Description du système de reconnaissance automatique de la parole

Pour cette étude, nous avons utilisé un moteur de reconnaissance de la parole fondé sur le système Sphinx-3 (Seymore *et al.*, 1998) distribué par Carnegie Mellon University (CMU).

5.1. *Paramétrisation et modèles acoustiques*

Les modèles acoustiques que nous avons utilisés sont mis à disposition de la communauté par le LIUM pour la reconnaissance du français (Deléglise *et al.*, 2005 ; Estève, 2009). Ils ont été entraînés avec les bases d'apprentissage diffusées lors de la campagne ESTER2 (Galliano *et al.*, 2009), constituées d'enregistrements de tranches d'information diffusées sur plusieurs radios françaises. Il s'agit de modèles continus, composés de 5 725 sénones (phones en contexte), avec 22 mélanges de lois Gaussiennes par état. Ils sont destinés à traiter un signal à 16 KHz (bande passante utilisée : 133 à 6 855 Hz), avec une paramétrisation de type PLP (Hermansky, 1990) (avec

dérivées premières et secondes). Ils sont composés de 35 phonèmes et de 5 types de pauses.

5.2. *Adaptation au genre et à l'âge du locuteur*

Le contexte acoustique des enregistrements originaux (non dégradés) ainsi que le genre et l'âge des locuteurs étant différents de celui des enregistrements radio pour lesquels les modèles acoustiques ont été réalisés au départ, il est attendu que les résultats de reconnaissance sur nos enregistrements présentent de moins bonnes performances. Pour vérifier cette hypothèse, de premiers essais de reconnaissance ont été réalisés avec un modèle trigramme calculé à partir du corpus de ESTER2 (Galliano *et al.*, 2009) et deux lexiques différents : les ressources lexicales proposées par le LIUM (62 k formes) et un lexique contenant uniquement les formes correspondant aux 60 mots du test. Comme attendu, ces essais ont montré que les enregistrements de voix d'homme sont bien mieux reconnus que ceux réalisés avec les deux autres voix (femme et enfant, cf. figure 3).

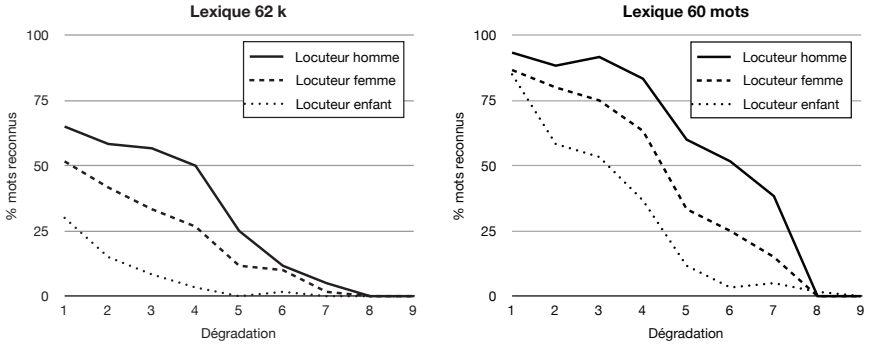

Figure 3. *Scores de reconnaissance de mots pour les 9 niveaux de dégradation, sans adaptation aux locuteurs (modèle de langage issu du corpus ESTER2)*

Notre travail initial a donc consisté à adapter le système pour minimiser les écarts de performances observés. Pour cela nous avons utilisé la technique de la normalisation en fonction de la longueur théorique du conduit vocal (*vocal tract length normalization – VTLN* ; Wegmann *et al.*, 1996). Cette technique repose sur l'idée, assez ancienne dans la littérature (Wakita, 1977 ; Bamberg, 1981), selon laquelle les fréquences des zones formantiques entretiennent une relation linéaire avec la longueur du conduit vocal du locuteur. Suivant cette idée il est donc possible de normaliser les signaux de parole produits par un locuteur particulier en vue de maximiser la vraisemblance entre les sons qu'il produit et les modèles acoustiques utilisés dans le système. L'essentiel des travaux sur la VTLN se sont concentrés sur la meilleure façon de déterminer ce facteur de distorsion fréquentielle optimal λ lors du traitement de signaux de parole inconnus *a priori* (Wegmann *et al.*, 1996) :

$$\lambda = arg\,max\ P(O|X, \lambda_k)$$

Dans notre cas, la détermination de ce facteur a pu être conduite directement à partir des stimuli prononcés par chacun des locuteurs, en observant le taux de reconnaissance obtenu pour chaque valeur de λ considérée. Nous avons ainsi testé différents facteurs de distorsion suivant la fonction linéaire inverse ($y = \frac{x}{\lambda}$) implémentée dans le moteur Sphinx-3 (Seymore *et al.*, 1998).

Ces différents facteurs de distorsion fréquentielle (*warping*) ont été appliqués par le système lors de la phase de paramétrisation des signaux de parole.

Comme pour le prétest décrit ci-dessus nous avons testé ces facteurs (1) avec un lexique de 62 k mots issu des ressources du LIUM puis (2) un lexique restreint aux 60 mots cibles (figures 4 et 5). En fonction de ces données, des équations de régression polynomiale d'ordre 3 ou 4 ont été calculées, avec pour contrainte d'obtenir un coefficient de détermination R^2 supérieur à 98 %. Les facteurs de distorsion donnant lieu aux maxima de ces fonctions ont ensuite été déterminés. Enfin, les facteurs de distorsion optimaux pour chaque locuteur dans les deux conditions de lexique ont fait l'objet d'un calcul de moyenne (tableau 2).

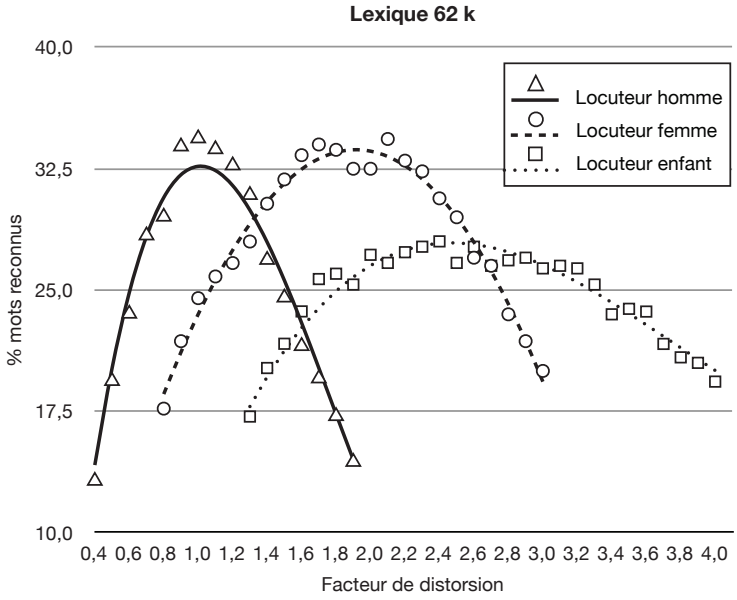

Figure 4. *Incidence du facteur de distorsion fréquentielle sur les scores de reconnaissance de mots (toutes dégradations confondues) : scores moyens et courbes de régression polynomiale, pour un lexique de 62 k formes*

Comme on peut le remarquer, le facteur de déformation optimal moyen obtenu pour le locuteur homme est de 1, ce qui signifie qu'aucune adaptation n'est néces-

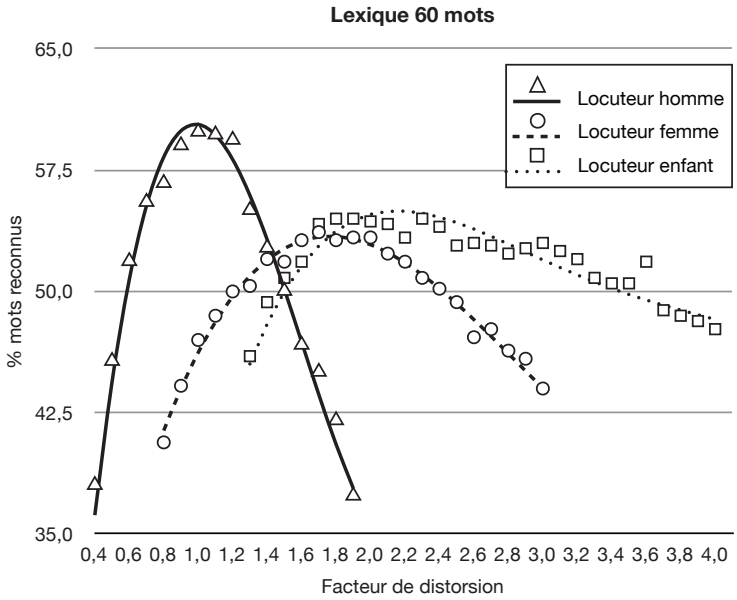

Figure 5. *Incidence du facteur de distorsion fréquentielle sur les scores de reconnaissance de mots (toutes dégradations confondues) : scores moyens et courbes de régression polynomiale, pour un lexique de 60 mots*

	Locuteur homme	Locuteur femme	Locuteur enfant
Lexique 62 k	1,02	1,91	2,32
Lexique 60 mots	0,99	1,77	2,16
Moyenne	**1**	**1,84**	**2,24**

Tableau 2. *Facteurs de distorsion fréquentielle optimaux observés pour les locuteurs homme, femme et enfant*

saire. En ce qui concerne les locuteurs femme et enfant, la technique VTLN permet d'améliorer les scores de manière importante. Le facteur de déformation optimal obtenu pour le locuteur enfant est supérieur à celui du locuteur femme, ce qui est cohérent avec le fait qu'un enfant a une longueur de conduit vocal généralement inférieure à celui d'une femme.

En conclusion, ce travail a permis de définir des facteurs de déformation optimaux moyens pour la paramétrisation des signaux de parole produits par les locuteurs

femme et enfant. Ces réglages sont utilisés pour la suite de l'étude, qui comprend deux phases :

– dans la première phase nous avons soumis les enregistrements dégradés à des tests perceptifs auprès d'auditeurs, ce qui nous a permis d'obtenir les scores décrits en section 6.3 ;

– dans la seconde phase ces mêmes enregistrements ont été soumis au système de reconnaissance pour disposer de résultats obtenus de manière automatique (section 6.4).

5.3. *Lexique et modèle de langage*

Les énoncés utilisés comme stimuli correspondent à des syntagmes nominaux élémentaires de type : article et nom. Un modèle de langage spécifique a été conçu afin :

– d'une part que les résultats du moteur de reconnaissance se présentent sous une forme linguistique qui soit la plus proche possible des réponses données par les sujets humains lors des tests d'intelligibilité ;

– d'autre part d'intégrer une dimension cognitive, c'est-à-dire obtenir un modèle qui ait une couverture lexicale assez large et qui puisse s'approcher de la connaissance lexicale générale d'un locuteur natif.

Dans ce contexte ont été considérées (i) les consignes lues aux sujets avant les tests et (ii) les réponses « alternatives » données par les sujets, c'est-à-dire leurs réponses non nulles mais erronées, qui sont au nombre de 375 dans cette étude. Dans la grande majorité des cas les sujets ont donné comme formes alternatives des noms masculins (96,2 %) et, parmi ces noms, des noms commençant par une consonne (97,9 %). Les sujets paraissent donc avoir été fortement influencés par la nature du matériel linguistique contenu dans les listes de Fournier (1951) que nous avons utilisées. Ces dernières sont en effet exclusivement constituées de noms masculins dissyllabiques commençant par une consonne (cf. annexe A).

Pour tenir compte de cet effet le modèle de langage a été conçu comme un modèle statistique de type bigrammes (construits sur des suites de séquences de type <*s*> *article nom* </*s*>) avec un vocabulaire limité à des noms masculins commençant par une consonne (15146 noms distincts). Afin que les statistiques reflètent au mieux la fréquence d'occurrence de ces formes en français oral, nous avons utilisé la base de données Lexique 3.8[2] (`http://www.lexique.org`).

Les représentations phonétiques des mots pris en compte dans notre lexique sont disponibles dans les ressources proposées par le LIUM (lexique de 62 k formes). Notons que la tâche initiale constitutive de ces ressources (l'analyse de tranches d'information radio) est différente de celle de notre étude et donc une partie importante des

2. Et plus précisément les fréquences des lemmes relevées dans des corpus constitués de sous-titres de films (New *et al.*, 2007).

mots de notre modèle de langage ne s'y trouve pas : le lexique utilisé documente les formes phonétiques possibles pour seulement 6 492 noms inclus dans le modèle. Les 60 noms du corpus de test sont néanmoins présents dans cet ensemble.

6. Production des mesures d'intelligibilité

À partir de cette section, tous les scores automatiques produits impliquent le travail d'adaptation au locuteur (VTLN – section 5.2) et reposent sur l'utilisation du modèle de langage bigramme issu de la base Lexique 3.8 et présenté dans la section 5.3.

6.1. *Annotations manuelles des réponses données par les auditeurs dans les tests subjectifs*

Afin de comparer les résultats des tests subjectifs et automatiques, les transcriptions orthographique et phonétique des 1 800[3] réponses données par l'ensemble des sujets pour chacun des différents stimuli entendus ont été réalisées en deux étapes successives. Un prétraitement automatique a permis d'obtenir un premier alignement des enregistrements de parole avec le système de reconnaissance de la parole décrit en section 5. L'alignement phonétique a été réalisé en utilisant les modèles acoustiques du LIUM décrits en section 5.1. Les mots hors vocabulaire ont été rajoutés au lexique 62 k en utilisant une phonétisation issue de BDLEX (de Calmès et Pérennou, 1998). Une seconde phase d'annotation manuelle de l'alignement automatique a ensuite permis de corriger les phonèmes erronés ou manquants. Les corrections ont porté principalement sur les phonèmes reconnus, et non pas sur la modification des frontières trouvées par le système automatique.

6.2. *Scores observés*

À travers les performances des sujets humains et du système de reconnaissance automatique de la parole, deux types de scores ont été calculés :

1) **scores de reconnaissance des mots**. Il s'agit de scores « classiques » d'intelligibilité, correspondant au pourcentage de mots correctement reconnus par les sujets humains et par la machine. Dans ce cadre, seule la reconnaissance du nom commun cité par le sujet est comptée. Cette reconnaissance doit être totale : tous les phonèmes doivent être correctement reconnus ;

2) **scores de distance phonologique des réponses**. Il s'agit de scores plus fins, traduisant la distance entre les réponses des sujets (ainsi que du moteur de RAP) et les

3. Les 1 800 réponses des sujets correspondent à la répétition des 1 620 stimuli simulant les effets de la presbyacousie ainsi qu'à la répétition de 180 stimuli non dégradés que nous n'utilisons pas dans le cadre de cette étude.

stimuli. Dans ce cadre deux procédures ont été effectuées pour calculer les distances entre les formes attendues et les formes reconnues :

- *le calcul de la distance de Levenshtein* (Levenshtein, 1966). L'algorithme de Levenshtein permet de quantifier la distance entre deux chaînes de symboles, et est utilisé pour quantifier la distance entre unités linguistiques, notamment en dialectologie (Beijering *et al.*, 2008 ; Heeringa, 2004). Le principe de l'algorithme de Levenshtein est de comptabiliser les modifications minimales nécessaires pour passer d'une chaîne a à une chaîne b, en termes d'ajout, de suppression ou de substitution de symboles (ici des phonèmes). Ainsi la distance entre les chaînes /p@ti/ (petit) et /apeti/ (appétit) est de valeur 2, car pour passer de la première chaîne à la seconde il faut effectuer deux opérations : 1) un ajout de phonème, et 2) une substitution de phonème (/@/ est remplacé par /e/). De manière mathématique, avec a et b deux chaînes de longueurs i et j, la distance de Levenshtein $\text{lev}_{a,b}$ peut se formaliser comme suit[4] :

$$\text{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{si } \min(i,j) = 0 \\ \min \begin{cases} \text{lev}_{a,b}(i-1,j) + 1 \\ \text{lev}_{a,b}(i,j-1) + 1 \\ \text{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{sinon} \end{cases}$$

où $1_{(a_i \neq b_j)}$ est une fonction renvoyant 0 lorsque $a_i = b_j$ et 1 lorsque $a_i \neq b_j$

- *le calcul de la distance de Levenshtein pondérée en fonction de la nature des phonèmes substitués*. Dans sa version première l'algorithme de Levenshtein attribue le même poids à toutes les substitutions de symboles. En suivant la procédure décrite dans Fontan (2012) nous avons modifié l'algorithme de Levenshtein afin de prendre en compte le fait que deux phonèmes peuvent être plus ou moins proches selon le nombre de traits distinctifs qu'ils partagent (Riegel, 1994). Dans ce nouveau calcul, que nous appelons *distance de Levenshtein pondérée* le coût de la substitution d'une consonne par une autre est équivalent au nombre de traits distinctifs que ces deux consonnes partagent, divisé par le nombre total de traits. Ainsi la substitution du phonème /p/ par le phonème /b/, au lieu de générer un coût de 1, compte 1/8 (soit 0,125) car, parmi les 8 traits phonologiques considérés, seul le trait du voisement distingue ces deux phonèmes. Le même calcul est effectué dans le cas de substitutions de voyelles, et le coût de la substitution d'une consonne par une voyelle, et *vice versa*, est maximal (c'est-à-dire égal à 1).

Ces deux derniers scores ont été normalisés en divisant la distance phonologique par le nombre de phonèmes du mot cible concerné.

Les scores de reconnaissance et les scores de distance phonologique ont été calculés aussi bien sur les données subjectives que sur les données objectives. Les détails relatifs à ces scores sont décrits dans les deux sections suivantes.

4. Cette formalisation est inspirée des travaux sur la distance d'édition de Wagner et Fischer (1974).

6.3. *Scores issus des tests subjectifs*

Dans cette partie de notre étude, nous avons calculé les scores de reconnaissance de mots (mesurant l'intelligibilité) ainsi que les distances phonologiques entre les stimuli et les réponses des sujets.

Dans un premier temps nous avons considéré un score de reconnaissance « binaire » : 1 lorsqu'un mot était totalement reconnu et 0 lorsque le mot n'était reconnu que partiellement ou pas du tout. Les scores obtenus pour chacun des 9 niveaux de dégradation numérotés de 1 à 9 sur l'ensemble des mots utilisés lors du test sont représentés en partie gauche de la figure 6.

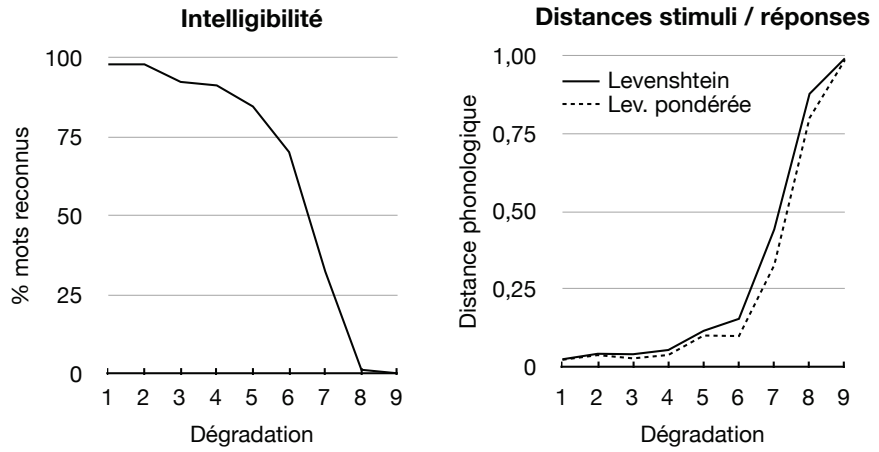

Figure 6. *Scores subjectifs pour les 9 niveaux de dégradation : scores d'intelligibilité (à gauche) et distances phonologiques moyennes entre les réponses des sujets et le mot attendu (à droite)*

La partie droite de la figure 6 montre l'évolution des distances phonologiques (distance de Levenshtein classique et distance de Levenshtein pondérée) au gré des dégradations. Ces distances sont calculées en comparant la transcription de la réponse du sujet avec la transcription du mot de référence (ou mot cible).

Sur cette figure, les scores correspondant aux stimuli originellement prononcés par les trois locuteurs ont été moyennés. En effet, lors d'une analyse précédant cette étude (Fontan *et al.*, 2014), il avait été observé un effet hautement significatif du niveau de dégradation sur les scores d'intelligibilité ($P < 0,001$) mais non significatif en ce qui concerne le locuteur (homme, femme, enfant) ou encore l'ordre de présentation des stimuli.

On peut remarquer les tendances inverses des courbes : plus la dégradation augmente, plus l'intelligibilité baisse et plus la distance phonologique moyenne entre les stimuli diffusés et les réponses des sujets augmente. Concernant les scores de recon-

naissance des mots, la différence entre les niveaux de dégradation les plus faibles (1 et 2) est minime, de même qu'entre les niveaux les plus élevés (8 et 9) pour ce qui concerne les scores d'intelligibilité.

6.4. *Scores issus des résultats de reconnaissance automatique*

6.4.1. *Scores binaires*

Le système de reconnaissance automatique utilisé offre la possibilité de récupérer les *n* meilleures hypothèses du processus de reconnaissance. Nous avons utilisé cette information pour voir si le fait d'avoir connaissance de plusieurs hypothèses pouvait avoir un impact sur les résultats. En conséquence, pour chaque stimulus, deux scores binaires de reconnaissance ont été produits :

1) un premier score indiquant si le mot cible a été reconnu par le système comme étant le mot le plus probable ;

2) un second score indiquant si le mot cible fait partie des 10 mots retenus comme les plus probables par le système.

Les pourcentages obtenus pour ces deux scores et les 9 dégradations sont représentés dans la figure 7. Contrairement aux données observées chez les sujets humains, nous avions dans notre étude préliminaire observé des différences pour les stimuli enregistrés par les trois locuteurs (cf. section 5.2), nous donnons les résultats de manière séparée pour chacun d'entre eux en figure 7.

Comme on peut l'observer, les scores obtenus de manière automatique sont généralement inférieurs à ceux relevés chez les sujets humains. Ces différences de performances du système sont plus ou moins marquées en fonction des locuteurs :

– concernant les mots reconnus et proposés en premier résultat (meilleure hypothèse), les scores du locuteur homme sont en moyenne de 35,2 %, du locuteur femme 31,3 % et du locuteur enfant 29,1 % ;

– concernant les mots reconnus et présents dans les 10 premières hypothèses, les scores du locuteur homme sont en moyenne de 49,4%, du locuteur femme 43,3 % et du locuteur enfant 42,8 %.

Par ailleurs, les scores automatiques semblent suivre une évolution plus linéaire que les scores subjectifs, ou plus précisément présenter des effets de plancher et de plafond moins marqués. En revanche, des chutes subites des scores apparaissent entre les niveaux 1 et 2, ainsi qu'entre les niveaux 4 et 5. Ces chutes s'observent pour les trois locuteurs.

6.4.2. *Scores de distance phonologique entre les stimuli et les mots reconnus par le système de RAP*

Ici également nous pouvons observer que les performances du système de RAP sont globalement inférieures à celles des sujets humains, avec des distances phonolo-

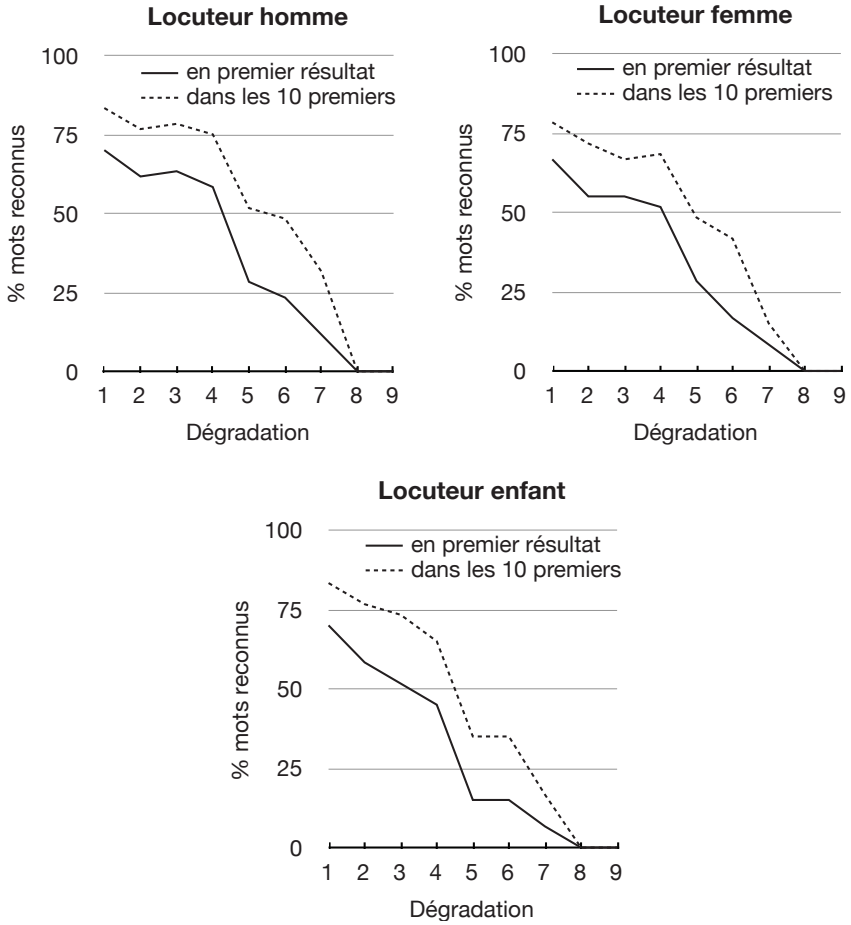

Figure 7. *Scores de reconnaissance de mots observés en sortie du système de RAP pour les 9 niveaux de dégradation*

giques supérieures entre mots attendus et mots produits. De même, ces distances sont plus ou moins importantes en fonction des locuteurs :

– concernant les distances de Levenshtein, les scores du locuteur homme sont en moyenne de 0,39, du locuteur femme 0,42 et du locuteur enfant 0,44 ;

– concernant les distances de Levenshtein pondérées, les scores du locuteur homme sont en moyenne de 0,27, du locuteur femme 0,31 et du locuteur enfant 0,30.

Tout comme pour les résultats de reconnaissance des mots, nous observons, quel que soit le locuteur, des augmentations abruptes des scores entre les niveaux 4 et 5.

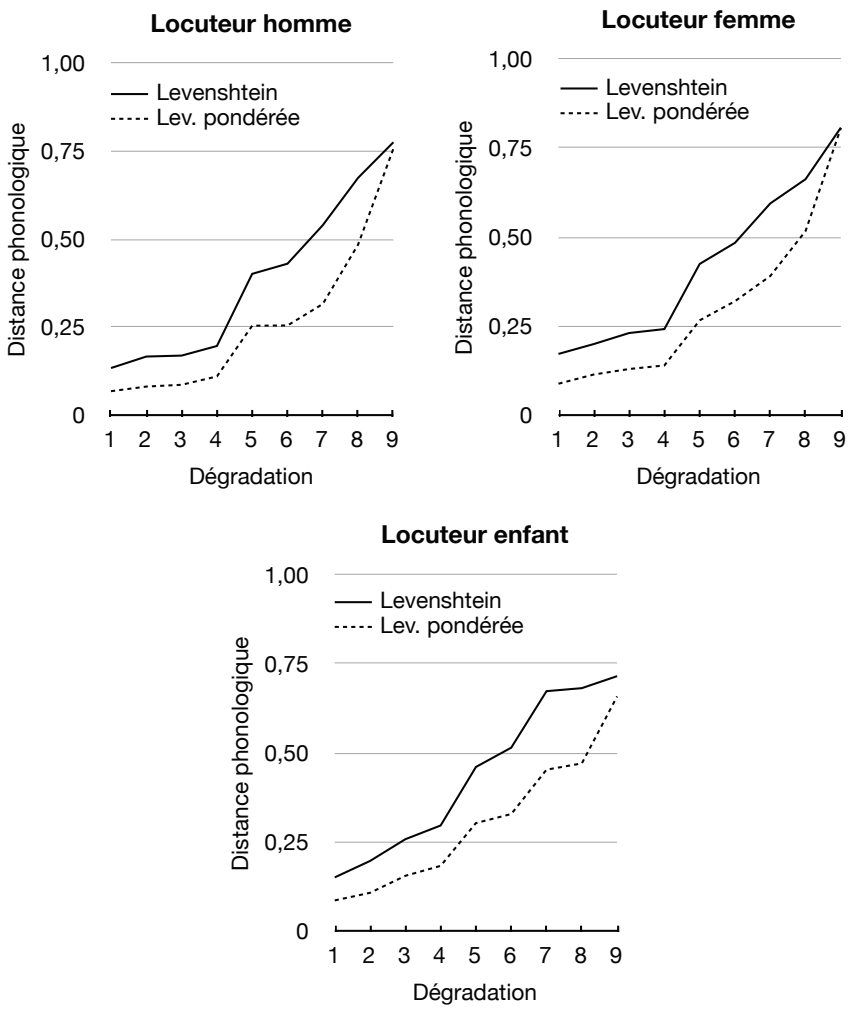

Figure 8. *Distances phonologiques moyennes entre la sortie du système de RAP et les mots attendus, pour les 9 niveaux de dégradation*

Les scores semblent ici aussi subir des effets de plancher et de plafond moins marqués que pour les scores subjectifs.

7. Analyse des résultats en vue de la prédiction des scores subjectifs par le système de traitement automatique de la parole

Pour avoir des indications concernant l'aptitude des scores issus du système de RAP à prédire les scores subjectifs, nous nous sommes intéressés aux scores les plus fins issus de cette étude et avons étudié la force d'association entre :

1) les distances de Levenshtein (resp. Levenshtein pondérées) entre les noms attendus et les productions des sujets humains (cf. graphique de droite dans la figure 6) ;

2) les distances de Levenshtein (resp. Levenshtein pondérées) entre les noms attendus et les productions du système de RAP (meilleure hypothèse – cf. figure 8) ;

3) les distances de Levenshtein moyennes (resp. de Levenshtein pondérées moyennes) entre les noms attendus et les productions du système de RAP pour les n meilleures hypothèses ($n \leq 10$).

Nous avons utilisé pour cela un calcul de corrélation de Pearson. Les résultats sont donnés dans le tableau 3. Toutes les corrélations observées sont de force moyenne à forte ($0,49 \leq r \leq 0,76$) et hautement significatives ($P < 0,001$). On peut noter que, globalement, les coefficients de corrélation sont plus élevés pour le locuteur homme (r moyen = 0,68) que pour les locuteurs femme (r moyen = 0,65) et enfant (r moyen = 0,57). De même du côté des scores du système de RAP, ce sont les scores de distances de Levenshtein *pondérées* qui présentent les corrélations les plus fortes (r moyen = 0,68), en comparaison des distances de Levenshtein usuelles (r moyen = 0,59). Si nous tenons compte des valeurs issues des mesures relatives aux stimuli prononcés par les trois locuteurs, alors la mesure automatique donnant lieu aux corrélations les plus fortes est celle de la distance de Levenshtein pondérée moyenne pour les n meilleures hypothèses (*Dist. de Lev. pondérée moyenne sur n-meilleures hypothèses* dans le tableau 3), qui donne un coefficient de corrélation moyen égal à 0,69.

Pour obtenir les résultats présentés dans le tableau 3, nous avons considéré chaque stimulus (c'est-à-dire chaque mot x prononcé par le locuteur y avec la dégradation z) comme une observation. Nous avons dans un second temps effectué des calculs plus généraux et observé la capacité des scores issus du système de RAP à prédire les scores d'intelligibilité moyens obtenus par les sujets humains pour les différentes dégradations de notre corpus. Pour cela nous avons testé plusieurs types d'équations de régression à partir des distances de Levenshtein pondérées moyennes obtenues pour les 10 dégradations. Le meilleur résultat a été obtenu grâce à une régression polynomiale d'ordre 5, qui donne un coefficient R^2 de 0,997.

8. Conclusion

Cette étude constitue le premier volet d'un programme de recherche visant à terme l'élaboration d'un système de mesure automatique de l'intelligibilité et de la compréhension de la parole. À cette fin des mesures automatiques observées *via* des systèmes

Locuteur homme

Humains \ Système RAP	Distance de Levenshtein	Distance de Levenshtein pondéree	Dist. de Lev. moyenne sur n-bests	Dist. de Lev. pondérée moyenne sur n-bests
Distance de Levenshtein	0,675**	0,757**	0,627**	0,706**
Distance de Levenshtein pondéree	0,642**	0,747**	0,606**	0,711**

** $P < 0{,}001$

Locuteur femme

Humains \ Système RAP	Distance de Levenshtein	Distance de Levenshtein pondéree	Dist. de Lev. moyenne sur n-bests	Dist. de Lev. pondérée moyenne sur n-bests
Distance de Levenshtein	0,588**	0,694**	0,628**	0,732**
Distance de Levenshtein pondéree	0,568**	0,687**	0,600**	0,731**

** $P < 0{,}001$

Locuteur enfant

Humains \ Système RAP	Distance de Levenshtein	Distance de Levenshtein pondéree	Dist. de Lev. moyenne sur n-bests	Dist. de Lev. pondérée moyenne sur n-bests
Distance de Levenshtein	0,521**	0,583**	0,558**	0,650**
Distance de Levenshtein pondéree	0,494**	0,576**	0,521**	0,632**

** $P < 0{,}001$

Tableau 3. *Coefficients de corrélation de Pearson observés en comparant les mesures de distances phonologiques relevées chez les sujets humains vs en sortie du système de RAP*

de reconnaissance automatique de la parole (RAP) sont comparées à des mesures subjectives de référence, issues de tests allant de la répétition de mots isolés à la compréhension de phrases en contexte.

Cette étude était consacrée à la prédiction de scores subjectifs d'intelligibilité de listes de mots, un test communément conduit dans le domaine de l'oto-rhino-laryngologie. À cette fin, un corpus simulant les effets de la presbyacousie sur la perception de la parole à différents niveaux de sévérité a été produit, et des données de référence (scores d'intelligibilité) ont été relevées auprès de 30 sujets qui ont eu pour tâche de répéter les mots du corpus. Enfin, un système de RAP spécialement adapté pour cette tâche a permis d'observer différents scores qui ont été mis en correspondance avec les mesures subjectives.

De manière générale, les résultats sont très encourageants dans la mesure où de fortes corrélations ont été observées entre mesures objectives et subjectives. Les scores subjectifs d'intelligibilité relevés pour les différents niveaux de dégradation appliqués sur le corpus peuvent être prédits de manière très précise à partir de l'observation de données issues du système de RAP (distances phonologiques entre les stimuli cibles et les réponses du système). Cette interprétation des résultats doit néanmoins être nuancée car, pour être véritablement à même de juger les performances du système à prédire les scores subjectifs, le système devrait être confronté à de nouvelles données de référence. La piste que nous privilégions dans ce cadre est le test du système en si-

tuation, c'est-à-dire en observant auprès de la population cible (personnes atteintes de presbyacousie) si les indications données par le système en fonction d'un réglage de prothèse particulier est corrélé avec les performances ou le ressenti du patient.

Par ailleurs, nos données font ressortir plusieurs éléments qui nécessitent de conduire des études ultérieures. Ainsi, malgré l'adaptation du système de RAP aux trois locuteurs de notre corpus *via* la technique de la VTLN, la sensibilité du système au sexe et à l'âge du locuteur persiste. Cela est certainement dû au déséquilibre initial dans le corpus audio ayant servi pour la création des modèles acoustiques utilisés dans cette étude, le corpus ayant servi à la campagne d'évaluation ESTER, qui est majoritairement composé de voix d'hommes (environ les deux tiers du corpus – cf. Galliano *et al.*, 2006). Puisque cette sensibilité au locuteur n'est pas observée dans les données subjectives de référence (un modèle linéaire mixte n'avait pas mis au jour d'effet significatif du locuteur sur les scores, cf. Fontan *et al.*, 2014), une partie du travail futur sera consacrée à la mise en place d'une meilleure adaptation du système pour les locuteurs femme et enfant ; dans ce cadre d'autres techniques d'adaptation comme la MLLR (Leggetter et Woodland, 1995) ou le maximum *a posteriori* – MAP, (voir Chengalvarayan et Deng, 2001) sont envisagées.

De même, contrairement aux scores observés chez les sujets humains, les scores issus du système de RAP démontrent des chutes relativement marquées lors du passage des niveaux de dégradation 4 à 5, ainsi que dans une plus faible mesure des niveaux 1 à 2 : les scores de reconnaissance ont tendance à chuter plus rapidement et les distances phonologiques entre les réponses attendues et les réponses fournies à augmenter subitement. Il est possible que ces chutes marquées dans les scores du système de RAP reflètent le passage des paliers définis dans l'algorithme de Nejime et Moore (1997) pour simuler la perte de sélectivité fréquentielle correspondant à différents grades de sévérité. Si l'on se réfère au tableau 1, les passages du niveau de dégradation 1 (60 ans) à 2 (66,25 ans) et du niveau 4 (78,75 ans) à 5 (85 ans) constituent précisément les paliers entre les différents degrés de sévérité : léger, moyen et important – paliers au-delà desquels les traitements simulant la perte de sélectivité fréquentielle sont appliqués de manière plus intense : le lissage effectué sur le spectre fréquentiel est renforcé. À chaque palier les pics formantiques deviennent donc moins prégnants, et il est logique que le système de RAP ait plus de difficulté à rapprocher les observations (MFCC) des modèles acoustiques correspondants. Ces chutes ne se retrouvant pas dans les scores subjectifs, il semble que, contrairement au système de RAP, les auditeurs parviennent à compenser la perte d'information engendrée par ces traitements par des mécanismes cognitifs de type *top-down*, c'est-à-dire faisant intervenir des éléments de connaissance supérieure. Dans leur étude Baer et Moore (1993) ont fait la même observation : leurs résultats montrent que le lissage du spectre fréquentiel n'a aucune incidence sur l'intelligibilité de la parole lorsque cette dernière est diffusée dans le silence ; en revanche lorsque la parole est diffusée dans du bruit, ce type de traitement du signal de parole a un effet très marqué sur les performances des sujets. Pour pallier cette différence entre les performances du système et les observations de référence, nous envisageons d'entraîner notre système sur des fichiers audio ayant subi un traitement de lissage fréquentiel, afin de le rendre moins sensible à ce

type de dégradation. De même, les observations de Baer et Moore (1993) confirment notre motivation à mener le même type d'étude sur la parole diffusée dans le bruit, d'autant plus que les patients presbyacousiques se plaignent de manière récurrente de leurs difficultés de compréhension dans de tels environnements (en particulier dans le brouhaha, cf. Moore, 2007).

Enfin, comme nous l'avons mentionné dans l'introduction d'autres tests ont été conduits afin de relever les performances d'auditeurs dans des tâches impliquant la répétition ou la compréhension d'énoncés plus complexes : répétition des phrases de la version française du *Hearing in Noise Test* (HINT, Vaillancourt *et al.*, 2005), et exécution de commandes verbales sur des images (Fontan *et al.*, 2013). Travailler à la prédiction des données issues de ces tests nous permettra à la fois de dépasser le cadre de la répétition de mots isolés pour aller vers des situations d'écoute plus valides d'un point de vue écologique, et également d'accroître le nombre de données à notre disposition pour confirmer les tendances observées dans cette étude.

Remerciements

Cette étude a été réalisée dans le cadre du projet numéro 12052648 « Mesure de la compréhension de la parole : dispositif électronique de préréglage des prothèses auditives basé sur une approche cognitive » financé par la région Midi-Pyrénées dans le cadre de l'appel à projets AGILE-IT 2012 et soutenu par le fonds européen régional de développement régional (FEDER). Le projet est porté par Archean Technologies en partenariat avec les laboratoires IRIT, OCTOGONE (*via* PETRA) et le service ORL de l'hôpital Purpan et fait suite à un dépôt de brevet européen (Aumont et Wilhem-Jaureguiberry, 2009).

9. Bibliographie

ANSI S3.5, *Methods for the Calculation of the Speech Intelligibility Index*, American National Standards Institute, 2007.

Ariöz U., *Developing Subject-Specific Hearing Loss Simulation to Apply Different Frequency Lowering Algorithms for the Enhancement of Sensorineural Hearing Losses*, Middle-East Technical University, 2012. Thèse de doctorat, accessible à l'adresse `http://etd.lib.metu.edu.tr/upload/12614929/index.pdf`.

Aumont X., Wilhem-Jaureguiberry A., *Brevet européen 2136359 – Procédé et appareil de mesure de l'intelligibilité d'un dispositif de diffusion sonore*, Institut National de la Propriété Industrielle, 2009.

Baer T., Moore B. C. J., « Effects of Spectral Smearing on the Intelligibility of Sentences in the Presence of Noise », *Journal of the Acoustical Society of America*, vol. 94, nº 4, p. 1229-1241, 1993.

Bamberg P., « Vocal Tract Normalization », 1981. Verbez. Internal Technical Report.

Beijering K., Gooskens C., Heeringa W., « Modeling Intelligibility and Perceived Linguistic Distances by Means of the Levenshtein Algorithm », *in* M. van Koppen, B. Botma (eds), *Linguistics in the Netherlands 2008*, John Benjamins, Amsterdam, p. 13-24, 2008.

Bouccara D., Ferrary E., Mosnier I., Bozorg Grayeli A., Sterkers O., « Presbyacousie », *EMC – Oto-Rhino-Laryngologie*, vol. 2, nº 4, p. 329–342, 2005.

Chengalvarayan R., Deng L., « A Maximum a Posteriori Approach to Speaker Adaptation Using the Trended Hidden Markov Model », *IEEE Transactions on Speech and Audio Processing*, vol. 9, nº 5, p. 549-557, 2001.

Collège National d'Audioprothèse, *Précis d'audioprothèse, Tome 1 : L'appareillage de l'adulte. Le bilan d'orientation prothétique (seconde édition)*, Les éditions du collège national d'audioprothèse, 2007.

Cruickshanks K. J., Wiley T. L., Tweed T. S., Klein B. E., Klein R., Mares-Perlman J. A., Nondahl D. M., « Prevalence of Hearing Loss in Older Adults in Beaver Dam, Wisconsin. The Epidemiology of Hearing Loss Study », *American Journal of Epidemiology*, vol. 148, nº 9, p. 879-886, 1998.

de Calmès M., Pérennou G., « BDLEX : A Lexicon for Spoken and Written French », *in* ELRA (ed.), *1st International Conference on Langage Resources & Evaluation (LREC1998)*, Grenade (Espagne), p. 1129-1136, 1998.

Deléglise P., Estève Y., Meignier S., Merlin T., « LIUM Speech Transcription System : A CMU Sphinx III-Based System for French Broadcast News », *Proceedings of Interspeech'05*, Lisbonne (Portugal), p. 1653-1656, 2005.

Estève Y., *Traitement automatique de la parole : contributions*, Université du Maine, 2009. Mémoire d'habilitation à diriger des recherches.

Fontan L., *De la mesure de l'intelligibilité à l'évaluation de la compréhension de la parole pathologique en situation de communication*, Université de Toulouse, 2012. Thèse de doctorat, accessible à l'adresse `https://tel.archives-ouvertes.fr/file/index/docid/797883/filename/Fontan_Lionel.pdf`.

Fontan L., Gaillard P., Woisard V., « Comprendre et agir : les tests pragmatiques de compréhension de la parole et EloKanz », *in* R. Sock, B. Vaxelaire, C. Fauth (eds), *La voix et la parole perturbées*, CIPA, Mons, p. 131-144, 2013.

Fontan L., Magnen C., Tardieu J., Gaillard P., « Simulation des effets de la presbyacousie sur l'intelligibilité et la compréhension de la parole dans le silence et dans le bruit », *30e édition des Journées d'étude sur la parole (JEP 2014)*, Le Mans, 2014.

Fontan L., Tardieu J., Gaillard P., Woisard V., Ruiz R., « Relationship Between Speech Intelligibility and Speech Comprehension in Babble Noise », *Journal of Speech, Language and Hearing Research*, accepté.

Fournier J.-E., *Audiométrie vocale : les épreuves d'intelligibilité et leurs applications au diagnostic, à l'expertise et à la correction prothétique des surdités*, Maloine, 1951.

Galliano S., Geoffrois E., Gravier G., Bonastre J. F., Mostefa D., Choukri K., « Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News », *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC 2006)*, Gênes (Italie), p. 315-320, 2006.

Galliano S., Gravier G., Chaubard L., « The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts », *Proceedings of Interspeech '09*, Brighton (Royaume-Uni), p. 2583-2586, 2009.

Heeringa W., *Measuring Dialect Pronunciation Differences using Levenshtein Distance*, Université de Groningen, 2004. Thèse de doctorat, accessible à l'adresse `http://www.let.rug.nl/~heeringa/dialectology/thesis/thesis.pdf`.

Hermansky H., « Perceptual Linear Predictive (PLP) Analysis of Speech », *Journal of the Acoustical Society of America*, vol. 87, n° 4, p. 1738-1752, 1990.

Humes L. E., Roberts L., « Speech-Recognition Difficulties of the Hearing-Impaired Elderly : The Contributions of Audibility », *Journal of Speech and Hearing Research*, vol. 33, n° 4, p. 726-735, 1990.

Hustad K. C., « The Relationship Between Listener Comprehension and Intelligibility Scores for Speakers With Dysarthria », *Journal of Speech, Language and Hearing Research*, vol. 51, n° 3, p. 562-573, 2008.

Leggetter C. J., Woodland P. C., « Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models », *Computer Speech and Language*, vol. 9, n° 2, p. 171-185, 1995.

Levenshtein V. I., « Binary Codes Capable of Correcting Deletions, Insertions, and Reversals », *Soviet Physics Doklady*, vol. 10, n° 8, p. 707-710, 1966.

Maier A., Haderlein T., Eysholdt U., Rosanowski F., Batliner A., Schuster M., Nöth E., « PEAKS – A System for the Automatic Evaluation of Voice and Speech Disorders », *Speech Communication*, vol. 51, n° 5, p. 425-437, 2009.

Moore B. C. J., *Cochlear Hearing Loss : Physiological, Psychological and Technical Issues*, Wiley, 2007.

Moore B. C. J., Glasberg B. R., « Simulation of the Effects of Loudness Recruitment and Threshold Elevation on the Intelligibility of Speech in Quiet and in Background of Speech », *Journal of the Acoustical Society of America*, vol. 94, n° 4, p. 2050-2062, 1993.

Nejime Y., Moore B. C. J., « Simulation of the Effect of Threshold Elevation and Loudness Recruitment Combined With Reduced Frequency Selectivity on the Intelligibility of Speech in Noise », *Journal of the Acoustical Society of America*, vol. 102, p. 603-615, 1997.

New B., Brysbaert M., Veronis J., Pallier C., « The Use of Film Subtitles to Estimate Word Frequencies », *Applied Psycholinguistics*, vol. 28, n° 4, p. 661-677, 2007.

Riegel M., *Les sons du français : phonétique et phonologie*, PUF, 1994.

Schuster M., Maier A., Haderlein T., Nkenke E., Wohlleben U., Rosanowski F., Eysholdt U., Nöth E., « Evaluation of Speech Intelligibility for Children With Cleft Lip and Palate by Means of Automatic Speech Recognition », *International Journal of Pediatric Otorhinolaryngology*, vol. 70, n° 10, p. 1741-1747, 2006.

Seymore K., Chen S., Doh S., Eskenazi M., Gouvea E., Raj B., Ravishankar M., Rosenfeld R., Siegler M., Stern R. *et al.*, « The 1997 CMU Sphinx-3 English Broadcast News Transcription System », *Proceedings of the 1998 DARPA Speech Recognition Workshop*, p. 55-59, 1998.

Vaillancourt V., Laroche C., Mayer C., Basque C., Nali M., Eriks-Brophy A., Soli S. D., Giguère C., « Adaptation of the HINT (Hearing in Noise Test) for Adult Canadian Francophone Populations », *International Journal of Audiology*, vol. 44, n° 6, p. 358-369, 2005.

Wagner R. A., Fischer M. J., « The String-to-String Correction Problem », *Journal of the Association for Computer Machinery*, vol. 21, n° 1, p. 168-173, 1974.

Wakita H., « Normalization of Vowels by Vocal-Tract Length and its Application to Vowel Identification », *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 25, n^{o} 2, p. 183-192, 1977.

Wegmann S., McAllaster D., Orloff J., Peskin B., « Speaker Normalization on Conversational Telephone Speech », *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Atlanta (États-Unis d'Amérique), p. 339-341, 1996.

A. Liste de mots utilisée dans l'étude

le parfum	le cheveu	le turbot
le cachet	le citron	le hoquet
le ravin	le rocher	le plastron
le dragon	le caveau	le raisin
le lilas	le soldat	le croyant
le récit	le muguet	le fourré
le couvent	le bouton	le taquin
le galon	le verrier	le morceau
le courrier	le fourneau	le normand
le crapaud	le bassin	le poisson
le rideau	le carton	le coupon
le tampon	le pruneau	le marché
le boudin	le regret	le doyen
le vacher	le dément	le torrent
le débit	le répit	le festin
le marteau	le colon	le cliché
le cadran	le respect	le drapeau
le requin	le bilan	le juron
le goudron	le dépôt	le pari
le clocher	le rachat	le sujet