Proximité rythmique entre apprenants et natifs du français : évaluation d'une métrique basée sur le CEFC

Sylvain Coulange

Proximité rythmique entre apprenants et natifs du français

Évaluation d'une métrique basée sur le CEFC

COULANGE
Sylvain

Sous la direction de Solange ROSSATO

Laboratoire d'Informatique de Grenoble – LIG

UFR LLASIC
Département d'informatique intégrée en Langues, Lettres et Langage (I3L)

Mémoire de master 2 recherche en Sciences du Langage - 20 crédits

Parcours : Industries de la langue

Année universitaire 2018-2019

Proximité rythmique entre apprenants et natifs du français

Évaluation d'une métrique basée sur le CEFC

COULANGE
Sylvain

Sous la direction de Solange ROSSATO

Laboratoire d'Informatique de Grenoble – LIG

UFR LLASIC
Département d'informatique intégrée en Langues, Lettres et Langage (I3L)

Mémoire de master 2 recherche en Sciences du Langage - 20 crédits

Parcours : Industries de la langue

Année universitaire 2018-2019

Remerciements

Je tiens à remercier Solange Rossato, ma directrice, pour m'avoir mis sur la voie de la prosodie à travers le stage qu'elle m'a proposé, et pour m'avoir guidé tout au long de ce travail de recherche.

Merci aussi à Romain Jourdan-Ōtsuka, qui m'a fait bénéficier des enregistrements de 29 étudiants, lesquels je remercie tous également.

Enfin, merci à tous les collègues du laboratoire GETALP pour leurs conseils et soutien.

UNIVERSITÉ Grenoble Alpes

Déclaration anti-plagiat
Document **à scanner** après signature
et **à intégrer** au mémoire électronique

DÉCLARATION

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

NOM : COULANGE

PRENOM : Sylvain

DATE : 6/09/2019

[redacted]

Sommaire

Introduction

Lorsqu'un locuteur parle dans une langue étrangère, il est souvent influencé par sa ou ses langues maternelles, et un auditeur natif de cette langue peut alors percevoir un « accent étranger ». Tout natif est capable de percevoir cet accent, et même de l'évaluer. Mais sur quels critères se base-t-on exactement pour juger du degré de cet accent ?

Si l'on peut souvent identifier les écarts de prononciation au niveau segmental, il reste une part non négligeable de l'accent due à quelque chose de moins tangible, une impression générale d'étrangeté, de distance par rapport à ce qui est communément admis par les locuteurs natifs. Dans ce mémoire de recherche, nous nous sommes intéressés à cette impression générale, en tentant de modéliser le rythme de la langue à travers un certain nombre de mesures acoustiques, et en mesurant la distance d'énoncés natifs et non-natifs par rapport à ce modèle.

Les trois premiers chapitres ont pour objectif de mettre en lumières les différents facteurs qui jouent un rôle dans la perception de l'accent étranger, et d'identifier les principaux paramètres prosodiques à l'origine de cette impression d'étrangeté. Nous nous sommes intéressés aux études de perception de l'accent, mais également aux classifications des langues par le rythme et aux techniques d'évaluation automatique de la prononciation.

Nous présentons ensuite le Corpus d'Étude pour le Français Contemporain (CEFC), sur lequel s'est basé l'apprentissage du modèle ; ainsi que les différents corpus de test utilisés, deux issus du même CEFC, et un constitué d'enregistrements d'apprenants japonophones du français.

Le 5ème chapitre détaille la méthodologie utilisée pour la modélisation du rythme. Le 6ème présente les différents tests statistiques mis en œuvre pour l'évaluer.

Les trois derniers chapitres sont consacrés à la présentation des résultats. Il s'agit dans un premier temps d'une analyse des scores obtenus par les locuteurs, puis nous nous concentrons sur les paramètres prosodiques choisis pour la modélisation, et leur efficacité à distinguer les locuteurs natifs et non-natifs. Enfin, nous observerons la corrélation qu'il peut y avoir entre le score rythmique des locuteurs japonophones et leur niveau de compétence en langue.

Première partie

Évaluation de la prononciation en L2

Chapitre 1

L'accent étranger

Dans ce chapitre, nous allons présenter le statut de l'oral dans l'enseignement des langues, et le rôle de la prononciation. Nous verrons rapidement sur quoi se concentrent les grandes théories d'acquisition de la prononciation en langue seconde (L2) puis nous analyserons le rôle des paramètres prosodiques dans la perception de l'accent étranger. Enfin, nous listerons les principaux facteurs interindividuels qui peuvent influencer cet accent.

1.1 La place de l'oral en didactique des langues

> « Occupez vous du sens, les sons se débrouilleront bien eux-mêmes »
>
> Lewis Carroll

L'oral n'a pas toujours eu la place qu'il a aujourd'hui dans l'enseignement des langues. Jusqu'à la fin du XIX^{ème}, l'apprentissage se faisait principalement à partir de textes issus de la littérature classique, et la traduction avait une place prépondérante sur les autres activités. Petit à petit on commence à considérer l'oral avec l'arrivée de la méthode directe et du développement de la phonétique (création de l'Association de phonétique internationale avec Paul Passy), puis surtout avec les méthodes audio-orales (MAO) dans les années 40 et les méthodes audio-visuelles (MAV) dans les années 60, où l'oral a une place prépondérante dans l'enseignement. Toutefois, il s'agissait alors essentiellement de répétitions et d'imitations, avec des exercices structuraux très cadrés et effectués en laboratoires de langues (Alazard 2013). Dans les années 70, les méthodes communicatives arrivent, et remettent la priorité sur l'écrit.

Concernant l'oral, elles proposent de nouvelles méthodes de travail, comme des jeux de rôles par exemple. L'objectif est maintenant d'utiliser la langue : on se concentre sur la compréhensibilité et l'intelligibilité[1] de l'apprenant, toujours dans un objectif purement communicatif, et la forme passe au second plan.

L'oral montre une variation importante aux niveaux intra- et interlocuteur, aux niveaux diatopique et diachronique, mais également diastratique et diaphasique (Alazard 2013, Detey 2007). Comment appréhender cette variabilité, et quel oral doit-on enseigner ?

Si l'on s'intéresse au Cadre Européen Commun de Référence pour les Langues (CECRL) de la Commission Européenne, qui est massivement utilisé par les institutions d'Europe et d'ailleurs, Alazard (2013) constate que la partie réservée à l'expression orale propose des descripteurs ciblant la compétence de communication, et qu'il est très rarement fait allusion à la qualité de prononciation. On se retrouve alors avec des grilles d'évaluation d'examens internationaux comme celles du DELF-DALF, n'attribuant que 3 points sur 25 à la maîtrise du système phonologique en production orale pour tous les niveaux. Dans les descripteurs de l'aisance à l'oral, on parle de « s'exprimer avec une certaine aisance » au niveau B1.2, et de « débit assez régulier » à partir du B2. Dans les niveaux C on trouve les mentions de s'exprimer « sans effort » et « avec aisance et spontanéité » (p. 100).

Le fait de ne pas accorder plus d'importance à la prononciation lors de l'évaluation est-il réellement un problème ? L'ennui, c'est de ne pas être capable de clairement diagnostiquer les difficultés de l'apprenant au niveau de sa prononciation, et cela parce qu'elles ne sont pratiquement pas abordées dans les référentiels actuels. Des apprenants de niveau avancé peuvent encore avoir d'importantes lacunes phonologiques, et ne pas avoir la possibilité de réellement les travailler par manque de diagnostic et de méthodes adaptées d'enseignement ou de remédiation.

Derwing et Munro (2015) définissent l'*accent* au sens général par l'ensemble des aspects de la prononciation qui distinguent les membres de différentes communautés de parole. Ces variations font alors souvent ressortir les différences régionales, ethniques ou de classes sociales. Lorsqu'il y a interférences entre le système phonologique de deux langues, notamment pour les locuteurs non-natifs, peuvent alors émerger des accents dits étrangers. Ces accents sont souvent expliqués comme le résultat de l'influence seule de la langue première sur la L2 (Boula de Mareüil et Vieru-Dimulescu 2006), mais nous verrons qu'ils sont influencés par de nombreux facteurs. Alazard (2013) décrit l'accent étranger comme « l'écart de prononciation

1. D'après le glossaire de Derwing et Munro (2015), la compréhensibilité renvoie au degré d'effort que doit fournir un auditeur pour comprendre un énoncé ; tandis que l'intelligibilité correspond à la mesure du degré de compréhension par l'auditeur.

commis par les apprenants vis-à-vis de la norme de prononciation attendue et partagée par les natifs d'une langue donnée » (p. 35). En tant que locuteur natif nous restons capables, plus ou moins inconsciemment, d'évaluer le degré d'accent d'un locuteur. Alazard ajoute que l'accent n'est pas une valeur "palpable" mais plutôt une « impression d'authenticité », évaluée perceptivement par des natifs (p. 40).

Nous avons vu que, depuis l'avènement des méthodes communicatives, l'intelligibilité de l'apprenant est devenu le principal objectif de l'enseignement de la prononciation. Abercrombie disait déjà à la fin des années 40 que la plupart des apprenants de langue *« need no more than a comfortably intelligible pronunciation »*[2] (Abercrombie 1949, p. 120), et beaucoup de chercheurs sont allés dans ce sens par la suite.

Pourtant, il y a de nombreuses situations où le locuteur avec un fort accent étranger est pénalisé, voire discriminé. Le milieu professionnel étant certainement le plus difficile, notamment lorsqu'il s'agit de passer un entretien d'embauche ou convaincre une assemblée. Plusieurs recherches (comme De Meo et al. 2012, Pettorino, De Meo et al. 2012, De Meo 2012) montrent que l'accent étranger peut fortement impacter la crédibilité du message. On a donc à l'autre extrémité des chercheurs considérant que l'accent est une chose à éradiquer telle une pathologie, comme Griffen pour qui *« the goal of instruction in pronunciation is that the student (or patient) should learn to speak the language as naturally as possible, free of any indication that the speaker is not a clinically normal native »*[3] (Griffen 1991, p. 182).

Bien que la priorité puisse être donnée à l'intelligibilité, il reste important de pouvoir fournir un enseignement de la prononciation de qualité lorsque celui-ci est demandé ou jugé nécessaire. Pour cela, il est important de comprendre quels sont les paramètres acoustiques pour lesquels on observe un écart par rapport aux locuteurs natifs, et savoir à quoi ces écarts sont dus, mais également connaître les facteurs interindividuels qui peuvent avoir un impact sur la prononciation.

1.2 Les grands modèles linguistiques en acquisition de la L2

L'accent étranger est donc un écart de prononciation par rapport à une norme commune aux natifs de cette langue. Mais sur quels paramètres acoustiques nous

2. n'ont besoin que d'un niveau confortable de compréhensibilité de prononciation (notre traduction).

3. L'objectif de l'enseignement de la prononciation est que l'étudiant (ou patient) apprenne à parler la langue aussi naturellement que possible, sans la moindre indication qu'il ou elle n'est pas un natif cliniquement normal (notre traduction).

basons-nous réellement pour évaluer cet écart ?

Jusque dans les années 2000, la plupart des études sur la perception de la parole non-native se sont focalisées sur les aspects segmentaux (Pellegrino 2012, Boula de Mareüil et Vieru-Dimulescu 2006, Piske et al. 2001). Les grandes théories et les modèles linguistiques en acquisition se focalisent souvent sur des phénomènes phonétiques, comme c'est le cas par exemple des catégories perceptuelles de phonèmes de Flege.

Le *Speech Learning Model* présenté par Flege (1995), développe la notion de similarité phonétique et prétend que tout apprenant peut apprendre à percevoir les propriétés phonétiques de n'importe quel phonème si l'input est suffisant et adéquat, et ce quelque soit son âge. Il propose l'idée qu'il existe un espace phonétique commun entre la L1 et la L2, et que leur système phonologique s'influencent mutuellement. Plus un phonème de la L2 est loin du phonème le plus proche correspondant dans la L1, plus il sera facile pour l'apprenant de le percevoir (phénomène de dissimilation). À l'inverse, si deux catégories de phonèmes se chevauchent, elles auront tendance à s'assimiler (Flege 2003).

C'est également l'idée que défend le modèle d'assimilation perceptuelle de Best *(Perceptual Assimilation Model)*, pour qui la précision d'acquisition du nouveau phonème de la L2 dépend de si il est assimilé à un phonème de la L1, et si oui, comment. Selon ce modèle, si deux phonèmes de la L2 sont assimilés au même phonème de la L1, l'apprenant les discriminera mieux si l'un partage plus de traits phonétiques avec le phonème d'assimilation (Best et al. 2001).

Cependant, on peut penser que tous les membres d'une même catégorie ne sont pas perçus de la même manière. C'est l'argument principal de la théorie des aimant perceptifs de Kuhl *(Native Language Magnet)*. Elle défend l'idée que la perception des propriétés acoustiques des phonèmes est définie tôt dans le développement, influencée par la langue maternelle (Kuhl 2000). Kuhl propose la notion de prototype, qui serait un son représentatif de sa catégorie. Les sons proches de ce prototype seront alors plus difficiles à discriminer que les sons éloignés, même au sein d'une même catégorie.

On retrouve dans ces modèles l'idée du crible phonologique de Trubetzkoy (1939) que constitue la L1 lors de l'apprentissage de la L2. Le système phonologique de la langue maternelle filtre et déforme la perception – et donc la production – des sons de la L2, entraînant certaines assimilations de catégories phonémiques. Mais les éléments segmentaux ne sont pas les seuls à influencer la perception/production de la L2. La prosodie de la langue maternelle peut également jouer le rôle de crible prosodique et influencer la perception de l'accent étranger (Boula de Mareüil et Vieru-Dimulescu 2006), et on commence aujourd'hui à reconnaître son importance dans la perception de la parole non-native (Piske et al. 2001).

1.3 Rôles des paramètres prosodiques

> La prosodie est l'ossature mais également le fondement et le développement de la compétence orale
>
> Guimbretière (2012)

La prosodie peut se découper en quatre constituants : le débit de parole, la segmentation de la chaîne sonore, l'accentuation et la mélodie (Guimbretière 2012). On peut mesurer ces constituants avec l'intonation, l'intensité ou le nombre et la durée des segments. Les segments peuvent être des phonèmes ou des syllabes par exemple. Quelle importance a chacun de ces paramètres dans la perception de l'accent étranger ?

Pellegrino (2012) cherche spécifiquement à déterminer quels paramètres impactent le plus la perception de l'accent étranger. Elle propose à 56 italophones natifs d'évaluer le degré d'accent d'énoncés de 8 locuteurs sinophones italianisants et de 2 natifs de l'italien. La tâche d'énonciation consiste en la lecture d'un article de magazine d'une cinquantaine de mots. Les 2 locuteurs natifs sont issus de la même région que les 56 auditeurs, et les apprenants ont tous étudié l'italien pendant 3 ans en Chine jusqu'au niveau B1, puis sont arrivés à Naples où ils ont suivi une formation intensive en compréhension et expression orale. Aucun auditeur ne parle mandarin, et personne n'a d'expérience particulière avec l'accent que peuvent avoir les sinophones en italien. Les participants sont amenés à évaluer le degré d'accent sur une échelle de Likert à 4 niveaux, et un certain nombre de mesures acoustiques sont faites puis comparées en fonction de l'accent perçu.

Après avoir découpé le corpus en unités entre pauses (UEP), les auteurs calculent pour chaque segment le débit de parole (ratio du nombre de syllabes avec la durée totale de parole), le débit d'articulation (idem avec la parole sans les pauses), le registre intonatif (minimum et maximum de la fréquence fondamentale), le pourcentage de parole sans pauses, la nombre de silences et de disfluences, et la durée moyenne des pauses silencieuses. Pellegrino mesure également la durée des syllabes en fonction de leur accentuation et de leur structure syllabique, ainsi que les erreurs de prononciation.

L'autrice constate que les natifs ont des mesures systématiquement plus élevées que les apprenants pour les paramètres suprasegmentaux. Pour les énoncés jugés non-natifs, les débits d'articulation et de parole restent relativement stables quelque soit le degré d'accent perçu (entre 4 et 4,6 syllabes par seconde pour le débit d'articulation, et entre 3,3 et 4 pour le débit de parole). La fluence (nombre de syllabe par UEP)

Degré d'accent	Débit art. (syl/s)	Débit par. (syl/s)	Fluence (syl/UEP)	Var. F_0 (dt)
Natif	6.2	5.1	13.8	9.5
Moyen	4.6	4	13.2	8.2
Fort	4.1	3.6	10.7	7.7
Très fort	4	3.3	8.7	5.8

TAB. 1.1: *Valeurs moyennes des paramètres supra-segmentaux en fonction des degrés d'accent (PELLEGRINO 2012, p. 3, UEP : unité entre pauses, dt : demi-ton)*

et le registre intonatif varient de manière plus importante : entre 8,7 et 13,2 syllabes par UEP, et entre 5,8 et 8,2 demi-tons. Le tableau 1.1 présente le détail des résultats obtenus.

Le nombre moyen de pauses varie également en fonction du degré d'accent : 8 pour un degré "moyen", 10 puis 12 pour le degré "très fort". Toutes les pauses font en moyenne 300 ms. Pellegrino explique cette variation du nombre de pauses par différentes stratégies de lecture chez les apprenants. Nous ne connaissons malheureusement pas le nombre de pauses pour les 2 locuteurs natifs. Comme les pauses, le nombre de disfluences diminue avec le degré d'accent. Du côté segmental, les énoncés à fort accent étranger se sont révélés avoir des syllabes accentuées plus longues et plus d'erreurs de prononciation.

VITALE et al. (2014) constatent également que les énoncés prononcés par des non-natifs ont un registre intonatif plus faible que dans le cas d'énoncés prononcés par des natifs. Ils remarquent également que les débits de parole et d'articulation restent généralement inférieurs à ceux des natifs, mais augmentent avec le niveau des apprenants.

Ils réunissent 12 apprenants, 4 de chacun des niveaux débutant, intermédiaire et avancé, et 4 italophones natifs pour le groupe contrôle. Les locuteurs doivent lire en binôme de courts dialogues contenant des questions fermées et des affirmations. Les paramètres mesurés dans l'étude sont la F_0 moyenne, le nombre et la durée des UEP, le nombre de syllabes par UEP, la durée des silences et des disfluences, les débits d'articulation et de parole, le registre intonatif, le pourcentage de disfluences et la F_0 des 3 dernières syllabes de chaque question. Les auteurs remarquent que le registre intonatif des apprenants est systématiquement inférieur à celui des natifs (4 demi-tons en moyenne, sauf pour les apprenants débutants à cause de voyelles parfois prononcées en *creaky voice*, faisant chuter l'intonation). Les différences relatives au débit d'articulation et de parole sont peu élevées (en moyenne 4,7 syllabes par seconde pour le débit d'articulation des non-natifs et 5,4 pour celui des natifs ; et 4,6 contre 5,4 pour le débit de parole), mais on remarque que plus le niveau de compétence augmente, plus les débits augmentent et se rapprochent de ceux des natifs (3,8 à 5,5 pour le premier ; 3,3 à 5,5 pour le second). Presqu'aucune disfluence n'est observée :

3% seulement pour les questions prononcées par les 4 apprenants débutants, pour les autres le pourcentage est inférieur à 0,2%. Les auteurs considèrent comme étant une disfluence les répétitions et les autocorrections dues aux difficultés de lecture des petits niveaux.

Vitale et son équipe proposent alors de modifier artificiellement les énoncés en remplaçant la durées des phonèmes et la courbe intonative des natifs avec celles des non-natifs et vice versa. Ils soumettent alors les nouveaux énoncés à un groupe de 40 auditeurs natifs en leur demandant d'évaluer le degré d'accent sur une échelle de 0 à 5, et de déterminer si l'énoncé est une question ou une affirmation. Parmi eux, 34 auditeurs sont experts dans le domaine des langues et/ou en phonétique. L'accent perçu des énoncés non-natifs avec la prosodie des natifs est alors beaucoup plus faible que celui des énoncés natifs avec la prosodie des non-natifs (1,4 à 1,8 contre 1,8 à 3,6 en fonction des niveaux). L'étude confirme donc un impact important de la prosodie sur la perception de l'accent étranger.

Cette expérience de transplantation d'informations prosodiques entre natifs et non-natifs a été réalisée dans de nombreuses études. On pourra citer en premier lieu Yoon (2007) qui propose un outil permettant de remplacer l'intonation, la durée des phonèmes et l'intensité d'un énoncé avec un autre. L'objectif étant d'améliorer les *feedbacks* dans les logiciels d'apprentissage de la prononciation, en renvoyant à l'utilisateur son enregistrement avec la prosodie d'un natif. L'outil est notamment repris par Rognoni et Busà (2014), Pettorino, De Meo et al. (2012) ou De Meo et al. (2012).

De Meo et al. (2012) proposent à 12 locuteurs mandarinophones italianisants, de niveau B1 à C2, et 4 italophones natifs de s'enregistrer en binôme sur un dialogue d'environ une minute. Ils isolent ensuite chaque énoncé et les soumettent à 50 experts de didactique de l'italien langue étrangère, natifs de la langue, en leur demandant d'évaluer le degré perçu d'accent étranger (fort, faible, natif), l'efficacité communicative (nulle, suffisante, bonne), et d'indiquer quel paramètre était le plus pertinent pour prendre cette décision (qualité articulatoire, intonation, débit et silences).

La figure 1.1 présente le paramètre choisi par chaque auditeur pour chaque locuteur, comme étant le plus pertinent pour juger l'efficacité communicative de l'énoncé. On s'aperçoit que l'intonation arrive systématiquement en tête (choisie par 40 à 50% des auditeurs), suivie par la qualité articulatoire (30 à 40%), le débit de parole (10 à 30%) et enfin les silences (moins de 10%). Parmi ces quatre facteurs, l'intonation serait donc la plus corrélée avec l'efficacité communicative. La figure 1.2 montre quant à elle la relation entre le degré d'accent étranger perçu et l'efficacité communicative. Le rapport semble clair ici, plus le degré d'accent est élevé, moins l'efficacité est bonne. Deux locuteurs (CIN11 et 12) se situent "à mi-chemin" entre les locuteurs natifs et les autres non-natifs, ils sont tous les deux de niveau C2 et ont passé plus de 10 ans

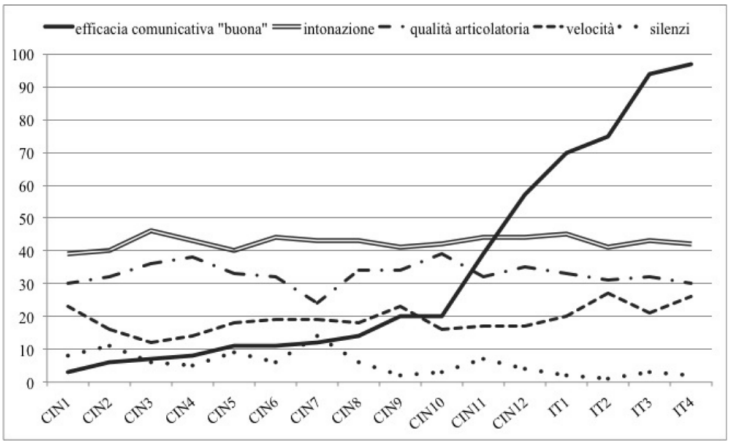


Fig. 1.1: *Paramètres jugés pertinents pour l'évaluation de l'efficacité communicative (en %) (De Meo et al. 2012, p. 121)*

en Italie.

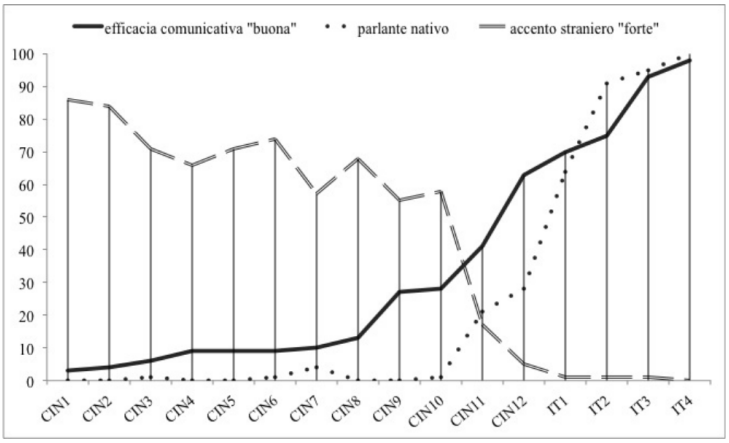


Fig. 1.2: *Relation entre le degré d'accent étranger et l'efficacité communicative (De Meo et al. 2012, p. 122)*

Afin de s'assurer que l'intonation a bien l'importance que lui donne les auditeurs, l'équipe de chercheurs italiens va conduire une deuxième expérimentation, dans laquelle ils interchangent l'intonation des deux natifs dont l'efficacité communicative est jugée la meilleure (IT3 et 4) avec celle des deux non-natifs dont le score est le plus bas (CIN1 et 2). Aucun paramètre n'est modifié par ailleurs. Ils effectuent ensuite le même test perceptif que précédemment et obtiennent les résultats présentés dans la figure 1.3. On constate que les énoncés des natifs ne sont jugés natifs plus qu'à 44% (comparé à 96% avant la transplantation), et les énoncés des non-natifs sont maintenant jugés à 51% d'accent fort (contre 74% initialement).

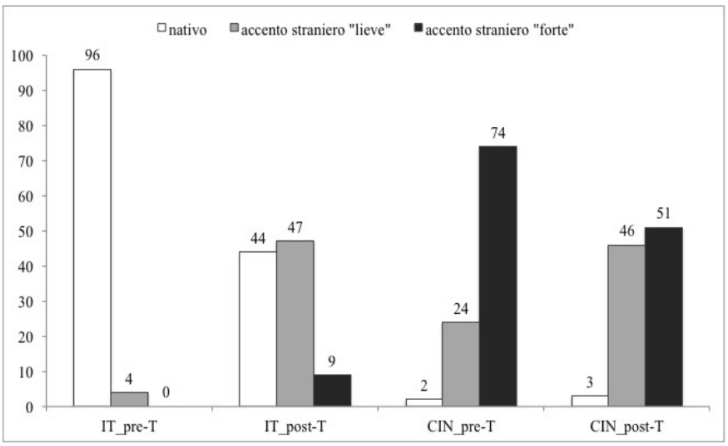


Fig. *1.3:* *Perception du degré d'accent avant et après transplantation prosodique (De Meo et al. 2012, p. 123)*

Les auteurs observeront la même tendance pour l'efficacité communicative, et concluront qu'une réduction de l'accent étranger permet une amélioration de la compétence communicative.

Dans une autre étude (Pettorino, De Meo et al. 2012), la même équipe de chercheurs monte une expérience de transplantation d'informations prosodiques pour étudier le lien entre accent étranger et crédibilité du message. Ils essaient alors une transplantation des paramètres suivants : la durée des phonèmes, l'intonation et les pauses. Le corpus est constitué de 18 enregistrements de faits divers *(bizarre-but-true)* en italien, prononcés par 4 non-natifs (A2-B1) de langues maternelles différentes et un italophone natif. Parmi ces enregistrements, 4 sont prononcés par le natif et les autres par les apprenants. 10 de ces derniers sont modifiés artificiellement : suppression des disfluences et transplantation des pauses des natifs (4), transplantation de l'intonation et des durées de voyelles (4) et suppression des erreurs segmentales pour les 2 enregistrements A2. Les 18 enregistrements sont soumis à 265 auditeurs natifs, qui doivent évaluer la compréhensibilité, le degré d'accent perçu, et la crédibilité du message. Les auteurs parviennent à la conclusion que plus la compréhensibilité du message est altérée, moins la crédibilité est élevée.

La suppression des disfluences et la transplantation des pauses réduit de 20% la perception de l'accent « fort » et augmente de 5% la perception de l'accent « natif ». Elle améliore également de 26% la crédibilité du message pour arriver pratiquement au niveau des natifs. La transplantation des durées de phonèmes et de l'intonation a plus d'impact encore : -60% pour l'accent "fort" et +30% pour l'accent « natif », et +16% pour la compréhensibilité « bonne ». Le niveau de crédibilité du message quant à lui se voit amélioré de 31%, à peine plus que dans le cas de la suppression des

disfluences et la transplantation des pauses.

La durée des phonèmes et l'intonation semble donc avoir un impact considérable dans la perception de l'accent étranger. Dans une étude qui traite séparément la durée des segments et l'intonation (Rognoni et Busà 2014), il semble que les deux paramètres influencent de manière égale la perception de l'accent étranger, mais que la combinaison des deux est toujours plus efficace. Notons que dans cette étude, l'amélioration due à la transposition de la prosodie native (intonation et/ou durées de segments) sur les segments non-natifs est assez faible comparé à celle du pattern inverse (segment natif et prosodie non-native, *cf.* figure 1.4).

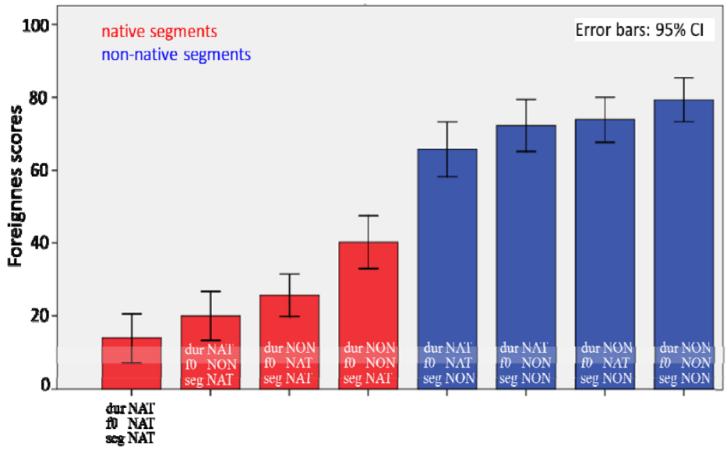


Fig. 1.4: *Degré d'accent étranger selon les critères de transplantation (Rognoni et Busà 2014, p. 556)*

Les mesures acoustiques comme les expériences de transposition et de modification artificielle du signal montrent donc que l'intonation et la durée des segments ont un impact important sur la perception de l'accent étranger, parallèlement aux informations segmentales. Au niveau prosodique, l'accent étranger se caractérise de manière générale par un registre intonatif réduit, moins de syllabes par UEP, et un débit de parole plus lent. Transposer l'intonation et la durée des segments des natifs sur de la parole non-native a tendance à réduire la perception de l'accent étranger (De Meo et al. 2012), et à l'inverse, la parole native avec l'intonation et la durée de segments non-natifs est jugée moins compréhensible et avec un accent étranger important (Pettorino, De Meo et al. 2012). Il semblerait toutefois que, pris individuellement, ces deux paramètres prosodiques n'aient pas tellement d'impact sur le degré d'accent perçu (Rognoni et Busà 2014). Il s'est avéré également que le rôle des disfluences et des pauses n'est pas aussi important que l'intonation et la durée des segments (Pettorino, De Meo et al. 2012). On peut dire en conclusion que l'accent étranger se matérialise à travers l'ensemble des paramètres prosodiques, avec une

prépondérance de l'intonation et de la durée des segments.

1.4 Facteurs influençant l'accent étranger

Dans cette section, nous tentons de répertorier les différents facteurs interindividuels pouvant impacter la prononciation de la L2. L'identification de ces facteurs peut se révéler utile pour éviter au maximum, ou au moins être conscient des biais sociologiques que peuvent comporter les études sur l'accent ou le rythme.

L'âge de première exposition à la langue semble être l'un des facteurs dont l'impact est le plus important (Flege, Frieda et al. 1997 ; Flege, Munro et al. 1995 ; Piske et al. 2001). On a longtemps parlé d'une période critique à partir de laquelle il deviendrait presque impossible d'apprendre à parler une langue sans accent (Lenneberg 1967), à cause d'une perte de plasticité cérébrale avec l'âge. Il y aurait même plusieurs périodes affectant différents niveaux de la langues (Fathman 1975, Seliger 1978, Walsh et Diller 1981, Long 1990, Hurford 1991, entre autres, cités par Piske et al. 2001). Mais cette théorie est largement remise en question car elle n'a jamais réellement été démontrée (Flege 2005). L'âge de première exposition n'en demeure pas moins un facteur important.

La durée de séjour dans un environnement où la L2 est prédominante peut également être un facteur à prendre en compte. Toutes les études ne montrent pas une corrélation entre le degré d'accent et la durée de séjour (Oyama 1976, Flege 1988 ou encore Piske et al. 2001), les résultats sont même assez disparates, preuve que ce facteur est très influencé par d'autres. Flege (1988) propose la théorie selon laquelle la durée de séjour n'a pas d'impact sur l'accent lorsque le locuteur a appris la L2 de manière intensive et sur une courte période, avec une première exposition à l'âge adulte. Piske et al. (2001) ajoutent que les locuteurs de niveau déjà avancé ne sont pas nécessairement meilleurs en prononciation s'ils ont une durée de séjour plus longue : à partir d'une certaine durée, il n'y a plus nécessairement d'amélioration de l'accent due à la durée de séjour.

Il peut sembler évident qu'un locuteur qui utilise quotidiennement sa L2 aura plus de chance d'améliorer sa prononciation qu'un locuteur qui ne l'utilise qu'occasionnellement. Dans Flege, Munro et al. (1995), les facteurs d'utilisation de la L2 expliquent 15% de la variance, après 59% expliqués par l'âge de première exposition ; des immigrants italiens au Canada devaient estimer leur pourcentage d'utilisation de l'anglais au travail, dans la vie quotidienne et à la maison. Piske et al. (2001) montrent aussi que les immigrants italiens au Canada qui continuent à parler fréquemment leur langue maternelle (et donc moins l'anglais) ont un accent généralement plus fort en anglais que ceux qui parlent moins souvent leur langue maternelle – et ce indépen-

damment de l'âge de première exposition.

La durée d'apprentissage formel de la L2 ne semble pas être très corrélée avec le degré d'accent perçu, d'après la plupart des études qui ont pris en compte ce facteur (Flege, Munro et al. 1995 ; Flege, Yeni-Komshian et al. 1999 ; Piske et al. (2001) cite aussi Thompson 1991, Elliott 1995). On ne mesure une différence significative que pour les locuteurs qui ont eu un entraînement spécifique à la phonologie de la L2, par rapport à ceux qui n'en n'ont pas eu (Missaglia 1999).

Qu'en est-il de la capacité du locuteur à apprendre une langue étrangère ? Flege, Yeni-Komshian et al. (1999) demandent à leurs locuteurs d'autoévaluer leur *"sound processing ability"*, c'est à dire leur capacité à imiter des sons, à retenir comment est prononcé un mot en anglais et leur compétence musicale. Ce facteur se révèle expliquer 2% de la variance du degré d'accent perçu.

Peu d'études semblent s'intéresser au genre comme facteur impactant l'accent étranger. Dans celles qui montrent une influence, c'est généralement les locutrices qui ont un moins d'accent : Flege, Munro et al. (1995) montrent que parmi les locuteurs qui ont été exposés tôt à la L2, le degré d'accent est perçu moins important chez les locutrices que chez les locuteurs ; c'est en revanche le cas inverse pour les locuteurs exposés tardivement à la L2. Dans l'étude de Piske et al. (2001), la moyenne du degré d'accent perçu pour les locutrices et les locuteurs ne varie pas de manière significative, et le facteur ne semble pas non plus corrélé avec les autres facteurs, qui sont l'âge de première exposition à la L2, la durée de séjour en immersion et le pourcentage d'utilisation de la langue.

On constate que tous ces paramètres sont très intercorrélés. Piske et al. (2001) montrent que la durée de séjour et l'autoévaluation de la compétence en L1 expliquent une part significative de la variance que si l'on considère également l'âge d'exposition. Tous les paramètres ne sont donc pas nécessairement indépendants, mais cela ne veut pas dire qu'ils ne sont pas pertinents pour autant : ils peuvent toujours être à l'origine d'une partie de la variance du degré d'accent perçu. Piske et al. (2001) considèrent comme paramètres indépendants dans leur étude l'âge de première exposition et le pourcentage d'utilisation de la langue maternelle et de la L2 ; le genre, la durée de résidence dans le pays de la langue cible et l'autoévaluation de la compétence en langue étant des facteurs secondaires.

Gardons à l'esprit qu'un certain nombre de facteurs côté auditeur peuvent également influencer la perception de l'accent étranger. Schoonmaker-Gates (2012) cite notamment la langue maternelle de l'auditeur, l'expérience que celui-ci peut avoir avec l'accent en question, s'il connaît ou non la langue maternelle du locuteur ou encore la durée passée dans un environnement où elle est parlée. Lorsqu'une étude perceptive est menée, il est donc nécessaire de documenter les principales caractéris-

tiques des auditeurs, et faire en sorte de minimiser les biais possibles. On constate par exemple que De Meo et al. (2012) choisissent des auditeurs tous expérimentés en didactique de l'italien langue étrangère ; ils ont donc en principe une oreille plus sensible et/ou habituée aux variations non-natives. Pellegrino (2012) quant à elle fait en sorte de choisir des auditeurs qui ne parlent pas la langue maternelle des locuteurs, et qui n'ont pas d'expérience particulière avec l'accent étranger des mandarinophones. Schoonmaker-Gates (2012) est une des rares études sur la perception de l'accent étranger par des auditeurs non-natifs de la langue cible. Elle remarque que le niveau de compétences grammaticales de ses auditeurs influe sur leur perception de l'accent. Les auditeurs de niveau avancé arrivent mieux à identifier l'accent étranger que les auditeurs de niveau plus faible, certainement parce que leur aisance dans la L2 leur permet de moins avoir à se focaliser sur le sens, et plus apprécier la forme. Ses travaux vont dans le sens de Flege (1988), pour qui les auditeurs mandarinophones ayant vécu 5 ans aux États-Unis distinguaient mieux l'accent que ceux n'ayant vécu qu'une année.

1.5 Conclusion du chapitre

Lorsque l'on travaille la prononciation en classe de langue, la focale est souvent mise sur la phonétique et peu sur la prosodie. Pourtant, les techniques de synthèse vocale et de modification artificielle du signal de parole ont permis d'étudier et de confirmer l'impact de la prosodie sur la perception de l'accent étranger.

Grâce à un certain nombre d'études présentées dans la section 3, nous avons pu mettre en avant deux paramètres prosodiques qui ont une grande influence sur la perception de l'accent : l'intonation et la durée des segments. Les autres paramètres, comme le nombre de pauses ou leur durée ont également un impact, mais moins important. L'intonation et la durée des segments sont des caractéristiques qui peuvent découler de la structure prosodique de la langue maternelle. En effet, les variations rythmiques qui peuvent exister entre les langues ont fait l'objet de nombreuses études que nous allons maintenant présenter.

Chapitre 2

Langues & rythmes

Ce chapitre se concentre sur la variation rythmique qu'il peut exister entre les langues. En effet, depuis les années 50, de nombreuses études se sont essayées à faire une typologie des langues en fonction de leurs caractéristiques rythmiques, c'est-à-dire la plupart du temps en fonction des variations de durées des segments qui la constituent. Nous présenterons les grandes théories de classifications, ainsi que les principales métriques utilisées pour quantifier cette variation. Enfin nous alerterons le lecteur sur la variabilité rythmique qui peut exister au sein d'une même langue et d'un même locuteur.

2.1 Classifications rythmiques des langues

Gibbon et Gut (2001) définissent le rythme comme une récurrence de patterns de marquages forts ou faibles d'éléments dans un environnement temporel. Ces éléments peuvent être des alternances de syllabes longues et courtes, mais aussi de hauteur ou de segments vocaliques et consonantiques. Di Cristo et Hirst (1997) le définissent comme « l'organisation temporelle des proéminences ». Arvaniti (2009) part de définitions du rythme en psychologie : « perception de patterns de similarités et de répétitions dans des séries de stimuli ». Alazard (2013) fait référence au rythme cardiaque ou respiratoire, au rythme en danse et en musique, ou encore pour la marche ou l'écriture, en considérant le rythme verbal comme un phénomène physique et psychobiologique. Toutes ces définitions font ressortir l'idée de pattern et de temps.

Les premières études sur le rythme de la parole proposent l'hypothèse selon laquelle les langues varient en fonction de leurs caractéristiques rythmiques (James

1929 et Pike 1945 cités par Dellwo et Fourcin 2013). La durée des syllabes de certaines langues comme l'arabe ou l'anglais étaient alors considérées particulièrement irrégulières : *« similar to the irregular timing pattern of Morse code »*, et ce contrairement à d'autres, comme le yoruba ou le français, considérées comme plus régulières : *« similar to the regularity of bullet sounds from a machine-gun »* (Dellwo et Fourcin 2013, p. 1, paraphrasant James 1929).

La théorie dite de l'isochronie se développe. Il s'agit d'une notion déjà introduite par Steele en 1779, qui constituera le fondement de la plupart des travaux sur le rythme à partir des années 1950, car ils seront basés sur une catégorisation des langues dont l'isochronie est à l'origine. Selon cette théorie, la chaîne parlée d'une langue peut être découpée en unités de durées à peu près équivalentes (Tortel 2009). En fonction des langues on trouve différents patterns de segmentation et de durées, et il est alors possible de classifier les langues selon ces caractéristiques. Cette théorie s'est développée et a été étudiée principalement sur des langues occidentales, et notamment le français et l'anglais, qui ont des caractéristiques rythmiques bien distinctes. Pike (1945) et Abercrombie (1967) remarquent que l'anglais a tendance a avoir des pieds[1] de longueur égale, quelque soit leur nombre de syllabes ; tandis que le français est plutôt constitué de syllabes se produisant à intervalles réguliers (Tortel 2009). Ils définissent alors la classe d'isochronie accentuelle, ou isoaccentuelle *(stress-timed)*, pour les langues ayant des caractéristiques similaires à celles de l'anglais, et la classe d'isochronie syllabique, ou isosyllabique *(syllable-timed)*, pour les langues rythmiquement plus proches du français.

Les langues isoaccentuelles Les intervalles entre les syllabes accentuées sont réguliers. Les syllabes accentuées sont plus longues que celles qui ne le sont pas, et la durée du pied est toujours plus ou moins la même quelque soit le nombre de syllabes, ce qui implique un raccourcissement des syllabes non-accentuées. Tortel (2009) parle de compression syllabique. Cette catégorie regroupe communément l'anglais, l'allemand, l'arabe, le néerlandais, le russe, le polonais ou le thaï (Ramus et al. 1999, Grabe et Low 2002, Arvaniti 2009).

Les langues isosyllabiques Dans cette classe, les syllabes sont dites revenir à intervalles réguliers, sans grande variation de durée contrairement aux langues isoaccentuelles. On y trouve principalement le français, l'espagnol, l'italien, le grec, le catalan, le yoruba ou l'hindi *(op. cit.)*.

Pour Abercrombie, toutes les langues ont un rythme et s'inscrivent donc obli-

1. Le pied est une unité rythmique qui correspond à un ensemble de syllabes dont la première est accentuée, et les suivantes ne le sont pas. Le pied suivant commence à la syllabe accentuée suivante.

gatoirement dans l'une de ces deux catégories (Tortel 2009). On parle toutefois souvent d'une troisième catégorie rythmique : les langues isomoraïques, dont font partie le japonais, l'estonien ou le télougou (Bloch 1950, Otake et al. 1993, Grabe et Low 2002). Pour ces langues, ce sont les mores qui reviennent à intervalles réguliers, avec toujours peu de variation de durée.

Les langues sont ainsi catégorisées depuis les année 50, mais les études empiriques n'ont jamais réellement montré l'évidence de cette isochronie, ni la pertinence de cette classification (Arvaniti 2009). Ce manque d'évidences mesurables mène à considérer la régularité rythmique comme un phénomène de perception et non véritablement un phénomène physique. L'isochronie subjective est d'ailleurs étudiée par de nombreux chercheurs comme Fraisse dès 1956 pour qui *« toute rythmicité est subjective »* (Fraisse 1956, p. 9 cité par Tortel 2009, p. 38).

Cette classification dichotomique des langues est remise en question (Arvaniti 2009). Wenk et Wioland (1982) montrent par exemple que la durée moyenne des syllabes du français n'est pas constante, contrairement à ce qu'il est communément admis, et à ce qui constitue la caractérisation même des langues isosyllabiques. Ils montrent aussi que l'allongement de la syllabe en finale des groupes rythmiques en français n'est pas, et ne peut pas être pris en compte dans la théorie de l'isochronie syllabique. Ils proposent une autre classification, centrée sur l'accent : les langues codachrones *(trailer-timed languages)*, comme le français, ont une syllabe accentuée en finale du groupe rythmique ; tandis que les langues capochrones *(leader-timed languages)*, comme l'anglais, ont leur syllabe accentuée en tête du groupe. On retrouve la même dichotomie anglais/français, mais cette fois au niveau de la position de l'accent dans le groupe rythmique.

Dauer (1987) s'oppose à une catégorisation absolue des langues, et propose une liste de 8 critères binaires pour placer les langues sur un continuum de rythmicité. Les critères sont par exemple la présence ou l'absence de différence de durée entre les syllabes accentuées et non accentuées, ou encore si la hauteur a une fonction lexicale dans la langue. Dauer part du principe que toutes les langues sont plus ou moins accentuelles, et son continuum permet de les positionner en fonction de l'importance de leur accent. Sur ce continuum également, Arvaniti (2009) fait remarquer que l'anglais et le français semblent s'opposer à travers plusieurs paramètres. Mais cette opposition est-elle due au fait que les deux langues sont réellement opposées rythmiquement, et que l'ensemble des autres langues se situerait plus ou moins entre les deux ; ou bien est-elle due au fait que Dauer est partie de la comparaison entre ces deux langues pour rassembler les paramètres qui lui semblent pertinents pour caractériser le rythme de la langue ?

Bertinetto (1989) propose une approche scalaire assez similaire au continuum de Dauer, mais où chacun des critères est un facteur qui se voit attribuer une va-

leur, une importance, pour chaque langue. Pour lui, toutes les langues ont une base rythmique commune mais elles varient en fonction de leurs facteurs phonético-phonologiques. Parmi ces facteurs, on trouve par exemple le taux de réduction vocalique dans les syllabes inaccentuées, celui de la compression syllabique, ou de certitude dans le décompte des syllabes ou bien dans la position des frontières intersyllabiques. Certains de ces facteurs correspondent aux critères utilisés dans le continuum de Dauer.

À la fin des années 1990, on change de focus pour s'intéresser non plus aux syllabes, mais aux intervalles vocaliques et consonantiques. Ramus et al. (1999) font partie des premiers à comparer les mesures d'intervalles vocaliques et consonantiques. Ils comparent 8 langues sensées appartenir aux trois catégories rythmiques traditionnelles : isoaccentuel, isosyllabique et isomoraïque. Les mesures sont effectuées sur 5 phrases d'une vingtaine de syllabes, lues par 4 locuteurs pour chaque langue. La figure 2.1 montre les résultats obtenus, et semble différencier effectivement trois groupes de langues : d'un côté l'anglais, le polonais et le néerlandais avec un écart type de durée consonantique plus grand (ΔC, entre 5,14 et 5,35) et un pourcentage vocalique plus faible ($\%V$, entre 40,1 et 42,3), que le français, l'italien, l'espagnol et le catalan (ΔC: entre 4,39 et 4,81 ; $\%V$: entre 43,6 et 45,6). Loin de ces langues et isolé, on trouve le japonais, avec un écart type consonantique très bas (3,56) et un pourcentage vocalique très élevé (53,1) par rapport aux autres. Les auteurs confirment que l'existence des catégories rythmiques, jusque là intuitives, est prouvée par ces mesures. Toutefois, ils n'excluent pas que ces catégories soient remisent en question avec plus de langues.

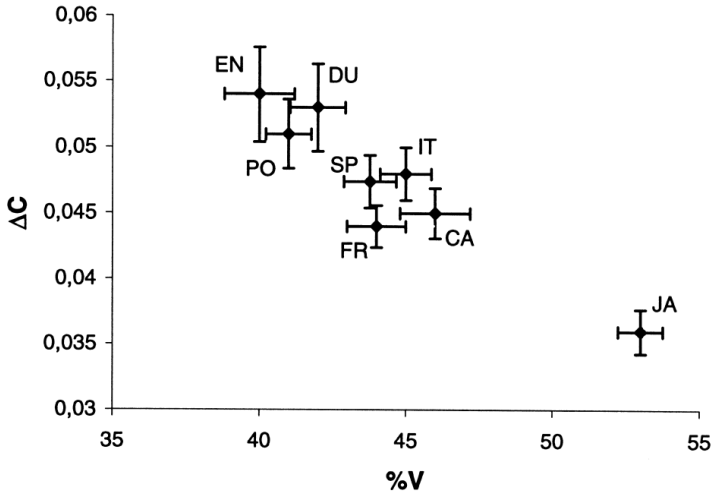


Fig. 2.1: *Répartition des 8 langues en fonction du pourcentage de vocalisation* ($\%V$) *et de l'écart type de la durée des consonnes* (ΔC) *(Ramus et al. 1999, p. 273)*

2.2 Disparité des résultats

Et c'est effectivement ce que montrent Grabe et Low (2002). Ils réitèrent les mesures sur 18 langues, sur la lecture de la fable d'Ésope *« North Wind and the Sun »* par un locuteur de chaque langue. Ils intègrent également une nouvelle métrique : la comparaison de paires successives d'intervalles vocaliques ou consonantiques ($PVI-V/C$). Ils montrent que les deux métriques classifient les langues différemment, et n'obtiennent pas les mêmes résultats que Ramus et al. (1999) avec le $\%V-\Delta C$. Le thaï est classé isosyllabique par le $\%V-\Delta C$ mais isoaccentuel par le PVI. C'est le cas inverse pour le luxembourgeois et le japonais, qui n'est plus isolé comme dans l'étude de Ramus et son équipe. Le catalan se retrouve clairement parmi les langues isoaccentuelles avec le $\%V-\Delta C$ (*cf.* figure 2.2).

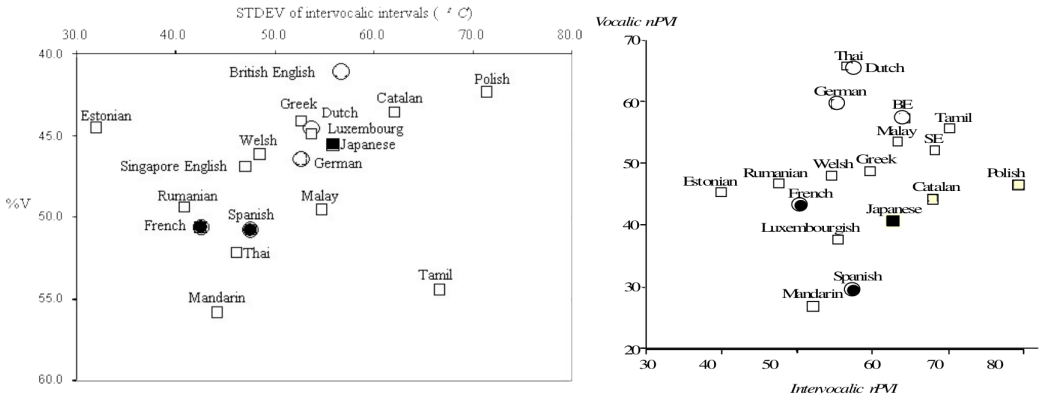


Fig. 2.2: *Résultats obtenus par Grabe et Low (2002) : $\%V-\Delta C$ à gauche et $rPVI-C$ $nPVI-V$ à droite (p. 7 et 9) (catégories traditionnellement assignées : cercle noir = isosyllabique, cercle blanc = isoaccentuel, carré noir = isomoraïque, carré blanc = mixte ou non-classé)*

Leur étude est l'une des rares qui traite un grand nombre de langues, et ils aboutissent à la conclusion que les métriques utilisées ne sont pas performantes pour classer les langues non prototypiques. L'anglais britannique, le néerlandais, l'allemand, le français et l'espagnol, toutes prototypiques, sont classifiées comme attendu isoaccentuelles pour les trois premières et isosyllabiques pour les deux dernières, néanmoins seules 4 des 13 autres langues sont classées sans ambiguïté par le PVI[2] mais sont classées différemment par le $\%V-\Delta C$; le catalan, l'estonien et le polonais sont classées comme *« mixed rhythm »* et les autres langues restent inclassables[3].

1. *Cf.* section 2.3 pour le détail des calculs.
2. Le thaï en isoaccentuelle, le japonais, le mandarin et le luxembourgeois en isosyllabique.
3. Le grec, du malais, du roumain, de l'anglais de Singapour, du tamoul et du gallois.

Arvaniti (2009) remet complètement en question la classification rythmique des langues tant qu'elle se base uniquement, et c'est le cas de la majorité des études, sur des mesures de durées. Pour elle la durée des unités est un élément important de la caractérisation du rythme, mais cela ne suffit pas pour classer les langues. Elle met en garde sur le fait que bien trop de facteurs peuvent influencer la durée des segments, comme la présence de consonnes géminées ou l'allongement vocalique, les focus lexicaux, les accents, le contexte de voisement ou de position syllabique, ou encore l'allongement de fin de groupe rythmique. Elle met en place avec Tristie Ross un protocole de comparaison de différentes métriques sur des langues de rythmiques différentes qui rentrent ou non dans les catégories isochroniques traditionnelles, et utilise un corpus de phrases lues isolées, d'une lecture d'un texte et de discours spontané (Arvaniti et Ross 2010). Elles montreront que les scores des différentes métriques sont très variables, surtout avec le discours spontané, alors que les différences interlangues sont minimes, là encore surtout pour le discours spontané.

Pour elles, les métriques existantes permettent seulement de fournir une mesure brute de la durée des segments pour un échantillon donné, mais pas de déterminer l'origine de ces fluctuations de durée. On manque d'une méthode claire d'interprétation des différences et similarités des scores. On se contente de juger les métriques en fonction des scores qu'elles obtiennent pour des langues préclassifiées ; mais on se rend compte que certains scores pour une même catégorie varient parfois plus que des scores de langues de catégories différentes.

	nPVI score differences	rPVI score differences
Stress-timed languages	$nPVI_{Th} - nPVI_{BE} = 8.6$	$rPVI_{BE} - rPVI_{Gm} = 8.8$
Syllable-timed languages	$nPVI_{Fr} - nPVI_{Mn} = 16.5$	$rPVI_{Jp} - rPVI_{Fr} = 12.1$
Languages of different types	$nPVI_{BE} - nPVI_{Fr} = 13.7$	$rPVI_{Sp} - rPVI_{Gm} = 2.4$
	$nPVI_{Gm} - nPVI_{Fr} = 16.2$	$rPVI_{Sp} - rPVI_{Th} = 1.2$

BE = British English, D = Dutch, Fr = French, Gm = German, Jp = Japanese, Mn = Mandarin, Sp = Spanish, Th = Thai.

Fig. 2.3: *Scores PVI entre des langues appartenant ou non à la même catégorie rythmique (tableau d'Arvaniti 2009, p. 56, d'après les résultats de Grabe et Low 2002)*

La figure 2.3 montre que la différence de PVI normalisé au débit de parole ($nPVI$)[4] entre des langues de même catégorie peut être supérieure à celle de langues de catégories différentes. C'est le cas par exemple du français et du mandarin, toutes deux sensées être isosyllabiques, dont la différence de nPVI est de 16,5, tandis qu'elle est de 13,7 entre le français et l'anglais britannique (qui est isoaccentuel). De plus, Grabe et Low (2002) obtiennent le japonais comme langue la plus proche rythmiquement à l'anglais en observant la distance euclidienne des scores $\%V - \Delta C$ (4.5pt), ce qui contredit tout à fait la classification traditionnelle. Arvaniti (2009) fait remar-

4. *Cf.* section 2.3.

quer qu'ils se basent sur la classification isochronique pour en conclure que ce sont les métriques ou leur corpus qui ont un problème. Cela a certainement un impact sur les résultats, mais peut-être est-il également pertinent de remettre en question cette classification qui n'est basée que sur des présupposés et des observations de langues prototypiques, à l'instar d'Arvaniti (2009).

Dans ses travaux, Arvaniti (2009) cherche des critères indépendants pour mesurer le rythme des langues, se suffisant à eux mêmes et ne nécessitant pas de comparer les scores obtenus entre les langues pour être interprétables. Arvaniti et Ross (2010) tentent de déterminer comment des auditeurs catégorisent en classes rythmiques des énoncés de différentes langues filtrés à 500 Hz, mélangés avec des pseudo-énoncés ne correspondant à aucun schéma rythmique connu. La tâche s'avère difficile pour tous les auditeurs, et les résultats ne correspondent pas à la classification rythmique traditionnelle. Les auteurs en concluent que la classification subjective n'est pas plus fiable que les métriques mesurant les variations de durée des intervalles vocaliques et consonantiques, et que la notion même de classe rythmique reste floue.

Beaucoup de facteurs inter-locuteurs et inter-situationnels influencent les résultats (Gibbon et Gut 2001, Arvaniti 2009) et les langues peuvent se retrouver classifiées différemment selon les métriques utilisées pour faire les mesures. L'objectif de ces métriques est de mesurer la variabilité de durée des intervalles (vocalique, syllabique *etc.*) en fonction des langues. Or, le rôle de ces durées nous intéresse comme élément constitutif de l'accent étranger.

2.3 Les principales métriques utilisées

Nous proposons ici de répertorier les principales métriques rencontrées dans la littérature, en détaillant le calcul lorsque c'est nécessaire.

On pourra commencer par les traditionnels pourcentage d'intervalles vocaliques dans la parole ($\%V$) et écart type de la durée d'intervalles consonantiques (ΔC) initialement introduits par Ramus et al. (1999).

$$\%V = \frac{\sum_{i=1}^{n_v} V}{d} \qquad \Delta C = \sqrt{\frac{1}{n_c}\sum_{i=1}^{n_c}(C_i - \bar{C})^2} \tag{2.1}$$

avec V et C la durée des intervalles vocaliques et consonantiques, n le nombre d'intervalles vocaliques ou consonantiques, et d la durée totale du segment de parole. Ces deux métriques consistent à rendre compte de la proportion de voyelles dans

l'énoncé, et le degré de variation de la durée des consonnes. On peut également calculer l'écart type des intervalles vocaliques (ΔV), mais le pourcentage consonantique n'a pas d'intérêt si l'on considère déjà le pourcentage vocalique.

Plus tard, Fourcin et Dellwo (2013) trouveront des résultats sensiblement similaires en considérant les intervalles voisés et non-voisés ($\%VO - \Delta UV$). Cela présente l'avantage de ne pas avoir à annoter le corpus, puisque la détection du voisement peut se faire à partir du signal brut.

L'inconvénient des précédentes métriques est qu'elles sont très dépendantes du débit de parole des énoncés analysés. Aussi, Dellwo (2006) puis White et Mattys (2007a) proposent de normaliser l'écart type de durée d'un segment par la moyenne de ses durées. Ce ratio est communément appelé *coefficient de variation*, et peut être appliqué sur les intervalles vocaliques ($VarcoV$) ou consonantiques ($VarcoC$), comme sur les intervalles de voisement ($VarcoVO$, $VarcoUV$ introduits par Fourcin et Dellwo 2013). Rossato et al. (2018) mesurent également le coefficient de variation de paires d'intervalles (voisé + non-voisé).

Une autre métrique consiste à utiliser la durée des syllabes. Pour la mesurer, on peut soit utiliser une transcription et un alignement (Ferrer et al. 2015 ; Shahin et al. 2016), soit mesurer la durée entre deux noyaux syllabiques (Jong et Wempe 2009 ; Pettorino, Maffia et al. 2013). On peut ainsi calculer la moyenne des intervalles ($VtoV$) et leur écart type ($\Delta VtoV$). L'avantage de cette dernière méthode est qu'elle nécessite pas de modèle acoustique, puisque les noyaux syllabiques peuvent être détectés automatiquement à partir du signal seulement.

Grabe et Low (2002) proposent d'analyser les intervalles par paire, en faisant la somme des différences de durée de chaque paire divisées par le nombre d'intervalles. La métrique est appelée *raw Pairwise Variability Index* ($rPVI$) et est calculée comme suit :

$$rPVI = \sum_{k=1}^{m-1} \frac{|d_k - d_{k+1}|}{m-1} \tag{2.2}$$

avec m le nombre d'intervalles et d_k la durée du k$^{\text{ième}}$ intervalle. Les auteurs proposeront également une version normalisée du PVI pour rendre la comparaison indépendante du débit de parole ($nPVI$), où la différence de durée des deux intervalles courants est divisée par leur moyenne :

$$nPVI = \left[\sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{\frac{d_k + d_{k+1}}{2}} \right| / (m-1) \right] \tag{2.3}$$

Les intervalles peuvent être vocaliques ou consonantiques, comme dans l'étude initiale de Grabe et Low (2002), voisés (Fourcin et Dellwo 2013), ou encore combiner un intervalle vocalique suivi d'un intervalle consonantique comme dans l'étude de Liss et al. (2009).

Ringeval et al. (2012) développent une variante du PVI en remplaçant les durées d'intervalles par les coefficients de variation $(pPVI)$. Gibbon et Gut (2001) en font un pourcentage plus facile à manipuler, qu'ils appellent le *Ratio Rythmique* (RR).

Le *Control/Compensation Index* (CCI), est une autre normalisation du PVI par le nombre d'éléments qui composent chaque segment. Il est introduit par Bertinetto et Bertini (2008), qui proposent de normaliser les intervalles consonantiques par le nombre de consonnes successives qu'il peut contenir. L'équation qu'il suggère devient alors :

$$CCI = \frac{100}{m-1} \sum_{k=1}^{m-1} \left| \frac{d_k}{n_k} - \frac{d_{k+1}}{n_{k+1}} \right| \tag{2.4}$$

avec n le nombre d'éléments contenus dans l'intervalle k.

Scott et al. (1986) font un autre calcul : ils proposent de considérer chaque paire d'intervalles de l'énoncé, consécutifs ou non, en faisant la somme du log de leur ratio. Le résultat est ce qu'ils appellent une *mesure d'irrégularité de paires* (PIM), qu'on nommera communément par la suite *mesure d'irrégularité rythmique* (RIM):

$$RIM = \sum_{i \neq j} \left| log \frac{d_i}{d_j} \right| \tag{2.5}$$

Les métriques qui se sont montrées jusque là les plus discriminantes et robustes aux variations de débit sont le $VarcoV$, le $\%V$ et le $nPVI-V$ (Wiget et al. 2010). Les moins contraignantes sont celles qui utilisent des segments voisés (les variantes proposées par Fourcin et Dellwo 2013), et les noyaux syllabiques (Jong et Wempe 2009 ; Pettorino, Maffia et al. 2013), puisqu'elles ne nécessitent aucune annotation du corpus. Le tableau 2.1 liste les métriques que nous avons décrites dans cette section.

2.4 Variabilité situationnelle et interindividuelle

On remarque que la majorité des études se basent uniquement sur du texte lu (entre autres Arvaniti 2012 ; Dellwo 2006 ; Fourcin et Dellwo 2013 ; Grabe et

$\%V$	Pourcentage d'intervalles vocaliques dans la parole (Ramus et al. 1999)
$\%VO$	Pourcentage d'intervalles voisés (VO) (Fourcin et Dellwo 2013)
ΔV	Écart type de la durée d'intervalles vocaliques (Ramus et al. 1999)
ΔC	Écart type de la durée d'intervalles consonantiques (Ramus et al. 1999)
ΔUV	Écart type de la durée des intervalles non-voisés (UV) (Fourcin et Dellwo 2013)
$\Delta syll$	Écart type de la durée des syllabes (Gibbon et Gut 2001)
$VarcoV$	Coefficient de variation des intervalles vocaliques (White et Mattys 2007a, Dellwo 2006)
$VarcoC$	Coefficient de variation des intervalles consonantiques (White et Mattys 2007a, Dellwo 2006)
$VarcoVO$	Coefficient de variation des intervalles voisés (Fourcin et Dellwo 2013)
$VarcoUV$	Coefficient de variation des intervalles non-voisés (Fourcin et Dellwo 2013)
$nPVI-V$	Comparaison normalisée de paires successives d'intervalles vocaliques (Grabe et Low 2002)
$rPVI-C$	Comparaison de paires successives d'intervalles consonantiques (Grabe et Low 2002)
$nPVI-VC$	Comparaison de paires successives d'intervalles vocaliques puis consonantique (Liss et al. 2009)
$nPVI-VO$	Comparaison normalisée de paires successives d'intervalles voisés (Fourcin et Dellwo 2013)
$rPVI-UV$	Comparaison de paires successives d'intervalles non-voisés (Fourcin et Dellwo 2013)
$pPVI$	Variante du PVI remplaçant les durées d'intervalles par les coefficients de variation (Ringeval et al. 2012)
CCI	*Control and Compensation Index* (Bertinetto et Bertini 2008)
RIM	Mesure d'irrégularité rythmique développée par Scott et al 1986 (cités par Gibbon et Gut 2001)
$VtoV$	Intervalle entre deux débuts de voyelles (voyel onset point) (Pettorino, Maffia et al. 2013)

Tab. 2.1: *Listes des principales métriques utilisées pour caractériser le rythme des langues*

Low 2002 ; Nazzi et al. 1998 ; Pettorino, Maffia et al. 2013 ; Piske et al. 2001 ; Ramus et al. 1999), et très rarement sur de la parole spontanée, alors qu'on sait que celle-ci donne des résultats beaucoup plus variables (Arvaniti et Ross 2010). Il ne s'agit souvent que de petits échantillons de langue (seulement 5 phrases par locuteur pour Fourcin et Dellwo 2013 ; Ramus et al. 1999 ; White et Mattys 2007b par exemple). On notera également le peu de locuteurs qui participent aux expériences (rarement plus de 6 ou 7 (parfois même un seul locuteur par langue pour Gibbon et Gut 2001 ou Grabe et Low 2002) et les biais possibles dus au sexe, à l'âge ou au profil des locuteurs peuvent biaiser les résultats. Gibbon et Gut (2001) déduisent de leur étude que le rythme de la parole n'est pas seulement déterminé par la langue mais aussi par des différences entre locuteurs. Plusieurs études ont ainsi mis en avant les différences rythmiques entre les locuteurs (Dellwo, Leeman et al. 2015 ; Rossato et al. 2018). Wiget et al. (2010) constatent que la disparité dans les résultats des études sur la classification des langues par leur rythme est due à plusieurs facteurs : les locuteurs, les protocoles d'élicitation, le matériaux linguistique élicité, les métriques ou encore les techniques de segmentation. Ils considèrent toutefois que la variation interlangue sera toujours plus importante que la différence interlocuteur.

Or le rythme de parole dépend beaucoup du style de parole, et ce style peut avoir des formes très variées selon la situation d'énonciation, le locuteur ou son intention. Le terme de *phonostyles* est déjà utilisé par Trubetzkoy à la fin des années 30, puis analysé en profondeur par Léon dans les années 90 (Léon 1993). Les phonostyles sont l'ensemble des variations phoniques dues à la situation ou à l'individu. Ils sont à la parole, ce que les genres de textes sont à l'écrit, et sont omniprésent dans la langue. Ils distinguent par exemple le commentaire sportif, du discours en public ou du langage adressé aux jeunes enfants (Fagyal et M.-A. Morel 1996). Il peut également s'agir du style utilisé par le locuteur pour apporter des informations extralinguistiques, comme l'emphase, une émotion particulière, de l'ironie ou un caractère informel à l'énoncé (Gumperz, 1982 cité par Simon et Lacheret 2016). Il peut encore s'agir d'une façon de parler propre à une communauté linguistique ou à un individu, comme l'art oratoire de Charles de Gaulle ou la voix charmeuse de Brigitte Bardot (Léon 1971 et 1981, cités par Fagyal et M.-A. Morel 1996).

Toutes ces variations de style peuvent fortement impacter les analyses du rythme. Ces variations sont également difficilement contrôlables, on peut tenter de réduire une population à quelques locuteurs d'une même communauté linguistique, avec un profil similaire, et travailler sur un genre situationnel bien délimité ; mais quelle sera la valeur de l'étude lorsqu'il s'agira de généraliser à une communauté de locuteurs plus grande ou un genre plus large ? Fagyal et M.-A. Morel (1996) insistent sur le fait qu'une grande quantité de données prélevées de façon systématique est la seule chance de réduire ces risques de variations et garantir la fiabilité des résultats.

Ajoutons pour finir que l'on peut également observer des différences rythmiques en fonction de la langue maternelle des locuteurs. White et Mattys (2007b) ont comparé le score de différentes métriques[5] sur de l'anglais lu par des natifs (sud-britanniques) et des non-natifs (néerlandais et espagnols), et les ont comparé avec le degré d'accent perçu par 12 natifs anglophones selon une méthodologie similaire à l'étude de Piske et al. (2001). Le $VarcoV$ est le plus corrélé avec le degré d'accent (p<0,01), suivi par le $\%V$ (idem) puis le $nPVI-V$ et le débit de parole (p<0,05 pour les deux). Les autres métriques n'ont pas de corrélation significative. En ce qui concerne l'anglais, la durée des voyelles peut donc varier selon que le locuteur est natif ou non. En outre, plus l'accent est jugé fort, plus les mesures sont éloignées de celles observées chez les natifs (corrélation uniquement pour les locuteurs espagnols, $r = -0,440, p < 0,05$). Les auteurs montrent également que l'écart des mesures augmente en fonction de l'éloignement des langues au niveau rythmique. Aussi, les locuteurs néerlandais ont des mesures $VarcoV$ et $\%V$ plus proches des natifs que les locuteurs espagnols (61 contre 54, par rapport à 64 pour les natifs avec le $VarcoV$, et 40 contre 41 par rapport à 38 pour les natifs avec le $\%V$).

2.5 Conclusion du chapitre

Les langues suivent des schémas rythmiques divers, et de nombreuses études ont tenté et tentent encore de les rassembler en familles rythmiques. Les différences entre les langues comment l'anglais, l'allemand ou le russe, avec d'autres comme le français, l'italien ou l'espagnol on été maintes fois démontrées, par différentes métriques ; mais lorsqu'il s'agit de classifier des langues dites moins prototypiques, la frontière entre les catégories s'estompe. On en vient alors à remettre en question la traditionnelle dichotomie isoaccentuel-isosyllabique au profit de différences sous la forme de continuums de variation (Arvaniti 2009).

De nombreuses métriques ont été imaginées pour mesurer le rythme des langues. La plupart des études se basent sur les segments vocaliques et consonantiques, mais certaines se basent sur des segments voisés ou des syllabes. Il ressort de ces études que les schémas rythmiques propres à chaque langue évoluent aussi de manière importante selon les locuteurs et les situations d'énonciation. Il est donc primordial de mener des études sur les caractéristiques rythmiques d'une langue sur de plus gros corpus, prenant en compte la variation interindividuelle et situationnelle.

5. Les métriques comparées sont ΔV, ΔC, $\%V$, $VarcoV$, $VarcoC$, $nPVI-V$, $rPVI-C$ et le débit de parole.

Chapitre 3

Systèmes d'évaluation automatique de la prononciation

Les systèmes d'apprentissage assisté par ordinateur sont nombreux dès l'arrivée des ordinateurs dans les écoles et les foyers, mais il faut attendre la fin des années 80 pour que l'on s'intéresse à la production des apprenants de langues étrangères (Mostow et al. 1993). Les systèmes d'entraînement à la prononciation assisté par ordinateur (EPAO) se distinguent en fonction de leurs objectifs, de la technique d'évaluation, et du type de données qu'ils prennent en entrée.

Comme la majorité d'entre eux se concentrent sur la détection d'erreurs segmentales (Detey et al. 2016), nous proposons d'abord de présenter certains d'entre eux, en expliquant le fonctionnement global des techniques de reconnaissance automatique de la parole (ASR) qu'ils utilisent, puis nous aborderons les systèmes d'évaluation de l'accent lexical principalement en anglais, qui a constitué un grand intérêt ces dernières années. Nous présenterons les techniques de modélisations auxquelles ils recourent. Enfin nous présenterons deux études qui évaluent le score global de fluence d'apprenants et les comparent à l'évaluation d'examinateurs.

3.1 Les outils de détection d'erreurs segmentales

Les premiers systèmes d'évaluation de la prononciation se basent sur un texte de référence lu par l'utilisateur. À l'aide d'un modèle acoustique, ils comparent ce qui est réellement prononcé par l'apprenant, avec la référence attendue. Le score de prononciation est alors calculé comme un *Word Error Rate* (WER), à partir du nombre de substitutions, d'insertions et de délétions détectées par rapport au texte

de référence. C'est ce que permet justement le logiciel *Evelyn* (Mostow et al. 1993). Le système affiche un texte à l'écran, et un curseur suit dynamiquement la lecture de l'apprenant au moyen d'un système d'ASR (CMU Sphinx-II, Huang et al. 1993). Une fois la lecture terminée, le système compte les écarts entre ce qui est reconnu et la référence, et renvoie un *feedback* audio-visuel à l'apprenant. Ce dernier voit alors à l'écran quels sont les mots qui ont été remplacés ou supprimés. Un *feedback* oral prononce également les mots substitués par l'apprenant pour lui indiquer ce qu'il aurait dû prononcer.

Le système de Kim et al. (1997) cherche quant à lui à donner un *feedback* pour chaque réalisation des dix phonèmes cibles choisis par les auteurs. Le score pour chaque phonème est un *log-posterior probability*, soit la probabilité du segment de parole X d'appartenir au phonème cible W ($p(X|W)$ avec $X = x_1, x_2...x_M$ représentant le signal de parole). Les modèles acoustiques sont appris sur des phrases lues par 100 locuteurs francophones natifs, et le système est testé sur de la parole non-native, phrases lues également, par 100 anglophones apprenant le français. L'évaluation humaine est faite par 5 enseignants et consiste à évaluer la prononciation des phonèmes cibles indépendamment des phrases dans lesquelles ils apparaissent. La corrélation entre ces scores et les scores automatiques reste toutefois relativement basse ($r = 0,72$). On pourra citer également le système de Franco et al. (1997) qui calcule un *log-posterior probability* sur l'ensemble des phonèmes d'une phrase lue. Les scores automatiques sont encore moins corrélés avec l'évaluation humaine ($r = 0,58$).

Certains systèmes proposent également de renseigner à l'avance les erreurs attendues chez les apprenants d'une langue maternelle donnée (comme la substitution ou la délétion de certains phonèmes), et laisse le système de reconnaissance « choisir » parmi une liste limitée de réalisations. Le système devient alors dépendant de la L1 de l'apprenant. Coulange (2016) propose par exemple au système d'alignement EasyAlign (Goldman 2011) d'aligner ou non certains phonèmes cibles, en fonction de ce que l'apprenant prononce. Ces phonèmes ont été choisis en fonction de réalisations fréquemment observées chez des locuteurs japonophones en français. Harrison et al. (2009) complètent quant à eux le dictionnaire phonétisé de leur système d'ASR par les productions phonétiques attendues des apprenants, en fonction de leur L1.

Tous ces systèmes dépendent des performances du système d'ASR utilisé pour reconnaître les phonèmes ou les mots, et les résultats sont donc très influencés par le type de corpus et la quantité de données sur lesquelles ont été appris les modèles acoustiques. D'après Truong et al. (2004), il serait plus fiable d'entraîner un système à reconnaître certains types d'erreurs spécifiques, plutôt que calculer un score de la même manière quelque soit le phonème. C'est ce qu'ils proposent, en se focalisant sur les réalisations du /ʏ/, du /a/ et du /x/ en néerlandais, qui sont des phonèmes connus pour poser problème à un grand nombre d'apprenants. Les auteurs utilisent le corpus

DL2Ni (Cucchiarini et al. 2000), qui contient des phrases lues au téléphone par 20 locuteurs natifs et 60 non-natifs de L1 et de niveaux hétérogènes. Les phrases sont ensuite transcrites et alignées pour identifier les segments du signal correspondant aux phonèmes cibles. Ils entraînent un modèle par sexe et par phonème sur les natifs et le testent avec les non-natifs. Ils arrivent à une précision moyenne de 65%, 70% et 75% pour les hommes et 70%, 70% et 91,7% pour les femmes pour les trois phonèmes respectivement.

3.2 Les techniques d'ASR

Un grand nombre d'outils se basant sur les techniques d'ASR, nous allons présenter les principes généraux de ces systèmes. L'objectif d'un système de reconnaissance de parole est de trouver la suite de mots la plus probable sachant un signal donné et une certaine connaissance de la langue. Une première étape consiste à extraire les paramètres acoustiques du signal nécessaires à la reconnaissance. Ensuite, la reconnaissance consiste en réalité à trouver la suite de mots qui maximise la probabilité d'observer ces paramètres acoustiques. Un système de reconnaissance recourt traditionnellement à un modèle de langage, pour cibler les suites de mots les plus fréquentes dans la langue ; un lexique phonétisé, permettant de trouver les suites de phonèmes associés à chaque mot, et des modèles acoustiques permettant de modéliser les phonèmes à partir des paramètres du signal.

L'extraction des paramètres acoustiques a pour but de ne garder que les informations minimum nécessaires à la reconnaissance. Ces paramètres peuvent être de plusieurs types : les *Linear Prediction Cepstral Coefficients* (LPCC), la Prédiction Linéaire Perceptuelle (PLP), les *Linear Frequency Cepstral Coefficients* (LFCC) ou encore les *Mel Frequency Cepstral Coefficients* (MFCC) (Furui 1981). Tous ces coefficients présentent l'avantage d'être décorrélés entre eux, ce qui minimise le nombre de coefficients nécessaires pour reconnaître les phonèmes prononcés. Les MFCC sont les coefficients les plus répandus dans les systèmes d'ASR, nous les présentons plus longuement dans l'annexe 9.4.

Les modèles acoustiques représentent la probabilité d'observer des paramètres acoustiques sachant une suite de phonèmes donnée. Ils peuvent être sous la forme d'un réseau de neurones profond (DNN), combiner un DNN et des chaînes de Markov cachées (HMM), ou encore, et plus traditionnellement, combiner des HMM avec des mélanges gaussiens (GMM). Les HMM sont des automates à état fini qui représentent chaque phonème en 3 états : le début, le milieu et la fin. Cela permet de modéliser la partie « fixe » du phonème, mais également les phénomènes de coarticulation. Chaque état est caractérisé par une loi de densité de probabilité dans

l'espace des paramètres acoustiques (un GMM), que nous décrirons plus longuement dans le chapitre suivant. En bref, chaque partie du phonème est modélisée par une distribution de probabilités d'apparition, et cette distribution dépend du corpus sur lequel ont été appris les modèles acoustiques.

Le lexique phonétisé associe à chaque mot de la langue une ou plusieurs suites de phonèmes. Pour un mot, il est ainsi possible de calculer la probabilité $p(X|W)$ que le signal observé X soit produit par un locuteur prononçant le mot W en concaténant tous les modèles HMM correspondant à la suite de phonèmes de ce mot. Le lexique phonétisé permet ainsi de modéliser les variantes de prononciation d'un même mot. Pour une même suite de phonèmes, plusieurs mots peuvent correspondre, c'est pourquoi entre en jeu également un modèle de langage, appris sur une grande quantité de textes, qui permettra de choisir la suite de mots la plus probable. Ce modèle génère une probabilité d'observer W dans la langue, indépendamment du signal. Ainsi, une suite de mots plus fréquente aura plus de chances d'être reconnue.

On pourra résumer le processus de reconnaissance de la parole tel que décrit ci-dessus par la figure 3.1. La suite de mots optimale W étant décrite par les arguments qui maximisent la probabilité d'obtenir cette suite de mots dans la langue $p(W)$ avec la probabilité de reconnaître les phonèmes de ces mots sachant une séquence X de paramètres en entrée.

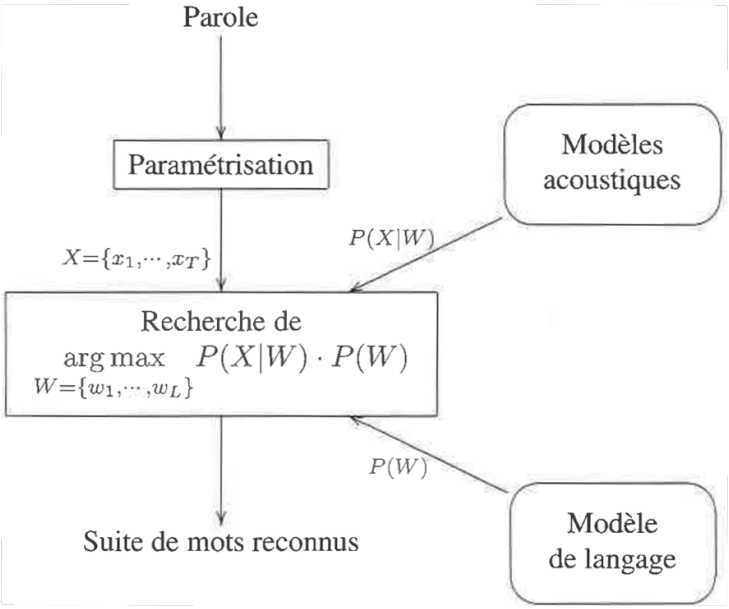


FIG. 3.1: *Principe de la reconnaissance bayésienne de la parole (HATON et al. 2006, p. 11)*

Il est important que les deux modèles soient adaptés au contexte que l'on souhaite reconnaître. Un modèle appris sur un corpus majoritairement masculin aura des

difficultés à reconnaître la parole des femmes ; s'il est appris sur un corpus de parole médiatique, il sera peu performant pour reconnaître une conversation entre amis.

Aussi on peut se demander si un modèle acoustique appris sur de la parole native permettra de reconnaître correctement de la parole d'apprenants, avec les différences segmentales et prosodiques qu'elle peut comporter. Certaines études proposent d'ailleurs d'apprendre ce modèle sur de la parole non-native, ou de mélanger natifs et non-natifs (Truong et al. 2004). Nous avons aussi parlé plus haut de la possibilité d'enrichir le dictionnaire phonétisé de réalisations attendues de la part d'apprenants (Harrison et al. 2009). On parle alors de réseau de reconnaissance étendu (ERN, N. Chen et H. Li 2016).

3.3 Les outils d'évaluation de l'accent lexical en anglais

Les systèmes de classification de l'accent lexical ont constitué un grand intérêt depuis le début des années 2000, d'une part car c'est une difficulté récurrente chez les apprenants de l'anglais L2, surtout pour ceux dont la langue maternelle n'a pas d'accent de ce type ; mais également parce que l'annotation manuelle est laborieuse et l'accord inter-examinateur est relativement faible. Poiré (2006) montre en effet que sur sept experts de prosodie à qui l'on demande d'annoter chaque syllabe accentuée sur 3 minutes de français spontané (165 syllabes), la proportion de syllabes jugées accentuées varie entre 19 et 49% selon les experts. Dans l'étude de Ferrer et al. (2015), le taux de désaccord entre les trois annotateurs atteint 21%. Il semblerait que chacun ne se base pas sur les mêmes paramètres pour juger si une syllabe est accentuée ou non. Goldman et al. (2007) expliquent qu'il est nécessaire de ne pas se limiter à un seul paramètre – comme la F0, qui est souvent choisie –, mais qu'il est également nécessaire de considérer d'autres paramètres comme la durée de la syllabe ou son intensité. Comme M. Morel et al. 2006, ils conseillent de toujours envisager une approche au moins partiellement automatique pour avoir une annotation systématisée des proéminences.

Dans ce contexte, de nombreux systèmes de classification de l'accent sont apparus. La plupart combine des mesures de durées, d'intonation et d'intensité (J.-Y. Chen et Wang 2010 ; L.-Y. Chen et Jang 2012 ; Deshmukh et Verma 2009 ; Tepperman et Narayanan 2005), et plus rarement des coefficients MFCC (Ferrer et al. 2015 ; C. Li et al. 2007 ; Shahin et al. 2016). Nous présentons ci-dessous deux systèmes récents de classification de l'accent lexical qui utilisent des paramètres prosodiques et spectraux, et qui comparent plusieurs techniques de modélisations.

Shahin et al. (2016) proposent un système de classification automatique des accents dans le but de l'intégrer à un système d'EPAO. Leur système permet de ca-

tégoriser chaque syllabe en deux classes (non-accentué, accentué) en anglais et en arabe standard. Ils testent deux architectures parallèles : un réseau de neurones profond (DNN) et un réseau de neurones convolutionnel (CNN).

Leurs systèmes sont appris sur trois corpus différents, le corpus TIMIT d'anglais adulte de Garofolo et al. (1993) (630 locuteurs, 8 dialectes américains, lecture), le corpus OGI d'anglais enfant de Shobaki et al. (2000) (1100 locuteurs de l'Oregon, du Kindergarten à la dixième année d'école, lecture et spontané), et l'*Arabic Speech Corpus* constitué des 6h d'audio du manuel *Al-kitaab Text Book*. L'accent est annoté automatiquement pour l'anglais à partir d'un dictionnaire d'accentuation, et manuellement pour l'arabe.

La transcription est alignée au signal avec un alignement forcé réalisée en fixant la suite de mots W de la transcription dans un ASR. Puis les paramètres sont extraits pour chaque syllabe : 27 coefficients cepstraux sur 30 trames de signal du milieu de chaque syllabe, et 7 paramètres temporels (la variation d'intensité dans le noyau syllabique, l'énergie moyenne, son maximum, la durée du noyau et celle de la syllabe, la F_0 moyenne et son maximum). Chaque vecteur comprend les paramètres d'une syllabe, de celle qui précède et de celle qui suit ; soit 2451 valeurs par vecteur.

La détection des syllabes accentuées par CNN s'avère légèrement plus performante que celle du DNN sur les corpus d'anglais, et comparable au DNN pour l'arabe : le taux d'erreur est de 6,52, 7,2 et 18 pour le corpus d'anglais enfant, adulte et le corpus arabe. Pour le DNN il est de 7,2, 7,7 et 17,9. Toutefois, les auteurs n'ont pas testé leur classifieur sur de la parole d'apprenants, ce pour quoi il est pourtant destiné.

Ferrer et al. (2015) conçoivent un classifieur pour le logiciel EduSpeak®, et évalueront sa performance avec des données d'apprenants. Leur système utilise également une combinaison de paramètres spectraux (MFCC et pente spectrale) et prosodiques (durée de la syllabe, F_0 et intensité normalisées). Les syllabes sont classifiées en trois catégories (non accentué, primaire, secondaire). La sortie se présente aussi sous la forme d'un postériogramme, donnant la probabilité d'appartenance de la syllabe à chaque classe d'accent.

Le corpus d'apprentissage est constitué de 157 888 syllabes lues par 329 enfants anglophones natifs (américain de la Côte Ouest) de 10 à 14 ans. L'annotation de chaque syllabe est faite automatiquement à partir d'un dictionnaire de prononciation. Le système est enfin testé sur un corpus d'anglais prononcé par des enfants japonais, constitué de 168 locuteurs de 10 à 14 ans et de niveaux hétérogènes, sur de la parole lue. Un total de 848 mots polysyllabiques est annoté manuellement par trois experts.

Les auteurs comparent les performances de trois types de modélisation : un GMM, un arbre de décision CART-Style (Buntine 1995) et un DNN, avec plus ou moins de paramètres en entrée. Le GMM avec l'ensemble des paramètres acoustiques

se révèle plus performant que les autres modèles et les autres combinaisons de paramètres. Le taux d'erreur du système est de 11,5 et 22,5 pour la parole native et non-native respectivement, contre 13,7 et 23,2 pour l'arbre de décision et 14,4 et 22,6 pour le DNN, toute configuration égale par ailleurs. On constate que dans tous les cas, il est plus difficile d'identifier correctement les syllabes des apprenants.

Ce genre de systèmes de classification de l'accent lexical permet ainsi d'évaluer l'accentuation de locuteurs non-natifs, en détectant automatiquement les syllabes qui auraient dû être accentuées mais qui ne sont pas reconnues comme telles. La comparaison avec les natifs est importante pour s'assurer que le système reconnaît bien les syllabes accentuées dans ce type de production chez les natifs. Leur inconvénient majeur est qu'ils nécessitent une grande quantité de parole transcrite pour apprendre les modèles de reconnaissance, ainsi qu'un système de reconnaissance de la parole pour aligner la transcription au signal. Ils demandent aussi un dictionnaire d'accentuation et supposent que les locuteurs natifs accentuent selon cette norme.

3.4 Description des principales modélisations utilisées

De nombreux types de modélisations ont été testés en classification automatique de l'accent lexical. Nous détaillons les principales techniques dans cette section.

Les séparateurs à vaste marge (ou machine à vecteur de support, SVM) ont pour objectif de trouver le plan aux marges maximales pour discriminer les syllabes en fonction de leur accent. Pour ce faire, les SVM transforment l'espace de représentation des données en un espace de plus grande dimension, jusqu'à trouver l'hyperplan de discrimination optimal. J.-Y. Chen et Wang (2010), Deshmukh et Verma (2009) et Zhao et al. (2011) ont utilisé cette technique pour détecter les accents lexicaux.

C. Li et al. (2007) ont préféré augmenter un modèle acoustique basé sur les chaînes de Markov avec une transcription prosodique. Ils alignent dans un premier temps le signal à une transcription sans accentuation, pour aligner ensuite des labels prosodiques aux syllabes. Ils réentraînent alors un HMM avec une liste de phonèmes augmentée des phonèmes accentués. Dans leur système, les auteurs utilisent un modèle triphone pour prendre en compte le contexte phonémique, et chaque état n'est modélisé que par une seule gaussienne à partir de coefficients MFCC et d'une normalisation de la F_0.

D'autres auteurs préfèrent créer un modèle par classe d'accent, indépendamment du type de syllabe. C'est le cas de Ferrer et al. (2015), qui apprennent un GMM pour chacune des trois classes d'accent de leur corpus. Les GMM sont en effet réputés performants pour modéliser une grande variation, c'est pourquoi ils ont longtemps été

utilisés en reconnaissance du locuteur notamment (Kahn 2011). Chaque modèle représente donc la densité de probabilité d'une classe sachant toutes les valeurs possibles des différents paramètres acoustiques mesurés, en fonction des tendances observées dans le corpus. Pour apprendre le GMM, les auteurs utilisent un algorithme d'Espérance Maximisation (EM, Laird 1993), qui cherche à optimiser les paramètres du modèle pour maximiser la vraisemblance des données avec celui-ci. En d'autres mots, l'algorithme calcule une courbe de densité de probabilité qui décrit au mieux distribution des données.

Pour palier au manque d'accents secondaires dans le corpus d'apprentissage, Ferrer et son équipe décident d'apprendre un modèle toutes classes confondues, et d'adapter ensuite ce modèle général avec les valeurs de chaque type d'accent. C'est une technique également utilisée en reconnaissance du locuteur, pour générer un modèle suffisamment robuste pour un locuteur avec seulement quelques minutes de parole (Ajili 2017 ; Kahn 2011). Les auteurs recourent à un algorithme de Maximum A Posteriori (MAP, Gauvain et Lee 1994) qui consiste à adapter les paramètres du modèle général (moyenne, covariance et poids de chaque gaussienne) en fonction des données de chaque classe, et permet de générer un modèle robuste pour chacune d'elles.

Shahin et al. (2016) avaient également l'objectif de traiter 3 classes d'accent, mais à cause du même problème du manque de données pour l'accent secondaire, ils abandonnent la distinction primaire/secondaire. Leurs deux systèmes (DNN et CNN) modélisent également chaque classe d'accent. Le nombre de couches cachées est ajusté empiriquement et les modèles sont entraînés avec une descente de gradient stochastique *mini-batch* (MSGD), jusqu'à ce que le taux d'apprentissage atteigne 0,0001 ou 200 époques, pour éviter le sur-apprentissage. On obtient en sortie une couche *softmax* donnant une probabilité d'appartenance de la syllabe à chaque classe d'accent. Dans le cas du CNN, le vecteur de 2451 valeurs est remappé en 3 spectrogrammes (2D), et sont traités comme des images. Les paramètres prosodiques ne sont ajoutés que par la suite dans une couche cachée.

La comparaison des performances de chaque système n'a pas réellement d'intérêt ici, dans la mesure où notre objectif n'est pas de travailler spécifiquement sur l'accent lexical, mais sur l'accent étranger dans sa globalité. De plus, les différences de corpus entre les études sont telles qu'il est difficile de comparer les performances des différentes techniques.

3.5 Évaluation de la fluence

L'évaluation de la fluence se fait au niveau de la phrase ou du texte avec des mesures globales. Il peut s'agir de mesures de durée de segments, comme celle des pauses,

des voyelles ou encore des syllabes, ou bien de mesures d'intonation, du nombre de pauses, ou de syllabes *etc*. Ces mesures globales ne recourent généralement pas à de la reconnaissance de la parole et peuvent donc se faire directement sur le signal.

Bhat et al. (2010) proposent de comparer les scores de prononciation issus d'une évaluation humaine avec ceux d'une évaluation automatique. Pour cela, ils utilisent un corpus de 136 minutes d'anglais L2 spontané issu de la version informatisée du TOEFL. Celui-ci est constitué de 28 locuteurs de 6 langues maternelles et 5 niveaux différents. Chacun d'eux doit décrire un film, un pays qu'ils veulent visiter, parler d'un problème social et indiquer un chemin sur une carte à un examinateur. Cet examinateur évalue alors la fluidité des énoncés sur une échelle de 0 à 4. L'évaluation d'un seul examinateur est utilisée pour l'expérience. Le corpus est ensuite divisé en 181 segments, soit légèrement moins d'une minute chacun, puis un certain nombre de mesures sont effectuées. Ces mesures sont le débit d'articulation, le débit de parole, le pourcentage de phonation, le nombre de silences par seconde, la moyenne de leur durée, et le nombre de pauses pleines par seconde. Les syllabes sont détectées grâce au script Praat de Jong et Wempe (2009) dont nous parlerons dans la partie méthodologique du mémoire. Les pauses pleines sont annotées manuellement.

Les auteurs apprennent ensuite au système à évaluer les segments de parole en fonction des scores de l'examinateur. L'apprentissage se fait au moyen d'un modèle de régression logistique, pour générer une probabilité de fluence. Pour s'adapter aux contraintes du modèle, tous les scores sont réduits à une valeur binaire fluent/non-fluent, selon que le score est supérieur ou inférieur à 2,5.

Après plusieurs combinaisons de paramètres, le débit de parole est abandonné pour ne garder que le débit d'articulation, calculé en enlevant les pauses. Le système final imite l'évaluation humaine avec une précision de 72,1%, et le degré d'accord entre l'homme et la machine est évalué à 0,66 avec le κ de Cohen. Le degré de dépendance entre chaque variable est calculé avec un *odds ratio*. Il indique que le débit d'articulation est de loin le paramètre le plus efficace (4,56), suivi par le pourcentage de phonation (1,23), la moyenne de durée des silences (0,14), puis le nombre de silences (0,0015) et le nombre de pauses pleines par seconde (0,0007).

Dans leur système CAPT-L2, Fontan et al. (2018) utilisent quant à eux une régression linaire pour générer un score automatique de fluence sur 12 sets de parole issus du Corpus Longitudinal Interphonologique de Japonais Apprenants de Français (CLIJAF, Detey 2011). Le corpus est constitué de 252 phrases lues par 8 étudiants, 4 hommes et 4 femmes entre 18 et 22 ans, dont 4 sont enregistrés après 4 mois puis réenregistrés après 19 mois d'apprentissage, et les 4 autres après un séjour d'une année en France. Le corpus est ensuite évalué par 3 experts en phonétique sur quatre dimensions : la fluence globale, le débit de parole, la régularité du débit, et la fluidité de la parole (au niveau de la coarticulation).

Les 252 phrases sont ensuite découpées en segments avec un algorithme *Forward-backward divergence segmentation (FBDS, André-Obrecht 1986)*, puis sont mesurés le débit de parole (nb de segments sur la durée de la phrase), la régularité du débit (écart type des durées de segments), la fluidité de parole (variation du premier formant entre les segments), le pourcentage de phonation, et enfin la longueur et l'écart type des silences. Les silences sont des segments d'au moins 250ms avec une intensité inférieure à 10% de l'intensité maximum de la phrase. La régression linaire prend pour variable dépendante le score de fluence globale moyen des 3 évaluateurs, et pour variables indépendantes les paramètres listés ci-dessus. Puisqu'il s'agit ici d'une régression linéaire et non logistique, il n'est pas nécessaire de réduire les scores à une valeur binaire.

Le meilleur modèle ne prend pas en compte la longueur et l'écart type des silences. Selon les auteurs, ils seraient trop corrélés avec le pourcentage de phonation et le débit de parole. Pour les autres paramètres, le débit de parole est le principal contributeur au modèle (coefficient β standardisé = 0,61), suivi par la régularité du débit (-0,19), le pourcentage de phonation (0,17) et la fluidité de parole (-0,15). Le score de fluence perçu et les scores prédits par le modèle une fois agrégés sont très corrélés ($r = 0,92$; $p < 0,001$). Les auteurs montrent également que la corrélation entre le score de fluence des évaluateurs et la fluence mesurée par les paramètres est forte dès 4 phrases par locuteurs ($r^2 > 0,80$) et se stabilise rapidement sans dépasser 0,85. La figure 3.2 présente l'évolution du r^2 moyen en fonction du nombre de phrases par locuteur.

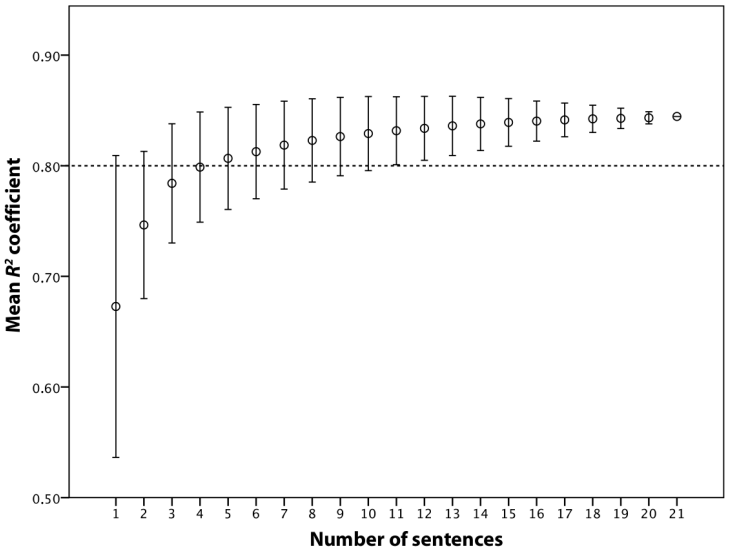


Fig. 3.2: *Coefficient de détermination r^2 moyen entre la fluence mesurée et perçue en fonction du nombre de phrases par locuteur (Fontan et al. 2018, p. 2547)*

On notera que l'avantage des deux systèmes décrits ci-dessus est qu'ils n'ont pas recours à la reconnaissance vocale, ni à une transcription. Toutes les mesures se font sur une segmentation entièrement automatique, soit par une détection des noyaux syllabique (Jong et Wempe 2009), soit par une segmentation de plus bas niveau comme le FBDS d'André-Obrecht (1986). Ces deux systèmes montrent que parmi les paramètres acoustiques utilisés, le débit et le pourcentage de phonation sont très corrélés avec les évaluations humaines. On remarquera toutefois que le débit de parole est abandonné dans Bhat et al. (2010) au profit du débit d'articulation. Pour Fontan et al. (2018) le débit d'articulation n'est pas mesuré, et le débit de parole arrive loin en tête devant les autres paramètres.

3.6 Conclusion du chapitre

Nous avons vu que de nombreux outils existent pour évaluer la prononciation. Certains cherchent à exploiter les techniques d'ASR et évaluent la précision de reconnaissance (Kim et al. 1997 ; Mostow et al. 1993), d'autres apprennent des modèles sur des phonèmes cibles spécifiques (Truong et al. 2004).

Concernant les paramètres suprasegmentaux, plusieurs travaux concernent la détection des syllabes accentuées en anglais (Ferrer et al. 2015 ; Shahin et al. 2016) ; seules deux études se concentrent sur l'évaluation de la fluence globale de l'énoncé (Bhat et al. 2010 ; Fontan et al. 2018).

Fontan et al. (2018) montrent que l'on peut obtenir une bonne corrélation entre l'évaluation humaine et les mesures automatiques de la fluence à partir de quelques phrases, et que les scores se stabilisent de plus en plus à mesure que la quantité de parole augmente. Ces mesures de fluence recoupent pour une large part les mesures rythmiques étudiées au chapitre précédent : les débits de parole et d'articulation, écart type de durée des segments, durée moyenne et écart type de durée des pauses silencieuses, pourcentage de parole *etc*.

Problématique

La perception de l'accent étranger est due à un écart de prononciation avec la norme attendue et partagée par les natifs de la langue (Alazard 2013). Les raisons de cet écart ont largement été décrites au niveau phonétique dans les théories de l'acquisition, mais l'importance de la prosodie, et notamment des durées de segments et de l'intonation, n'a été prouvée que plus tard dans les études de perception de l'accent. Il est clair aujourd'hui que les deux niveaux sont importants dans la perception de cet écart par des natifs.

Nous avons vu que, parmi les paramètres suprasegmentaux, le rythme varie sensiblement d'une langue à l'autre, et de nombreux auteurs ont ainsi envisagé une classification des langues seulement sur des paramètres rythmiques. La plupart des études du rythme se basent sur la durée des voyelles et des consonnes, et nécessitent donc une transcription et un alignement précis. Toutefois, de récentes études montrent des résultats similaires avec la durée des segments voisés ou des syllabes, qui peuvent être détectés et alignés automatiquement sans transcription (Dellwo et Fourcin 2013 ; Dellwo, Leeman et al. 2015 ; Fourcin et Dellwo 2013). La majorité des études fait également ressortir les limites de ses résultats, à cause de la taille des corpus utilisés et des méthodes d'élicitation. Il est nécessaire de traiter des corpus plus importants et sur de la parole plus diversifiée, notamment spontanée, car le rythme peut varier fortement selon les situations et les locuteurs, et les écarts sont plus importants encore avec la parole spontanée[1]. En 2016, Astésano parle encore de cette nécessité comme d'une « tâche urgente » (p. 85).

Plutôt que de se limiter à un type de discours, ou à une catégorie de locuteur, la solution est peut-être de travailler sur un maximum de diversité possible, pour étudier la langue dans sa grande diversité d'instantiation. Dans ce travail de mémoire, nous allons donc essayer de modéliser le rythme du français à partir du Corpus d'Étude pour le Français Contemporain (CEFC), qui est un corpus francophone particulièrement varié dans ses situations d'énonciation comme dans l'origine de ses locuteurs. C'est ce corpus qui a rendu possible notre expérience de modélisation du rythme.

Pour déterminer quels modèles utiliser pour modéliser le rythme de la parole tout en modélisant également la variation propre aux situations et aux locuteurs, nous nous sommes inspirés de Ferrer et al. (2015) qui utilisent les GMM pour modéliser des classes d'accent lexicaux, indépendamment du type de syllabe prononcée et du locuteur. Selon eux, les GMM sont adaptés à la modélisation de la variation, et semblent donc être une piste intéressante pour modéliser le rythme des locuteurs

1. ce qui explique peut-être les meilleurs résultats de Fontan et al. (2018) (lecture) par rapport à Bhat et al. (2010) (spontané).

dans leur diversité. À partir d'un modèle général du rythme, comme le modèle des accents de Ferrer et al. (2015), il devrait donc être possible d'évaluer convenablement le rythme d'un locuteur avec suffisamment de parole.

Nous tenterons de répondre aux questions suivantes : est-il possible d'obtenir un modèle du rythme à partir de locuteurs natifs ? Un score d'appartenance à ce modèle nous permettrait-il de distinguer des locuteurs natifs et non-natifs ? Si oui, quels sont les mesures du rythme les plus pertinentes pour faire cette distinction entre les productions de natifs et d'apprenants ? Nous examinerons également dans quelle mesure l'écart observé avec le modèle est corrélé avec le niveau de compétence en français du locuteur.

Deuxième partie

Méthodologie du travail de mémoire

Chapitre 4

Corpus de travail

Nous avons besoin pour cette étude d'un corpus de français qui nous permette de modéliser la langue dans sa diversité. Plus il y a de locuteurs et de situations d'énonciation différents, plus la parole pourra être prise en compte dans sa variation. Des corpus de parole exclusivement médiatique comme ESTER 1 et 2 (radio), ÉTAPE (radio, télévision) ou REPÈRE (télévision) ne pourront donc pas convenir, bien qu'ils fassent intervenir de nombreux interlocuteurs dans différents types d'émissions. Ces corpus sont également relativement petits : 100h transcrites pour ESTER 1, 129h pour ESTER 2, 42h pour ÉTAPE et 60h pour REPÈRE (Garnerin 2018). Récemment, le Corpus pour l'Étude du Français Contemporain (CEFC, Benzitoun et al. 2016) a mis à disposition une sélection de corpus uniformisés, totalisant plus de 2 500 locuteurs sur 300h de parole annotée manuellement et semi-automatiquement. C'est l'existence de ce corpus qui a motivé le présent travail de modélisation du rythme.

Ce chapitre se consacre à la présentation du CEFC, dont une partie des locuteurs natifs constitue notre corpus d'apprentissage, ainsi qu'à la constitution de trois corpus de test : un de locuteurs natifs et un deuxième constitué des locuteurs non-natifs du CEFC ; et le troisième constitué de deux classes d'apprenants japonophones du français langue étrangère (FLE) enregistrés à l'Université de Langues Étrangères de Kyōto en 2019.

4.1 Le CEFC

Le CEFC[1] est le produit du projet Outils et Recherches sur le Français Écrit et Oral (ORFÉO), financé par l'Agence Nationale de la Recherche dans le cadre de la campagne Corpus Données et Outils de la Recherche en Sciences Humaines et Sociales 2011. Ce travail est le fruit d'une collaboration de 7 laboratoires[2] et de chercheurs français, suisses, belges et japonais. Le CEFC rassemble dans un format unique et harmonisé plusieurs corpus sources, et totalise 4 millions de mots pour sa partie orale et 6 millions pour sa partie écrite. Nous ne décrirons ici que la partie orale.

Cette section a pour objectif de donner un aperçu de la diversité du CEFC. Nous commencerons par décrire le format des données, puis nous énumérerons les différents sous-corpus, les proportions des différents genres et situations de parole ainsi que les caractéristiques des locuteurs.

4.1.1 Mise à disposition du corpus

Le CEFC est disponible sous licence Creative Commons BY NC SA (Attribution, utilisation non-commerciale et partage dans les mêmes conditions), entièrement téléchargeable en une cinquantaine de giga-octets, ou bien consultables en ligne via deux interfaces. Chaque enregistrement consiste en un fichier audio de longueur variable au format WAVE, un fichier de transcription annotée et alignée dans un format natif ORFEO et un fichier XML de métadonnées indiquant les caractéristiques de l'enregistrement (titre, durée, nombre de mots, responsable, degré de planification du discours, date et lieu d'enregistrement, la langue, le médium, la qualité du son *etc.*), ainsi que les caractéristiques des locuteurs. Les enregistrements sont tous échantillonnés à 22,05 kHz, en 16 bits et en mono. Les XML suivent cependant différentes structures en fonction du corpus d'origine. Le lecteur trouvera un aperçu de transcription ORFEO et d'un fichier de métadonnées en annexe 9 et 10.

4.1.2 Les sous-corpus oraux

Le corpus oral regroupe 13 corpus sources, dont certains déjà diffusés par le passé (comme Valibel ou Clapi), et couvre une grande variété de français sur 315 heures d'en-

1. https://repository.ortolang.fr/api/content/cefc-orfeo/6/documentation/site-orfeo/index.html

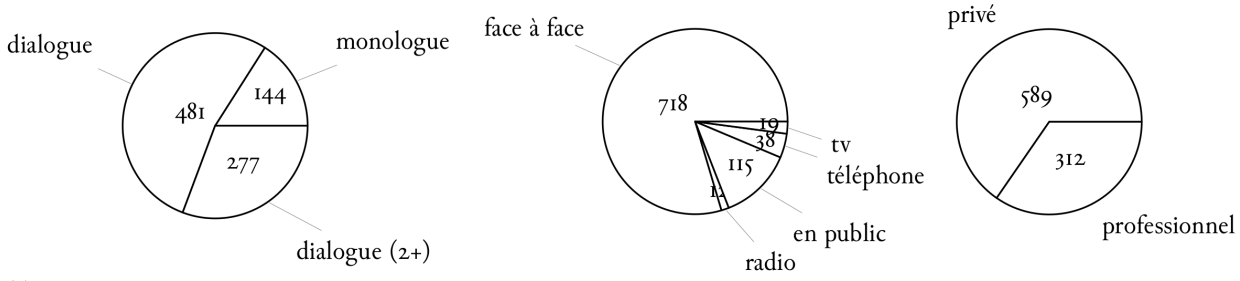
2. LATTICE, Paris 3 ; MoDyCo, Paris Ouest Nanterre la Défense ; ATILF, U. Lorraine ; LORIA, CNRS ; LIF, U. Aix-Marseille ; ICAR, Lyon 2 et CLLE-ERSS, U. Toulouse.

registrements récents de locuteurs adultes de toutes les régions françaises, de Suisse et de Belgique. Plus de 2 500 locuteurs sont enregistrés dans des situations d'énonciation diverses : conversations, réunions de travail, entretiens, interviews, interactions avec des services, conversations téléphoniques, contes, récits *etc*. L'ensemble du corpus est transcrit et aligné au mot et enrichi des parties du discours et des relations de dépendances. L'annexe 9.4 détaille les 13 corpus sources, avec leur durée et nombre de mots.

4.1.3 Statistiques générales sur le CEFC

Chaque enregistrement est accompagné d'un fichier de métadonnées. Ces fichiers nous permettent d'extraire un certain nombre de renseignements sur le contenu des enregistrements. Les statistiques de cette sous-section sont réalisées à partir des 902 fichiers de métadonnées du CEFC oral.

La figure 4.1 présente le nombre d'enregistrements en fonction de différents critères : le nombre de personnes impliquées dans la conversation (a), la situation d'énonciation (b) et le secteur (c). On constate que plus de la moitié des enregistrements sont des dialogues (481), mais beaucoup d'enregistrements impliquent aussi plus de deux locuteurs (277). La majorité des situations est un face à face (718), seulement 31 enregistrements proviennent de médias oraux (19 de la télévision et 12 de la radio). On trouve également 38 conversations téléphoniques et 115 discours en public. Près des trois quarts des enregistrements sont fait dans un environnement privé (589), souvent chez l'habitant[3].



(a) Nb d'enregistrements en fonction du nombre de personnes impliquées dans la conversation

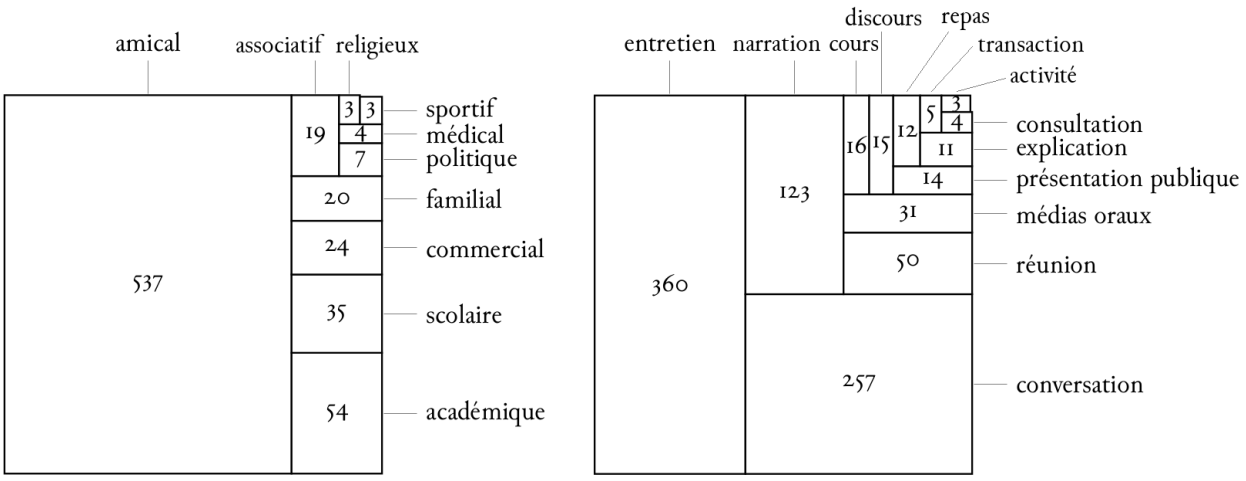
(b) Nb d'enregistrements en fonction de la situation d'énonciation

(c) Nombre d'enregistrements en fonction du secteur

Fig. 4.1: *Nombre d'enregistrements en fonction du nombre de locuteurs, de la situation et du secteur de la conversation*

Les conversation sont majoritairement en contexte amical (537), mais on trouve également des enregistrements en milieu académique (54) et scolaire (35), en contexte

3. Un enregistrement n'a pas de secteur renseigné (PUB-TOU-1).



(a) *Contexte de la conversation* ***(b)*** *Type de conversation*

FIG. *4.2:* *Nombre d'enregistrements en fonction du contexte et du type d'énonciation*

commercial (transactions à la boulangerie, dans une fromagerie *etc.*, 24) ou des repas en famille (20). Quant aux types de d'énonciation, on retrouve plus d'un tiers d'entretiens (360), presque un autre tiers de conversations (257), et des narrations (contes ou histoires lues, 123), des réunions (50), des médias oraux (31) ou encore des enregistrements de cours (16). La figure 4.2 présente le détail des proportions.

Pour manipuler plus facilement le corpus, nous avons constitué une base de données de l'ensemble des enregistrements et de leurs caractéristiques (*cf.* fichier enregistrements.csv). Le script python qui parse les XML pour la constitution de cette base de données s'intitule `infoEnregistrements.py`. Le CSV de sortie indique les renseignements suivants pour chaque enregistrement : nom du fichier, titre, nom du premier responsable, nombre de mots, commentaire date, commentaire son, corpus source, durée, date d'enregistrement, résumé, langue, type, secteur, milieu (contexte de la conversation), thème, channel (situation d'énonciation), modalité, nombre de locuteurs (1, 2 ou 2+), lieu d'enregistrement et le nom des locuteurs impliqués.

Les calculs précédents ont été faits sur le nombre de fichiers, il serait sans doute plus pertinent d'avoir les proportions en temps de parole ou en nombre de mots, étant donné que les fichiers ont des durées très variables (de 30 secondes pour `ftelpv28` et `ftelpv31` à 2h58 pour `commerce_fromagerie`). Néanmoins, ces premières statistiques permettent d'avoir un aperçu global de la constitution du corpus.

4.1.4 Croisement des données locuteurs

Dans les fichiers de métadonnées, chaque locuteur est introduit par la balise `<person xml :id="IdDuLocuteur">`. Les identifiants des locuteurs sont malheureusement uniques seulement pour l'enregistrement dans lequel ils apparaissent, aussi on retrouve de très nombreux locuteurs dont l'identifiant est L1, L2, ou A, B *etc.* sur l'ensemble du corpus. Il est donc impossible d'identifier les locuteurs par le seul nom qui leur est donné dans les métadonnées. Nous déciderons donc de les appeler de la manière suivante : `<nomFichier>_<IdDuLocuteur>-<sexe>`, pour être sûr d'éviter tout doublon et différentier les locuteurs en fonction des enregistrements. Voici un exemple : `PRI-POI-2_L3-F` pour la locutrice L3 de l'enregistrement `PRI-POI-2`.

Par la commande shell `grep '<person xml :id='`, on compte un total de 2587 locuteurs identifiés dans les fichiers de métadonnées. Certains locuteurs peuvent apparaître dans plusieurs enregistrements à la fois, mais étant impossible de les identifier automatiquement, nous traitons chaque locuteur de chaque enregistrement comme un locuteur unique.

Les métadonnées spécifiques aux locuteurs sont les suivantes : le sexe, la tranche d'âge, l'occupation, le niveau d'éducation, le lieu de naissance et le statut de la langue française (langue maternelle ou seconde). Selon les enregistrements et les locuteurs, certaines informations peuvent ne pas être renseignées, notamment lors de prises de parole en assemblée, ou lors de brefs passages de locuteurs dans un magasin par exemple. Afin d'exploiter plus facilement ces informations, nous avons constitué une base de données locuteurs combinant l'ensemble des champs mentionnés ci-dessus avec le nom de fichier et l'identifiant du locuteur (*cf.* fichier locuteursORFEO.csv). Le script python qui parse les XML pour la constitution de cette base de données s'intitule `locuteurs.py`. Malgré l'uniformisation des sous-corpus, de nombreuses différences persistent dans la structure des fichiers et le type d'informations renseignées[4]. Beaucoup de locuteurs ont donc des informations inconnues concernant leur sexe, leur tranche d'âge ou leur niveau d'éducation.

La figure 4.3 présente la répartition homme/femme (a) et le niveau d'étude (b) de chaque locuteur. On constate pour une fois une majorité de femmes (1373 pour 1048 hommes), contrairement aux corpus Ester, Étape et Repère (Garnerin 2018). Quant au niveau d'études, il va jusqu'au primaire pour 30 locuteurs, jusqu'au secondaire pour 262, et jusqu'au supérieur pour 1173 locuteurs. Dans le premier cas, approximativement la moitié des locuteurs a moins de 15 ans, et l'autre plus de 60

4. Notamment la gestion des espaces et des retours à la ligne, rendant difficile le parsage, le nom de certaines balises, la façon de renseigner une valeur (notamment la tranche d'âge), l'ordre des informations et leur nombre.

ans. La répartition des sexes en fonction du niveau d'études est bien équilibrée pour le primaire : 18 femmes pour 12 hommes, le collège : 58 femmes pour 57 hommes, et pour le lycée : 75 femmes pour 72 hommes. On trouve une majorité de femmes pour les études supérieures cependant : 752 pour seulement 421 hommes. Pour la plupart des locuteurs cependant, cette information n'est pas renseignée (1122 locuteurs, dont 470 femmes et 486 hommes).

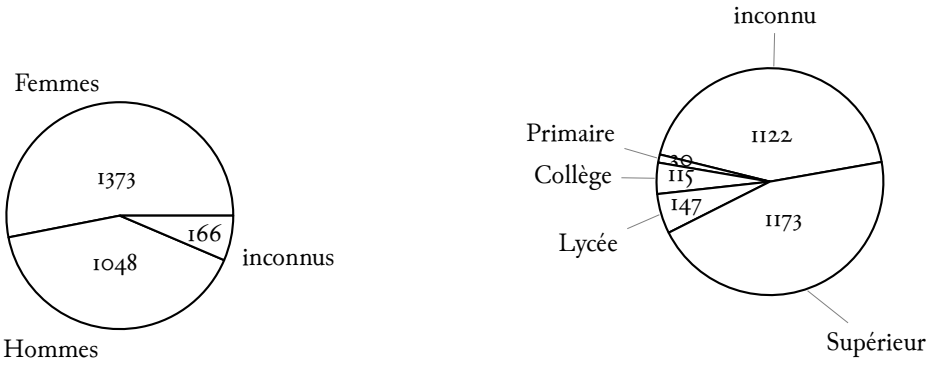


Fig. *4.3:* *Répartition des sexes et niveaux d'études des 2587 locuteurs*

Concernant les tranches d'âge, elles sont renseignées de manière à distinguer les enfants/pré-adolescents (0 à 15, 1,2%) des jeunes adultes (16 à 20, 5,3%), des adultes (21 à 60 ans, 43,4%), et des séniors (61+, 6,7%). 1002 locuteurs (38,7%) n'ont pas d'information d'âge, 121 locuteurs (4,7%) ont une tranche d'âge renseignée selon un autre format et ne sont pas compris ici.

On compte 348 occupations différentes enregistrées dans les métadonnées. 23% des locuteurs sont des étudiants (603), suivis par 6% d'enseignants-chercheurs (162). Les occupations suivantes par ordre de fréquence décroissante sont conteur (86), secrétaire (64), journaliste (56), enseignant (56) et retraité (52) *etc*.

Enfin, parmi les 2587 locuteurs renseignés dans les métadonnées, 57 sont identifiés comme francophones non-natifs et 207 ont un statut de français inconnu. Les locuteurs non-natifs constitueront une population intéressante pour comparer les mesures du rythme avec celles des locuteurs natifs. Les 207 locuteurs inconnus ont été ignorés pour la suite de l'étude.

Précisons que certains locuteurs présents dans les transcriptions ne sont pas indiqués dans les fichiers de métadonnées. N'ayant aucune information sur ces locuteurs, ils ont été ignorés dans la suite de l'étude. De plus, de nombreuses différences de formatage du nom des locuteurs sont présentes entre les transcriptions et les fi-

chiers de métadonnées, ce qui a provoqué la perte de 152 locuteurs au cours de la mise en forme des données, malgré une phase d'uniformisation des noms sur les caractères spéciaux et les espaces[5]. Précisons également que deux enregistrements sont manquants dans le corpus, on trouve pourtant un fichier xml et une transcription orfeo pour ces fichiers.

Pour la suite de ce travail, nous utiliserons uniquement les locuteurs correctement identifiés, soit 2 435 locuteurs, répartis sur 900 fichiers audio. Parmi eux, 2200 locuteurs sont renseignés natifs et 54 non-natifs.

4.2 Les corpus de test et d'apprentissage

À partir des données exploitables du CEFC, nous avons créé un corpus d'apprentissage composé uniquement de locuteurs natifs, un corpus de test de locuteurs natifs et un corpus de test avec tous les locuteurs non-natifs. Aucun locuteur des corpus de test n'a été utilisé pour l'apprentissage. En plus de cela, un troisième corpus de test composé de locuteurs de langue maternelle japonaise a été constitué pour croiser score rythmique et niveau de compétence en langue.

Précisons que 10% de l'ensemble des locuteurs du CEFC, soit 243 locuteurs sur 2435, ont été initialement mis de côté dans le but de constituer une partition de test comprenant des locuteurs natifs et non-natifs mélangés, dans l'objectif de les identifier automatiquement. Cet objectif a été abandonné en cours de route, car jugé trop peu pertinent, et les 243 locuteurs concernés n'ont pas été utilisés. Parmi eux, on compte 225 natifs, 9 non-natifs et 9 inconnus.

4.2.1 Le corpus d'apprentissage

Le corpus d'apprentissage se compose de 1 777 locuteurs natifs. Il s'agit des locuteurs restants après suppression des locuteurs de langue maternelle inconnue (172), récupération des non-natifs (45) et sélection aléatoire de 10% des locuteurs natifs pour la première partition de test (198).

Comme mentionné dans le chapitre précédent, les mesures rythmiques ont été

5. Des différences identifiées notamment dans la gestion des espaces, les tirets qui varient entre '-' et '_' pour les noms composés, ou encore les accents. Une phase d'uniformisation sur les tirets et espaces a permis de récupérer 87 locuteurs sur les 287 présents dans les métadonnées et non identifiés dans les transcriptions, et 135 locuteurs présents dans les transcriptions mais non-identifiés dans les métadonnées. Il semblerait que 152 locuteurs n'apparaissent pas dans les transcriptions ou n'ont pas le même identifiant.

effectuées sur des segments d'au moins 30 secondes de parole. Il arrive que des locuteurs n'aient pas de segments s'ils parlent moins de 30 secondes. Les locuteurs ayant au moins un segment de parole sont au nombre de 1 340 pour le corpus d'apprentissage.

4.2.2 Le corpus de test natifs

Le corpus de test natifs est composé de 198 locuteurs natifs sélectionnés aléatoirement parmi les natifs du CEFC. Parmi eux, 146 ont au moins un segment de parole. Les locuteurs ayant été sélectionnés aléatoirement, nous considérons qu'ils représentent la même diversité que le corpus d'apprentissage. Ils ne font donc pas l'objet d'une présentation spécifique.

4.2.3 Le corpus de test non-natifs du CEFC

Le corpus de test non-natifs est composé de 45 locuteurs. Parmi eux, 37 ont au moins un segment. Les statistiques suivantes décrivent ces 37 locuteurs.

La figure 4.4 présente la répartition homme/femme des locuteurs non-natifs, ainsi que leur niveau d'études. Nous avons ici 27 locutrices pour seulement 10 locuteurs. La grande majorité a un cursus scolaire allant jusqu'au supérieur (29), mais on trouve également un locuteur qui est allé jusqu'au collège et 4 jusqu'au lycée. On connaît la tranche d'âge de 16 locuteurs : 12 ont entre 21 et 60 ans, et 4 ont plus de 61 ans.

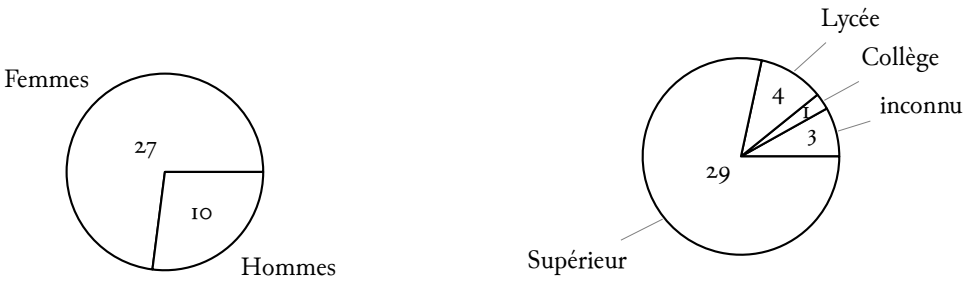


Fig. 4.4: *Répartition des sexes et niveaux d'études des 37 locuteurs non-natifs*

Les lieux de naissance sont assez divers : 6 locuteurs sont nés en Angleterre[6], 4 viennent de Suisse, 3 du Japon, 3 de Pologne et 3 d'Espagne. Deux locuteurs viennent d'un lieu non précisé en Europe, 2 non-natifs sont nés en France. La figure 4.5 présente les lieux de naissance des 37 locuteurs. Il est impossible de déterminer précisé-

6. Pour 4 d'entre eux, il est mentionné "Angleterre (?)"

ment la langue maternelle de ces personnes, ni leur niveau de langue ; on se contentera d'émettre l'hypothèse selon laquelle leur rythme est potentiellement influencé par leur langue maternelle, et donc possiblement différent de celui des locuteurs natifs.

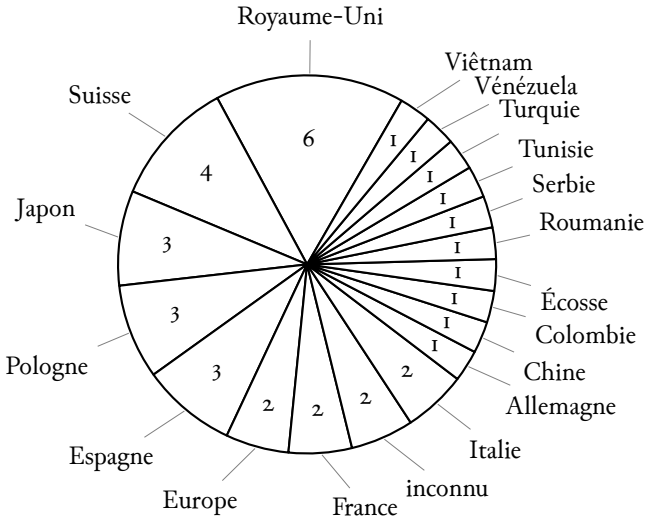


Fig. 4.5: *Lieux de naissance des locuteurs non-natifs*

Plus de la moitié des locuteurs non-natifs est étudiante au moment des enregistrements (24). On trouve ensuite différents métiers, d'agriculteur à maître céramiste en passant par juriste. Quatre locuteurs n'ont pas d'occupation renseignée. Ces indications laissent à penser que les 24 étudiants sont en échange universitaire et ont un niveau moins élevé que les 13 autres, qui vivent probablement en France. Toutefois, cela ne reste qu'une supposition.

La figure 4.6 donne un aperçu de la quantité de parole par locuteur. La quantité de parole est présentée en nombre de segments et en trames de voisement. On constate que la plupart des locuteurs parle assez peu, mais nous tenterons d'avoir une idée de leur rythme même avec 30 secondes de parole.

4.2.4 Corpus d'apprenants japonophones de FLE

Grâce à la coopération de Romain Jourdan-Ōtsuka, collègue enseignant de FLE à l'Université de Langues Étrangères de Kyōto, au Japon, nous avons pu bénéficier de l'enregistrement de 29 apprenants Japonais, lors d'une évaluation de production orale de fin de semestre. Pour cet échantillon, et contrairement à celui du CEFC, nous disposons du niveau de langue de chaque apprenant, ce qui permet de comparer les score de rythme obtenu au niveau de compétence, mais aussi et surtout cibler

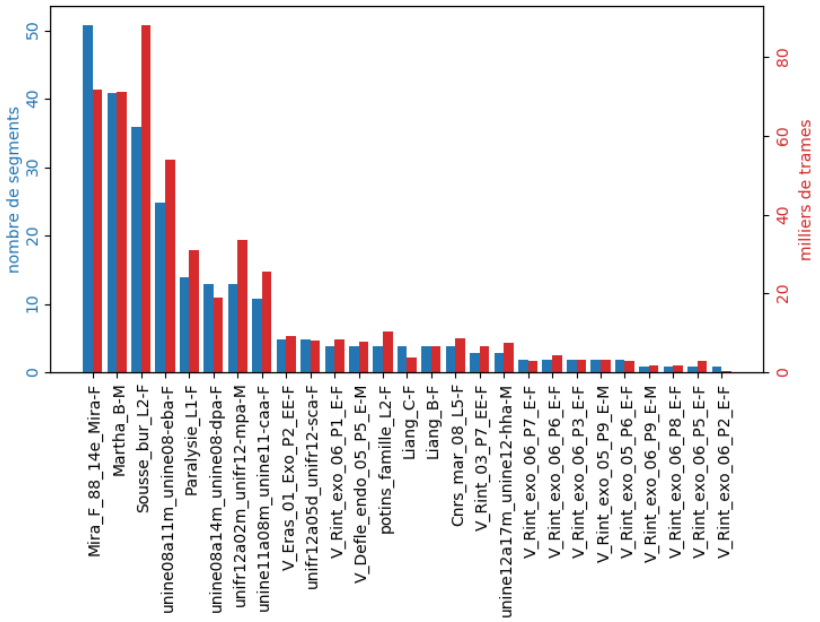


Fig. *4.6:* *Nombre de segments et de trames par locuteur*

une population en cours d'apprentissage de manière certaine. Dans cette partie, nous allons détailler les caractéristiques de la population, la tâche de production et les conditions d'enregistrement.

Nous disposons de l'enregistrement de deux classes d'étudiants de français de spécialité en 4$^{\text{ème}}$ année. Il s'agit des groupes C et D, soit les apprenants qui ont le plus de difficultés. Parmi eux, 18 étudiants ont fait 3 ans de français, les 11 autres en ont fait 4 pour cause de redoublement. Le niveau général de la classe est environ A2. Les étudiants ont entre 20 et 23 ans et on compte 17 femmes pour 12 hommes. Depuis leur arrivée à l'université, ils disposent de 1h30 à 3h de communication en français par semaine, en plus d'autres cours orientés sur la grammaire ou l'écrit. Deux d'entre eux ont effectué un séjour de 6 mois en France (Junko et Masaya). Nous ne disposons pas d'information précise pour les autres étudiants.

La figure 4.7 présente la note de chaque étudiant à l'examen de fin de semestre (4 compétences, évaluées sur 100 points) et la notes de production orale (PO, sur 25 points). On constate que le niveau de PO est relativement corrélé avec le niveau global en français. Certains étudiants sont bons surtout à l'oral, comme Hayato, d'autres ont une bonne compétence globale mais des difficultés à l'oral, comme Naoto. La grille d'évaluation de la PO est constituée de 5 catégories sur 5 points chacune : parler de soi, interroger l'autre, restituer l'information, précision (justesse) et aisance (clarté, fluidité). Nous disposons également du détail de ces notes.

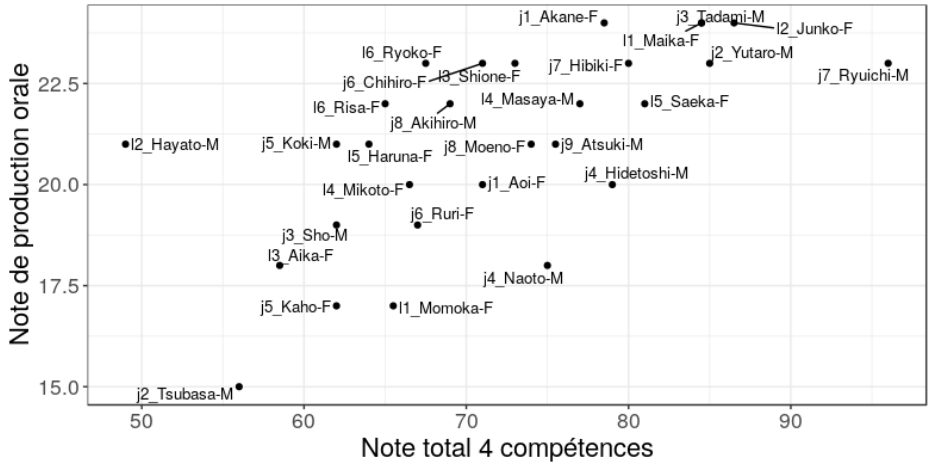


Fig. 4.7: *Notes obtenues en production orale en fonction de la note globale à l'examen (toutes compétences confondues)*

Les enregistrements se déroulent dans la classe, pendant l'examen sur table. Les étudiants viennent par groupe de deux à la table de l'enseignant pour l'évaluation de la production orale. De ce fait, les conditions d'enregistrement ne sont pas optimales ; on peut entendre des bruits de table ou de chaise, des toux ou des éternuements, ou parfois des bruits dans le couloir (notamment l'enregistrement j7). La voix des locuteurs est souvent assez basse (notamment Kaho (j5) et Shione (l3)), bien que l'enseignant leur répète régulièrement de parler plus fort. L'enregistrement a été fait au moyen d'un téléphone portable posé directement sur la table. Le son reste toutefois assez clair et on comprend globalement tout ce qui est dit.

La tâche de production se divise en deux temps. Lorsqu'il arrive, le binôme tire un sujet au hasard qui va lui indiquer le thème de la conversation. L'un des deux commence à poser des questions à son partenaire sur ce sujet, puis devra restituer à l'enseignant tout ce qu'il a compris. Ensuite les rôles sont inversés. Les sujets concernent le quartier dans lequel habite l'étudiant, l'âge des membres de sa famille, leurs occupations, leurs goûts, ou encore le temps que met l'étudiant pour aller à l'université et les moyens de transport qu'il utilise.

Nous disposons d'un enregistrement par classe, dans lequel s'enchaînent les binômes. L'enseignant nous a également fourni les temps de passage, ainsi que le nom des étudiants de chaque binôme. À partir de cela, nous avons découpé les enregistrements en 15 fichiers de 13 minutes en moyenne, un par binôme. Nous avons ensuite annoté manuellement chaque enregistrement en unités entre pauses et en identifiant les locuteurs. Après cela, la chaîne de traitement des fichiers est identique à celle des enregistrements du CEFC. Ajoutons que l'enseignant a également participé à un des

dialogues. Il s'agit d'un 30ème enregistrement, et il sera utilisé comme témoin natif.

L'annotation manuelle des tours de parole nous a permis de constater que le débit de parole est généralement très lent, et que les unités entre pauses sont nombreuses, assez courtes et séparées par de longues pauses d'hésitation. Cela est dû au niveau relativement bas des apprenants, mais également à la tâche d'énonciation qui implique de nombreuses questions-réponses.

4.3 Conclusion du chapitre

Nous avons choisi d'exploiter le corpus CEFC pour la diversité de parole qu'il présente, aux niveaux situationnel comme individuel, mais également pour sa taille et la présence d'une transcription alignée permettant d'identifier les locuteurs dans les enregistrements. Comme ce corpus présente également une cinquantaine de locuteurs identifiés comme non-natifs, nous pourrons évaluer leur rythme par rapport aux francophones de langue maternelle.

Toutefois, nous n'avons pas d'information sur le niveau de ces locuteurs, et leurs profils sont très hétérogènes. Aussi, nous proposons d'intégrer un second corpus de locuteurs non-natifs, pour lesquels le niveau de compétence en langue et en production oral est connu, et dont les profils sont homogènes (même langue maternelle, même conditions d'apprentissage etc.).

En résumé, notre corpus d'apprentissage se compose de 1777 locuteurs natifs et 17k segments de parole. Le corpus de test natif est constitué de 200 locuteurs et 2k segments. Le corpus de test non-natif du CEFC est composé de 44 locuteurs et 270 segments. Et enfin le corpus d'apprenants japonais de 29 locuteurs et 96 segments (*cf.* figure 4.8).

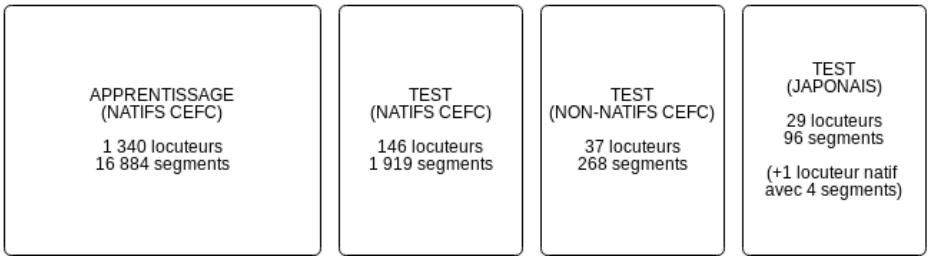


Fig. 4.8: *Constitution du corpus d'apprentissage et des 3 corpus de test*

Chapitre 5

Un modèle statistique du rythme pour le français

Pour développer un modèle statistique du rythme pour le français à partir d'un corpus constitué de nombreux locuteurs et de situations de communication variées, nous nous sommes inspirés des modèles utilisés en reconnaissance automatique du locuteur (RAL), et notamment du système ALIZÉ/SpkDet (Bonastre et al. 2005) qui recourt à l'approche UBM-GMM.

La première étape de notre travail a consisté à apprendre un modèle UBM-GMM sur notre corpus, à l'image des modèles utilisés en RAL. La plupart des systèmes existants se basent sur des paramètres issus d'analyses spectrales (Haton et al. 2006 ; Kahn 2011), et donc principalement sur des informations segmentales. Dans un second temps, nous avons appris un modèle similaire mais cette fois avec des paramètres mesurant le rythme de la parole. Nous nous inspirerons alors des métriques décrites dans le chapitre 2 ainsi que les mesures de Bhat et al. (2010) et Fontan et al. (2018) du chapitre 3.

Dans ce chapitre, nous présentons d'abord le modèle UBM-GMM utilisé en RAL, en décrivant plus en profondeur le principe du GMM et comment nous allons l'utiliser. Nous détaillons ensuite les choix de paramétrisation faits pour nos deux modèles.

5.1 Le modèle UBM-GMM

5.1.1 Modèle du monde

L'*Universal Background Model (UBM)* ou « modèle du monde » est une approche utilisée en RAL qui consiste à modéliser la parole indépendamment du locuteur dans un premier temps, pour l'adapter ensuite à un locuteur spécifique. Cette technique permet d'obtenir un modèle suffisamment robuste pour un locuteur dont on ne dispose que de quelques minutes de parole.

Le modèle du monde est un modèle appris sur une grande quantité de locuteurs dans différentes conditions, pour représenter « la parole » dans ce qu'elle a de commun entre les individus et les situations. Kahn (2011, p. 39) explique que son intérêt est d'« obtenir une représentation précise de ce qu'est la parole afin de structurer l'espace des paramètres autour des lieux où se concentrent les échantillons de parole ». Si ce modèle est unique pour tous les locuteurs, il arrive cependant qu'un modèle par sexe soit constitué.

Le concept de « modèle du monde » nous intéresse dans la mesure où nous souhaitons ici concevoir un modèle du rythme qui représente au mieux la diversité des rythmes du français, et bien-sûr indépendamment du locuteur. C'est par rapport à ce modèle que nous souhaitons évaluer le rythme des locuteurs. Il ne sera pas nécessaire de créer un modèle spécifique au locuteur. De plus, un modèle par genre ne semble pas très pertinent ici, étant donné l'impact insignifiant du sexe du locuteur sur la perception de son accent étranger (Flege, Munro et al. 1995 ; Piske et al. 2001). Nous constituerons donc un modèle unique du rythme.

Nous proposons d'apprendre un modèle $\lambda_{\bar{N}}$ à partir de la parole d'une grande quantité de locuteurs natifs $(\bar{N})$. Nous calculerons ensuite la vraisemblance d'échantillons de parole de locuteurs non-natifs NN par rapport à ce modèle, et nous la comparerons avec celle de la parole d'autres locuteurs natifs N. A priori, la vraisemblance des natifs avec le modèle devrait être plus élevée que pour les non-natifs, étant donné que le modèle a été appris sur de la parole exclusivement native.

5.1.2 Modélisation GMM

Les GMM ont longtemps été la modélisation la plus utilisée en RAL, bien que des approches plus récentes obtiennent aujourd'hui de meilleurs résultats, notamment les i-vecteurs (Dehak et al. 2011) ou les réseaux de neurones (Variani et al. 2014). Étant plus simples à prendre en main, et ayant des performances reconnues pour modéliser la variation, nous avons choisi d'utiliser les GMM pour modéliser le rythme

dans notre étude.

Un GMM est une densité de probabilité constituée d'une somme de gaussiennes pondérées. La somme de ces gaussiennes permet d'obtenir une fonction qui suit au mieux la distribution de nos données d'apprentissage, et l'apprentissage du GMM consiste à trouver les paramètres optimaux des gaussiennes pour représenter nos données. La probabilité d'un vecteur de mesures $\vec{x}$ étant donné un modèle défini par les paramètres $\{w_k, \vec{\mu}_k, \Sigma_k\}_{k=1}^{K}$ se calcule donc par la loi de densité de probabilité suivante :

$$p(\vec{x}) = \sum_{k=1}^{K} w_i \mathcal{N}(\vec{x}|\vec{\mu}_k, \Sigma_k) \tag{5.1}$$

où K est le nombre de gaussiennes du modèle, w_k le poids attribué à la gaussienne k et tel que $\sum_{k=1}^{K} w_k = 1$, et enfin $\mathcal{N}(\vec{x}|\vec{\mu}_k, \Sigma_k)$ la fonction normale de $\vec{x}$ selon le vecteur de moyennes $\vec{\mu}$ et la covariance Σ de k. Nous utiliserons ici une covariance diagonale pour alléger l'apprentissage, bien que certaines mesures prosodiques soient corrélées entre elles. Nous aurons ainsi une moyenne et un écart type par dimension de $\vec{x}$.

Pour faire notre apprentissage, nous avons choisi d'utiliser la classe GaussianMixture de la librairie Scikit-learn, en Python. Cette classe permet d'estimer les paramètres du GMM à l'aide d'un algorithme d'Espérance Maximisation.

Plus les données d'apprentissage sont variées et nombreuses, plus un nombre élevé de gaussiennes est conseillé. Dans le cas du système ALIZÉ/spkDet, ce nombre varie entre 256 et 2048 (Larcher et al. 2010), mais le temps et la mémoire vive nécessaire pour l'apprentissage augmentent de manière exponentielle avec le nombre de gaussiennes. Après plusieurs essais, nous avons décidé d'utiliser 1024 gaussiennes.

Une fois notre modèle appris, nous souhaitons calculer la proximité de locuteurs natifs et non-natif par rapport à ce modèle. Chaque locuteur sera représenté par l'ensemble des vecteurs de mesures calculées sur sa parole, avec un vecteur par trame de signal ou par segment de parole (*cf. sections 5.2.1, 5.2.2*). La probabilité d'observer l'ensemble des vecteurs $\vec{x}_n$ d'un locuteur sera donc égale au produit des probabilité d'observer chaque vecteur $\vec{x}_n$:

$$p(X) = \prod_{n=1}^{N} p(\vec{x}_n) = \prod_{n=1}^{N} \sum_{k=1}^{K} w_k \mathcal{N}(\vec{x}|\vec{\mu}_k, \Sigma_k) \tag{5.2}$$

Afin de simplifier le calcul, on transformera le produit en calculant $log\ p(X)$,

et on le normalisera par le nombre N de vecteurs par locuteur. La log-vraisemblance moyenne d'un locuteur sera calculée comme suit :

$$log\ p(X) = \frac{1}{N}\sum_{n=1}^{N} log\ p(\vec{x}_n) = \frac{1}{N}\sum_{n=1}^{N} log\left(\sum_{k=1}^{K} w_k \mathcal{N}(\vec{x}|\vec{\mu}_k, \Sigma_k)\right) \tag{5.3}$$

5.2 Les paramètres utilisés

Nous créons deux modèles : le premier avec des coefficients MFCC comme utilisés en RAL, le second avec des mesures rythmiques. Nous détaillons chacun des paramètres dans la présente section.

5.2.1 Coefficients cepstraux

Comme Ferrer et al. (2015), nous choisissons d'extraire 12 coefficients MFCC avec leurs dérivées premières et secondes et leur log énergie. Ces 29 paramètres constitueront les vecteurs en entrée pour notre modèle UBM-GMM. Toutefois, nous souhaitons nous concentrer sur la parole, en évitant les pauses et les bruits parasites des enregistrements. Aussi, une fois les coefficients calculés sur chaque enregistrement, nous les filtrerons de manière à ne garder que les trames contenues dans les intervalles voisés.

Différents outils existent pour extraire les MFCC. Nous avons choisi d'utiliser OpenSMILE[1] (Eyben et al. 2013), qui est un outil de reconnaissance de patterns avec un puissant extracteur de paramètres acoustiques. Il permet de calculer les dérivées des MFCC, ce que ne propose pas Praat par exemple. OpenSMILE (Speech and Music Interpretation by Large-space Extraction) est *open source* et utilisable gratuitement dans le cadre institutionnel.

L'installation et l'utilisation d'OpenSMILE est rapide et très bien documentée. L'outil dispose d'un fichier précompilé (SMILExtract) qu'il suffit d'exécuter pour accéder aux fonctionnalités d'extractions basiques du logiciel. Pour l'extraction des MFCC, nous avons utilisé un des fichiers de configuration proposés avec l'outil (`MFCC12_E_D_A.conf`). Lancé sans autres arguments que le fichier son, il génère une sortie au format HTK (Hidden-Markov Toolkit parameter file). Nous souhaitons récupérer les valeurs MFCC brutes, par trame, et nous avons donc spécifié une sortie

1. https://www.audeering.com/opensmile/

au format CSV[2].

En sortie, on obtient un fichier où chaque ligne donne les mesures d'une trame : son temps de début, 12 valeurs MFCC, 12 dérivées premières, 12 dérivées secondes, et trois logs énergie associés. Les trames ont une durée de 25ms et un pas de 10ms. On peut alors lancer une boucle d'extraction sur les 900 enregistrements du corpus.

Le filtrage se fait ensuite avec un script python qui prend en entrée le fichier de vecteurs MFCC, un textgrid d'intervalles voisés extraits automatiquement à partir de Praat et un textgrid d'unités entre pauses (UEP) par locuteur (`locVO2mfcc.py`). Ce script génère un fichier de vecteur par locuteur. Les UEP par locuteur sont préalablement extraites avec le script `orfeo2textgrid.py`, qui aligne des segments de parole par locuteur sur les enregistrements à partir des temps d'alignement des fichiers ORFEO. Toute pause supérieure à une seconde ou tout changement de locuteur met fin au segment de parole.

5.2.2 Paramètres rythmiques

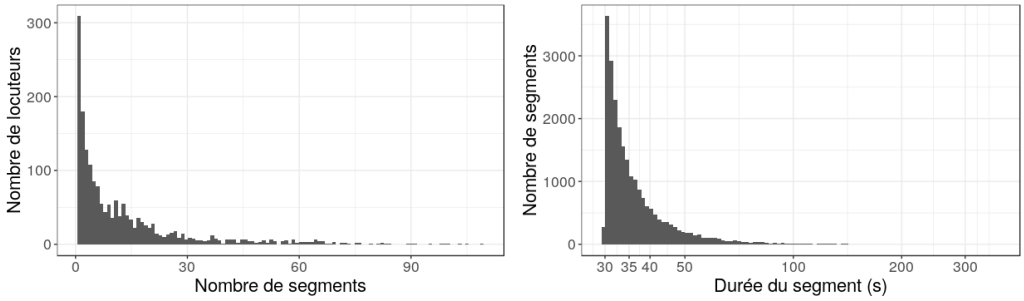
Les paramètres rythmiques sont calculés sur des segments de parole d'au moins 30 secondes. Nous pensons que cette durée permet d'avoir suffisamment de parole pour obtenir des mesures fiables de débit de durée de syllabes. Si le segment est plus court, les variations locales du rythme dues aux hésitations par exemple risquent d'impacter trop fortement les mesures.

Ces segments sont constitués par la concaténation d'UEP consécutives d'un même locuteur, jusqu'à atteindre au moins 30 secondes. Aucune UEP n'est coupée avant sa fin, il arrive donc que certains segments soient assez longs.

Chaque locuteur a donc un certain nombre de segments, parfois un seul s'il parle très peu, mais toujours supérieur ou égal à 30 secondes. Un total de 23 225 segments a ainsi été généré sur 2435 locuteurs. Parmi eux, 658 parlent moins de 30 secondes, et n'ont donc pas de segment.

La figure 5.1a présente la répartition du nombre de locuteurs en fonction du nombre de segments. La moyenne est de 13,04 segments par locuteur ayant au moins un segment. 309 locuteurs n'ont qu'un seul segment ; 5 locuteurs en ont plus de 100. La figure 5.1b montre quant à elle la distribution des durées des segments. Leur durée moyenne est de 37,31 secondes et leur médiane de 33,62. Certains segments peuvent être très longs, jusqu'à 374 secondes pour `PUB-TOU-1_L2-M_seg1`, qui contient effectivement un long monologue du locuteur L2, pratiquement sans interruption. L'annexe **??** liste les 20 locuteurs qui parlent le plus dans le corpus.

2. Plus d'informations page 36 du guide d'utilisation d'OpenSMILE.



(a) *Nombre de segments par locuteur (sur 1777 locuteurs ayant au moins 1 segment)*

(b) *Durées de segment (sur 23225 segments, moyenne = 37,31 s.)*

FIG. *5.1:* *Statistiques sur les segments de parole du CEFC*

Sur chacun de ces segments, nous avons calculé les 16 paramètres rythmiques suivants :

- $speechRate$: le débit de parole, ratio entre le nombre de syllabes et la durée du segment ;
- $\%dV$: le pourcentage de voisement, calculé par le ratio entre la durée de voisement et la durée du segment ;
- μdV, σdV et $VarcoV$: la moyenne, l'écart type et le coefficient de variation des durées des intervalles voisés ;
- μdU, σdU et $VarcoU$: la moyenne, l'écart type et le coefficient de variation des durées des intervalles non-voisés ;
- μdVU, σdVU et $VarcoVU$: la moyenne, l'écart type et le coefficient de variation des durées des intervalles voisés suivi d'un intervalle non-voisé ;
- $rPVI\_dV$ et $nPVI\_dV$: l'indice de comparaison brut et normalisé de paires successives d'intervalles voisés ;
- $\mu\Delta t$, $\sigma\Delta t$ et $Varco\Delta t$: la moyenne, l'écart type et le coefficient de variation des durées intersyllabiques ;

Le débit de parole est un indicateur pertinent dans la mesure où les énoncés non-natifs ont tendance à être plus lents que ceux des natifs. Il est calculé par le ratio entre le nombre de syllabes et le temps total de parole (avec les pauses) d'un segment de parole.

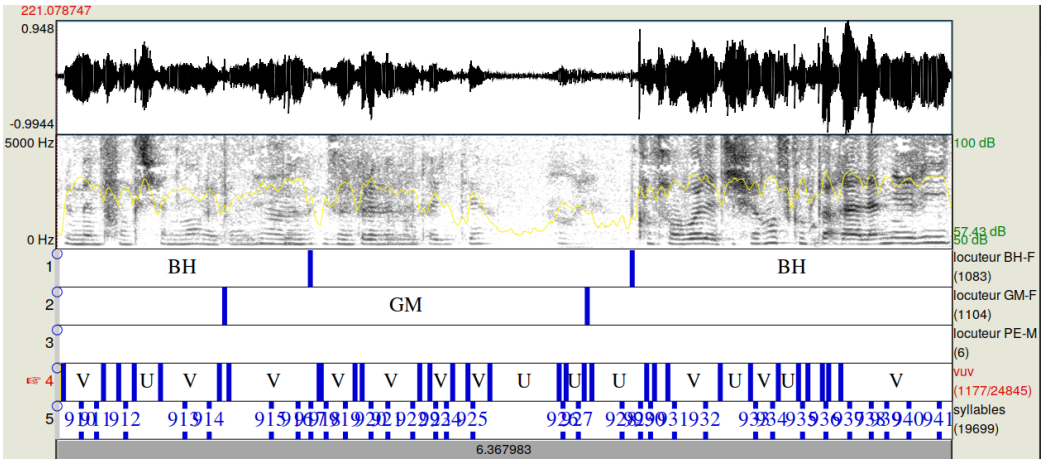


FIG. 5.2: *Aperçu d'un extrait d'enregistrement avec les UEP locuteurs (tires 1, 2 et 3) les segments voisés et non-voisés (tire 4) et les noyaux syllabiques (tire 5). Trois textgrids différents sont fusionnés ici*

Le pourcentage de voisement est calculé par le ratio de la durée totale de voisement sur la durée totale de parole avec pauses. Plus il y a de pauses dans le segment, plus le coefficient est faible. Rappelons que seules les pauses inférieures à 1 secondes sont conservées dans les UEP, et donc dans les segments de parole. Si un locuteur parle de manière continue, en voisant les hésitations ou les ralentissements, le pourcentage sera plus élevé que s'il fait de nombreuses pauses silencieuses. L'efficacité de ce paramètre en classification des langues n'est pas remis en question, et ne demande pas de normalisation avec le débit d'articulation (FOURCIN et DELLWO 2013).

Les intervalles de voisement ont été détectées automatiquement avec Praat (BOERSMA et WEENINK 2019). L'extraction en chaîne de ces intervalles est faite à partir du script `vuv.praat`, qui génère un textgrid d'intervalles voisés (V) et non-voisés (U) par enregistrement. Ces intervalles sont visibles sur la tire n°4 de la figure 5.2. Les intervalles de voisement par locuteur sont extraits dans un second temps à partir des UEP de chaque locuteur et concaténés en segments de parole (*cf.* `vuv2seg`). Les intervalles de voisement qui chevauchent une frontière d'UEP sont réduits au temps de début ou de fin de l'UEP en question.

Les coefficients de variation nous informent sur la variance des durées d'intervalles en normalisant sur le débit d'articulation. Ils sont calculés par le ratio entre l'écart type et la moyenne des durées de chaque type d'intervalle. Nous conserverons tout de même les écarts types non normalisés.

L'indice de comparaison normalisé de paires successives d'intervalles voisés est aussi un paramètre qui a prouvé son efficacité dans (WHITE et MATTYS 2007b), bien

qu'il soit moins discriminant que le coefficient de variation. Nous conservons également l'indice brut $rPVIdV$. Ces indices sont calculés comme indiqué dans le chapitre 2, section 2.3; le détail du code est visible en annexe **??**.

Comme Bhat et al. (2010), nous avons détecté les syllabes à partir de leurs noyaux vocaliques. Ceux-ci sont détectés à partir des pics d'intonation suivis d'une chute, avec une adaptation du script `syllable-nuclei_v2.praat` de Jong et Wempe (2009). On peut voir les noyaux syllabiques détectés sur la tire n°5 de la figure 5.2. Comme pour les segments voisés, le script génère un textgrid de syllabes par enregistrements. La durée intersyllabique est calculée par la différence entre deux noyaux consécutifs. À partir de 300ms d'écart entre deux noyaux, nous considérons qu'il y a une pause entre les deux syllabes ; aussi, la durée moyenne $(\mu\Delta t)$ et l'écart type $(\sigma\Delta t)$ des syllabes du segment sont calculés seulement à partir des écarts inférieurs à 300ms. Nous appelons ces durées *deltas intersyllabiques* (Δt).

L'écart type des deltas intersyllabiques nous indiquera la variation de durée de l'écart entre deux syllabes successives au sein du segment. Nous nous attendons à observer une variation des deltas plus importante chez les locuteurs natifs que chez les non-natifs.

Le script python qui calcule ces paramètres est `seg2vec.py` (*cf.* annexe **??**). Pour chaque locuteur, il prend en entrée la liste des segments (concaténations d'UEP) et le textgrid des noyaux syllabiques. En sortie, le script génère un fichier de vecteurs de 16 paramètres par locuteur, indiquant le résultat des mesures pour chacun de ses segments de parole.

Précisons enfin que les valeurs de chaque paramètre sont d'ordres de grandeur très différents. Si les moyennes de durée de segments varient entre 0 et 0,25, le PVI normalisé monte jusqu'à 128. Cela n'a pas d'influence sur les performances de notre outil, mais pourra aboutir à des log-vraisemblances supérieure à zéro[3].

5.3 Conclusion du chapitre

Nous avons donc créé deux modèles pour cette étude. Le premier est un modèle UBM-GMM calculé sur des 39 paramètres spectraux (coefficients MFCC, deltas, delta-deltas et logs énergie). Le second modèle est appris de la même manière, mais sur 16 paramètres rythmiques. La modélisation est faite à l'aide de mélanges gaussiens à 1024 composants sur un corpus d'apprentissage exclusivement natif. Est

3. Lorsque la variance de certains paramètres est sur une petite échelle, la forme des gaussiennes devient alors un long pic vertical pouvant dépasser 1 sur l'axe des y. On se retrouve alors avec une log-vraisemblance pouvant être positive.

ensuite calculée la proximité à ces modèles de locuteurs natifs et non-natifs, issus d'une population de test. Cette distance est représentée par une log-vraisemblance moyenne des vecteurs de mesures pour chaque locuteur. Les paramètres spectraux sont calculés sur des trames de 25ms de parole voisée uniquement. Les paramètres rythmiques sont quant à eux calculés sur des segments de parole supérieurs à 30 secondes, constitués d'une concaténation d'unités entre pauses.

Chapitre 6

Analyses statistiques

Dans ce chapitre, nous présentons les tests statistiques que nous avons utilisés pour analyser nos résultats. Dans un premier temps, il a fallu déterminer si nos échantillons de test de locuteurs non-natifs ont un score rythmique significativement différent du score de la population de test de locuteurs natifs. Pour cela, nous avons utilisé le test de somme des rangs de Wilcoxon-Mann-Whitney. Ce test est l'équivalent d'un t-test lorsque la distribution normale des données n'est pas assurée.

Nous avons ensuite souhaité mesurer l'influence de chacun des 16 paramètres rythmiques utilisés pour modéliser le rythme dans notre modèle. Pour cela, nous avons utilisé le test des êta-carrés (η^2). Ce test nous a permis d'identifier les paramètres qui contribuent le plus à la modélisation de l'écart rythmique entre les locuteurs natifs et non-natifs.

Enfin, nous avons également recouru à différents tests de corrélation, pour examiner la relation entre le score rythmique obtenu par les apprenants japonais, et leur niveau de compétence en langue. Nous avons utilisé le coefficient de corrélation linéaire de Pearson (r) et le coefficient de détermination r^2, ainsi que les rangs de Spearman (ρ).

6.1 Test de Wilcoxon-Mann-Whitney

Le test qui permettra de comparer le score des locuteurs natifs et non-natifs est le test de Wilcoxon-Mann-Whitney. C'est un équivalent du χ^2 adapté pour des données ordinales, et l'équivalent non-paramétrique du test *t* de Student. Nous avons en effet deux échantillons d'observations continues (scores numériques) non appariés : d'un côté le score des natifs, de l'autre celui des non-natifs. Les scores sont peu

nombreux pour certains échantillons (29 pour les apprenants japonais, 37 pour les non-natifs du CEFC), impossible donc de décrire la distribution seulement par une moyenne et une variance, et comparer les deux échantillons avec ces paramètres. Nous avons donc utilisé un test non-paramétrique.

Les deux hypothèses que nous allons tester sont les suivantes : soit les deux échantillons appartiennent à la même population, il n'y a donc pas de différence significative entre les deux (H_0); soit les deux échantillons appartiennent à deux populations différentes (H_1). Pour notre test, (H_1) indiquera que les locuteurs natifs ont tendance à avoir un score plus élevé que les non-natifs. La p-value que nous obtenons à l'issu du test représente le risque d'erreur en rejetant l'hypothèse nulle.

Le test de Wilcoxon-Mann-Whitney consiste à classer les observations des deux échantillons analysés, puis faire la somme des rangs d'observation de chaque échantillon. Nous indiquons à titre indicatif la somme des rangs pour l'échantillon des non-natifs (U_{NN}).

L'implémentation du test est faite avec la librairie python `scipy.stats` et sa fonction `mannwhitneyu(X,Y,alternative="greater")` [1], avec X la liste ordonnée des scores natifs, Y celle des scores non-natifs, et en précisant un test unilatéral avec X > Y.

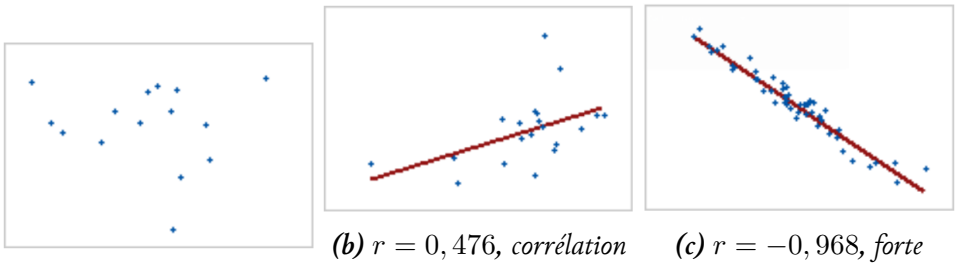
6.2 Test des êta-carrés

Le test des êta-carrés (η^2) nous permet de quantifier la part de variance de chaque paramètre expliquée par l'appartenance aux groupes des locuteurs natifs ou non-natifs. L'η^2 est calculé par le ratio entre la variance de chaque paramètre due au groupe d'appartenance natifs ou non-natifs ($SS_{between}$) et la variance du paramètre SS_{total}:

$$\eta^2 = \frac{SS_{between}}{SS_{total}} \tag{6.1}$$

Cohen (1988) propose de considérer l'effet d'un paramètre comme faible s'il est supérieur à 0,01, moyen à 0,06 et important s'il est supérieur à 0,14. De plus, la valeur de l'η^2 n'aura de sens que si la différence expliquée par le paramètre entre les deux échantillons est significative ; nous indiquerons donc également la p-value de cette différence. Seuls les η^2 des paramètres dont la p-value est inférieure à 0,05 seront

1. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html#scipy.stats.mannwhitneyu



(a) $r = 0$, *corrélation nulle*

(b) $r = 0,476$, *corrélation positive*

(c) $r = -0,968$, *forte corrélation négative*

FIG. 6.1: *Nuages de points avec différentes relations linéaires*

considérés.

6.3 Tests de corrélation

Afin d'examiner la relation entre le score rythmique obtenu par chaque apprenant japonais, et la note qu'il a obtenu à l'examen de fin de semestre, nous proposons d'abord de faire un test de corrélation de Pearson (r). La valeur de r est comprise entre -1 et 1. Plus elle est proche de 0, moins la corrélation est forte. Si la valeur est négative, la relation entre les deux variables est opposée ; si elle est positive, les deux variables évoluent dans la même direction. La figure 6.1 présente différents nuages de points plus ou moins corrélés[2]. Dans le cas où $r = 0$, les deux variables sont indépendantes.

Nous avons également calculé le coefficient de détermination r^2, pour mesurer le pourcentage de variation de Y expliqué par les valeurs de X. r^2 nous permet d'évaluer à quel point la régression linéaire est adaptée pour décrire la distribution des locuteurs, et nous indique ainsi le pouvoir de détermination de la droite. Le coefficient est compris entre 0 et 1, et plus sa valeur est grande, plus le pouvoir de prédiction est fort. La figure 6.2 montre trois nuages de points plus ou moins dispersés sur l'axe des Y, avec les coefficients de détermination associés[3].

Toutefois, étant donné que nous avons peu d'observations, il est risqué de se contenter d'un test paramétrique. Aussi nous proposons également un test non-paramétrique, basé sur les rangs d'observation, et donc indépendant de la distribution des données. Il s'agit du ρ de Spearman. Ce coefficient nous permettra notamment

2. source :https://support.minitab.com/fr-fr/minitab/18/help-and-how-to/statistics/basic-statistics/how-to/correlation/interpret-the-results/key-results/

3. source :http://work.thaslwanter.at/BSA/html/Fitting.html

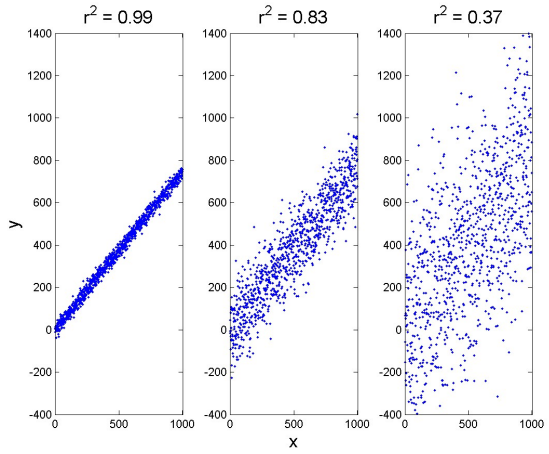
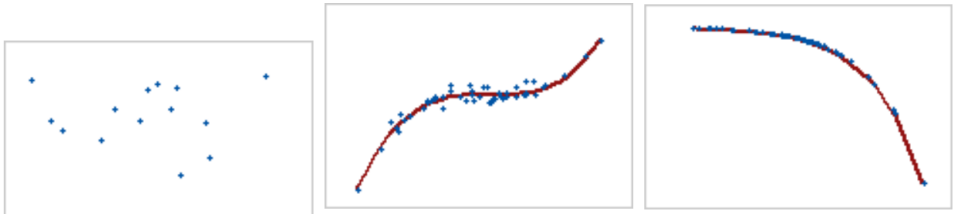


FIG. 6.2: *Coefficients de détermination en fonction de leur distribution sur Y*



(a) $\rho = 0$, *corrélation nulle*

(b) $\rho = 0,948$, *forte corrélation positive*

(c) $\rho = 1,0$, *corrélation parfaite*

FIG. 6.3: *Nuages de points avec différentes relations monotones*

de caractériser une relation non linéaire, où les deux variables n'évoluent pas forcément à la même vitesse. Il est aussi compris entre -1 et 1, et s'interprètent de la même manière que r. La figure 6.3 présente différentes relations non-linéaires[4].

Les tests de corrélation ont été faits avec le logiciel R (v3.4.4), par la fonction `cor.test(X,Y,method="<type>")`, où X et Y sont les listes de scores rythmiques et de notes des étudiants (elles doivent être de même longueur car couplées, mais pas forcément ordonnées), et <type> correspondant à pearson ou spearman.

Dans le cas du coefficient de détermination, nous avons d'abord effectué une régression linéaire avec la fonction `lm(X ~ Y)`, puis nous avons récupéré le coefficient et la p-value en faisant un `summary()` du résultat.

4. source :https://support.minitab.com/fr-fr/minitab/18/help-and-how-to/statistics/basic-statistics/how-to/correlation/interpret-the-results/key-results/

Troisième partie

Résultats & discussion

Chapitre 7

Scores rythmiques

Dans ce chapitre, nous présentons les log-vraisemblances obtenues par les locuteurs des différents corpus de test avec les deux modèles que nous avons conçus. Lorsque la log-vraisemblance est calculée pour un locuteur spécifique, nous parlons de score. Dans un premier temps, nous avons calculé les scores des deux corpus de test du CEFC pour le modèle UBM-GMM construit à partir des coefficients MFCC, afin de voir si, tels que la plupart des modèles de RAL sont conçus, les locuteurs non-natifs se distinguent des natifs. Nous avons ensuite calculé les scores pour le modèle du rythme. Dans un troisième temps, nous présentons le score rythmique des apprenants du corpus japonais.

Certaines log-vraisemblances obtenues avec le modèle du rythme sont étrangement basses et nous avons tenté d'en identifier les raisons dans la section 4. Enfin, nous discutons des résultats obtenus dans la dernière section.

7.1 Scores pour le modèle UBM-GMM

Le modèle UBM-GMM a été calculé sur 10 millions de trames choisies aléatoirement parmi les 37 millions de la partition d'apprentissage. Les besoins en mémoire vive du serveur pour calculer 1024 gaussiennes sur des vecteurs de 39 paramètres sont tels, qu'il n'a pas été possible de monter au-delà. Toutefois, 10 millions de trames ont suffi pour modéliser convenablement la population d'apprentissage. Pour nous en assurer, nous avons calculé 2 modèles parallèles avec des trames différentes. L'apprentissage de chaque modèle s'est fait sur un peu plus de 30h. La partition de test natifs est constituée de 3,6 millions de trames de 203 locuteurs ; celle des locuteurs non-natifs du CEFC de 600 mille trames de 49 locuteurs.

Dans un premier temps, nous avons calculé la log-vraisemblance moyenne des deux corpus. Pour le premier modèle, le score moyen du corpus natifs est de -98,18 et celui du corpus non-natifs est de -97,27. Les scores des deux échantillons sont très proches, et la vraisemblance est même légèrement plus élevée pour les vecteurs non-natifs, mais rien ne nous permet de dire si la différence est significative ou non. On observe la même chose avec le second modèle : -98,96 pour les natifs et -98,06 pour les non-natifs (*cf.* tableau 7.1).

Partition	Modèle 1	Modèle 2	Nombre de vecteurs	Nombre de locuteurs
CEFC natifs	-98,18	-98,96	3,6M	203
CEFC non-natifs	-97,27	-98,06	600k	49

Tab. ***7.1:*** *Log-vraisemblances moyennes de deux modèles appris sur les MFCC pour les vecteurs des partitions de test du CEFC*

Nous avons ensuite calculé la log-vraisemblance moyenne de chaque locuteur. Les scores vont de -84,86 à -113,44 pour les natifs et de -87,05 à -105,31 pour les non-natifs. L'écart plus large pour les natifs s'explique certainement par le plus grand nombre d'observations ; néanmoins les non-natifs ne semblent pas avoir des scores particulièrement inférieurs aux natifs. La figure 7.1 présente la projection de chaque score en fonction du nombre de trames par locuteur et du corpus, ainsi que la distribution des scores. Les scores natifs et non-natifs apparaissent complètement mélangés, les premiers peut-être même légèrement inférieurs aux seconds. Cette projection nous permet de constater que les locuteurs non-natifs n'ont pas un score inférieur à celui des natifs avec ce modèle, et que le nombre de trames par locuteur n'impacte pas le score obtenu.

Pour vérifier cette observation de manière statistique, nous recourons au test des rangs de Wilcoxon-Mann-Whitney. Avec l'ensemble des locuteurs pour chaque partition, on obtient un risque d'erreur de 0,678 ($U_{NN} = 4763$). H_1 est donc rejetée : les scores non-natifs ne sont pas significativement inférieurs à ceux des natifs. Nous avons également effectué le test avec le même nombre de locuteurs dans les deux corpus, en prélevant aléatoirement les scores de 50 natifs. Nous avons réitéré 10 fois le processus pour s'assurer de la régularité des observations, et avons obtenu une p-value oscillant de 0,370 ($U_{NN} = 1273$) à 0,952 ($U_{NN} = 988$). On constate que la p-value n'est jamais significative, on retiendra toujours H_0.

Il est surprenant de ne pas observer plus de différence entre les scores natifs et non-natifs avec un modèle appris sur des coefficients cepstraux, quand on sait que ces derniers sont une représentation de la prononciation au niveau segmental.

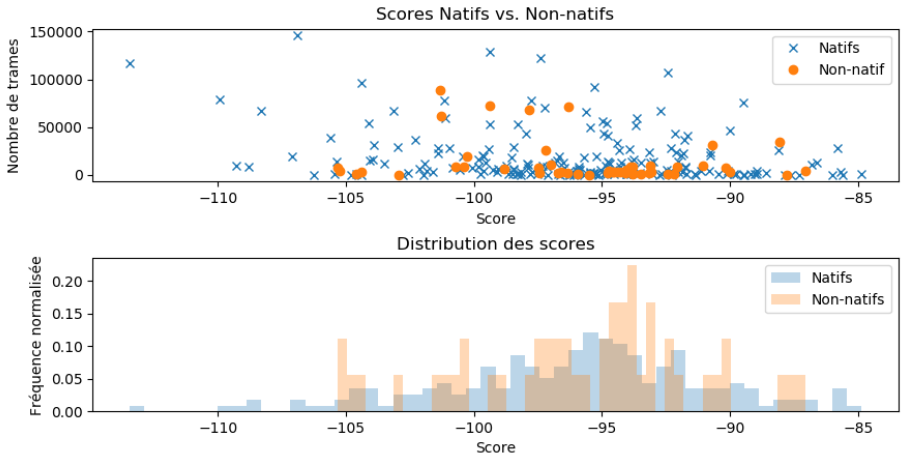


Fig. *7.1: Projection des scores par locuteur pour le modèle UBM-GMM, en fonction du nombre de trames et du statut du français des locuteurs*

7.2 Scores pour le modèle du rythme

Le modèle du rythme a été appris sur l'ensemble des 16 884 segments de la partition d'apprentissage (1 340 locuteurs), soit un minimum de 140 heures de paroles continue si l'on concatène tous les segments. La partition de test des locuteurs natifs totalise 1 919 segments (soit au moins 17h pour 146 locuteurs), et celle des non-natifs 268 segments (soit 2h pour 37 locuteurs). La log-vraisemblance moyenne pour la partition des natifs est 18,48 et 20,69 pour les non-natifs. Encore une fois, la valeur est plus élevée pour les non-natifs que pour les natifs. Toutefois, si l'on considère les scores de chaque locuteur individuellement, on peut voir que 5 scores natifs sont étrangement bas, et que la majorité des autres scores se trouve entre 0 et 40. Les scores natifs inférieurs à -50 sont les suivants :

- -74.33 pour `iljBB1r_iljBB1-M` (30 segments) ;
- -129.76 pour `Raphael_Lariviere_H_23_7e_Sonia-Branca-Rosoff-F` (2 segments) ;
- -462.68 pour `ilpBM1r_ilpSS0-F` (1 segment) ;
- -743.57 pour `ilpBM1r_ilpGV0-M` (1 segment) ;
- -858.67 pour `OF1_SeanceTravail_4dec07_L2-F` (9 segments).

Seuls deux locuteurs sont dans ce cas parmi les non-natifs :

- -57.83 pour `V_Rint_03_P5_EE-F` (1 segment) ;
- -110.92 pour `V_Rint_exo_06_P2_E-F` (1 segment).

Nous tentons d'identifier la raison de ces scores dans la section 7.4. Si l'on calcule à nouveau les log-vraisemblances sans ces locuteurs, on obtient alors 25,0 pour les natifs et 21,48 pour les non-natifs. Dans ce cas, on constate que les vecteurs des natifs sont effectivement plus proches du modèle du rythme que ceux des non-natifs. Le tableau 7.2 présente les log-vraisemblances moyennes en fonction des échantillons. La figure 7.2 donne la projection des scores individuels lorsqu'ils sont supérieurs à -50.

Partition	Log-vraisemblance	Nombre de vecteurs	Nombre de locuteurs
CEFC natifs	18,48	1 919	146
CEFC non-natifs	20,69	268	37
CEFC natifs-50	25	1 876	141
CEFC non-natifs-50	21,48	266	35

Tab. *7.2: Log-vraisemblances moyennes du modèle du rythme pour les vecteurs des partitions de test du CEFC (natifs et non-natifs tous locuteurs, et natifs et non-natifs sans les scores < -50)*

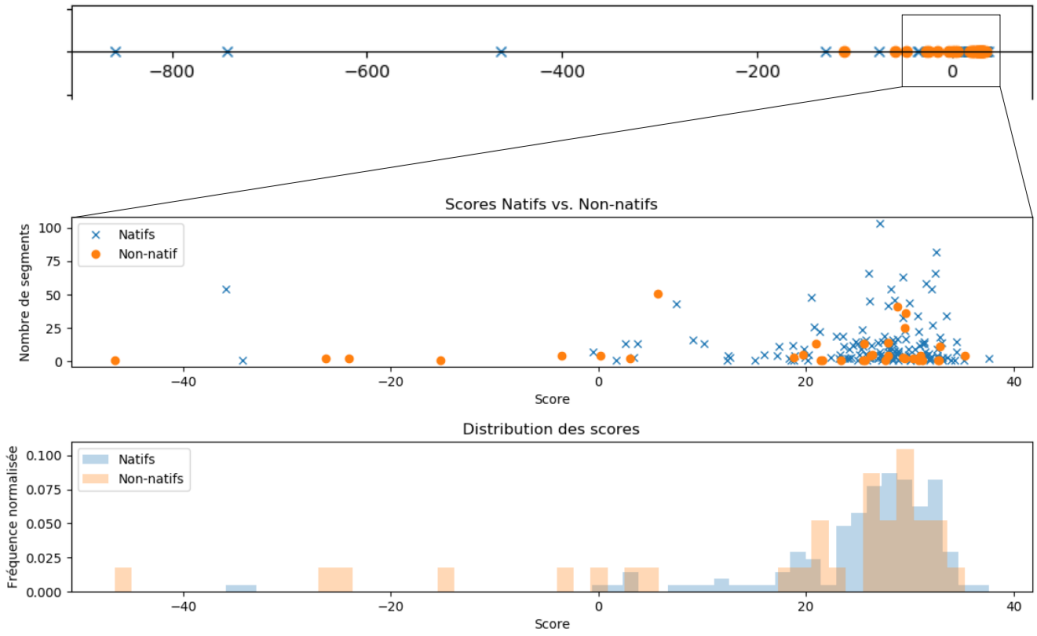


Fig. *7.2: Projection des scores par locuteur pour le modèle du rythme, en fonction du nombre de segments et du statut du français*

On constate que le score des non-natifs semble plus étalé que celui des natifs, et si beaucoup se situent entre 20 et 40, là où se concentre la plupart des scores natifs,

de nombreux scores sont aussi inférieurs à 20. Cela reste toutefois difficile à percevoir étant donné la disproportion des deux échantillons. Ces résultats sont cohérents si l'on considère que les locuteurs non-natifs ont probablement des niveaux de français hétérogènes, avec des locuteurs en cours d'apprentissage d'un côté, et des locuteurs experts de l'autre. De plus, les locuteurs ont des langues maternelles différentes, dont le rythme varie plus ou moins par rapport au français.

Le test de Wilcoxon-Mann-Whitney sur l'ensemble des locuteurs (146 natifs et 45 non-natifs) nous confirme la significativité de la différence à 10% (p-value de 0,067 ($U_{NN} = 3132$)). Il s'agit donc seulement d'une tendance selon laquelle les non-natifs ont un score moins élevé que les natifs. Sur 10 itérations avec un échantillon aléatoire de 50 locuteurs natifs, la p-value varie entre 0,027 ($U_{NN} = 1150$) et 0,437 ($U_{NN} = 944$). Selon l'échantillon de natifs sélectionné, la différence peut ne pas être significative. Il s'agit sûrement de l'influence du score très bas de certains natifs cités plus haut.

7.3 Scores rythmiques du corpus japonais

Dans cette section, nous présentons le score rythmique des apprenants du corpus japonais. Tous les locuteurs présentent entre 1 et 9 segments de parole, sur 96 segments au total.

Ici encore, on observe quelques scores très bas qui font tendre la log-vraisemblance globale vers une valeur ridiculement petite (-190,88). L'analyse des scores par locuteur révèle 6 locuteurs avec un score inférieur à -50 :

- -462.91 pour `l6_Risa-F` (3 segments) ;
- -705.63 pour `j6_Ruri-F` (3 segments) ;
- -708.42 pour `l4_Masaya-M` (1 segment) ;
- -1 101.02 pour `l4_Mikoto-F` (1 segment) ;
- -1 294.86 pour `j2_Tsubasa-M` (1 segment) ;
- -12 544.14 pour `j5_Kaho-F` (1 segment).

Sans ces 6 locuteurs, la log-vraisemblance globale monte à 0,74. La figure 7.3 montre la projection des scores par locuteur supérieurs à -50. En filigrane on peut également voir le score des locuteurs du CEFC que nous avons vus précédemment. Les scores des japonophones sont compris entre -26,85 et 22,48. Nous avons également projeté le score de l'enseignant, francophone natif, qui a un score de 19,96.

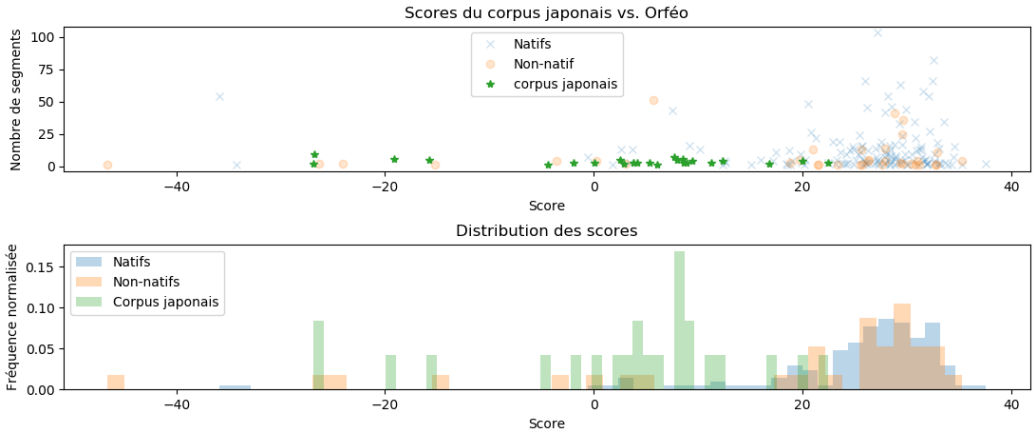


FIG. *7.3: Projection des scores par locuteur du corpus japonais pour le modèle du rythme (les 5 locuteurs <-50 ne sont pas affichés ; en filigrane les scores des deux partitions de test natifs et non-natifs du CEFC)*

On peut voir ici que les scores sont majoritairement concentrés entre 0 et 10. Trois locuteurs ont un score assez élevé : `j7_Hibiki-F` (22,48), `j9_Romain-M`, l'enseignant (19,96), et `j1_Akane-F` (16,77). Quatre locuteurs ont un score assez bas par rapport aux autres : `j8_Akihiro-M` (-15,75), `l5_Haruna-F` (-19,10), `l2_Hayato-M` (-26,74) et `j3_Sho-M` (-26,91).

Le test de Wilcoxon-Mann-Whitney valide l'hypothèse selon laquelle les locuteurs natifs de test du CEFC ont de meilleurs scores que les apprenants japonais avec une p-value de $8{,}47e^{-14}$ ($U_{NN} = 4064$) avec l'ensemble des locuteurs, soit une différence fortement significative. Sur 10 itération avec 30 natifs aléatoires, la p-value varie entre $8{,}88e^{-11}$ ($U_{NN} = 882$) et $4{,}76e^{-6}$ ($U_{NN} = 750$), fortement significative dans tous les cas.

7.4 Concernant les scores inférieurs à -50

Nous avons tenté d'identifier les raisons pouvant être à l'origine des scores particulièrement bas. Nous avons d'abord écouté les enregistrements des locuteurs concernés, pour vérifier la qualité du son, puis nous avons observé le détail des mesures de chaque paramètre rythmique.

L'écoute d'extraits de parole des locuteurs dont le score est très bas montre soit que la qualité du son en général est mauvaise (`OF1_SeanceTravail_4dec07_L2-F` (-858,67), réunion de bureau, parole assez difficile à distinguer ; `ilpBM1r_ilpSS0-F` (-462,68), son étouffé comme si le microphone était couvert), soit que la voix du lo-

cuteur en question est très basse (`V_Rint_exo_06_P2_E-F` (-110,92) ; `j5_Kaho-F` (-12 544,14) ; `j6_Ruri-F` (-705,63)).

Pourtant tous les locuteurs dont la voix est basse n'ont pas nécessairement un score faible. C'est le cas notamment de `l3_Shione-F` (8,87) ou `j6_Chihiro-F` (6,05). Idem pour la qualité du son, des enregistrements comme ceux de `j7_Hibiki-F` (22,48) ou `potins_famille_L2-F` (35,28) peuvent être particulièrement bruités et toutefois obtenir des scores très élevés.

La relation entre qualité d'enregistrement et score rythmique n'est donc pas forcément évidente, mais la voix de certains locuteurs a probablement provoqué des erreurs de mesures. On se rend compte par exemple que les pourcentages de parole *(pcdV)* mesurés dans le segment de `j5_Kaho-F` (-12 544,14) et le segment de `j2_Tsubasa-M` (-1 294,86) sont à 0,26% et 0,97% quand la médiane sur les 96 segments du corpus est à 16%. Sur ces mêmes segments, le débit de parole est respectivement de 0,06 et 0,11 syllabes par seconde (médiane à 1,34). On observe également de gros décalages dans les durées d'intervalles non-voisés : 14,93 et 4,30 secondes pour leur durée moyenne (μdU) (médiane à 0,58), et de mêmes décalages importants pour $\mu dVU, \sigma dU$ et σdVU.

Bien que certains segments présentent des valeurs plus incohérentes que d'autres, on constate que l'ensemble des segments pour un même locuteur ont tendance à se ressembler, il ne s'agit donc vraisemblablement pas d'erreurs locales de calcul. Il ne s'agit a priori pas non plus des caractéristiques de l'enregistrement, puisque ces derniers proviennent tous de deux uniques enregistrements `j` et `l`. Nous en concluons que c'est principalement le volume de la voix du locuteur qui est à l'origine de ces mesures biaisées.

Les longues durées d'intervalles non-voisés doivent être la cause d'une mauvaise détection du voisement pour ces locuteurs. Lorsqu'on observe les textgrids de voisement de l'enregistrement `j5`, beaucoup de parole de `j5_Kaho-F` n'est pas détectée, et l'on se retrouve avec de longs intervalles non-voisés sur toute la durée de certaines UEP (*cf.* figure 7.4). La détection fonctionne pourtant avec le second locuteur `j5_Koki-M`. Il y a effectivement une différence d'intensité moyenne sur les deux UEP (45,99 dB pour `j5_Kaho-F` contre 49,71 dB pour `j5_Koki-M`). De la même manière, on constate que les noyaux syllabiques ne sont pas non plus correctement détectés pour `j5_Kaho-F`.

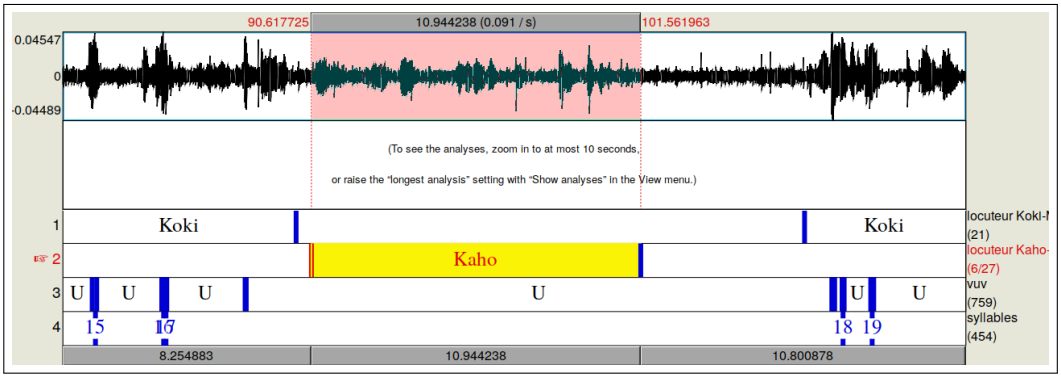


FIG. *7.4: Mauvaise détection du voisement sur une UEP du locuteur j5_Kaho-F*

7.5 Discussion

Nous avons calculé les log-vraisemblances moyennes des 2 corpus de test du CEFC avec le modèle UBM-GMM à base des MFCC, et celui appris sur les 16 paramètres rythmiques. Il s'avère que l'hypothèse selon laquelle les locuteurs non-natifs obtiennent un score inférieur à celui des natifs (et sont donc plus éloignés du modèle) n'est vérifiée que pour le modèle du rythme, et qu'il s'agit seulement d'une tendance ($p < 0{,}01$). Dans le cas du corpus d'apprenants japonais, la différence entre les apprenants et les mêmes locuteurs natifs est bien plus significative ($p < 0{,}0001$).

Nous nous sommes confrontés au fait que les locuteurs non-natifs du CEFC peuvent être de niveau très avancé, et avoir un rythme comparable à celui des locuteurs natifs. Nous obtenons donc des résultats très hétérogènes pour cette population (entre -46,57 et 35,28 après filtrage). En revanche, les scores du corpus d'apprenants japonais, dont tous les niveaux sont compris entre A2 et B1, sont répartis sur un intervalle plus restreint (entre -26,85 et 22,48 après filtrage).

Certains scores se sont avérés étrangement bas (13 inférieurs à -50 sur 213 locuteurs de test, toutes partitions confondues). Les mesures prosodiques effectuées sur ces locuteurs présentent des durées d'intervalles non-voisés excessivement longues, qui semblent s'expliquer par une voix dont l'intensité n'a pas permis à l'algorithme de détection de voisement de fonctionner correctement ; le locuteur ne parlait pas assez fort. Il aurait peut-être été judicieux de normaliser l'intensité de chaque locuteur avant d'exécuter la détection de voisement, ou bien d'exécuter le script indépendamment sur chaque UEP[1].

Ces scores inférieurs à -50 restent toutefois minoritaires, et de manière générale,

1. En effet, le script effectue un *To PointProcess (periodic, cc)* sur l'ensemble de l'enregistrement (du binôme). Exécuté sur une UEP seule, le voisement pourrait être alors correctement détecté.

le modèle du rythme semble correctement discriminer les natifs des non-natifs. Toutefois, étant donné les disparités de conditions d'enregistrement, certains paramètres sont probablement moins efficaces que d'autres. Dans le chapitre suivant, nous mesurons l'efficacité de chaque paramètre, pour identifier ceux qui expliquent le plus la différence entre les locuteurs natifs et non-natifs.

Chapitre 8

Analyse des paramètres

Dans ce chapitre, nous comparons d'abord les distributions des mesures de chaque paramètre en fonction des corpus des locuteurs (natifs, non-natifs du CEFC et apprenants japonais). Dans un second temps, nous présentons les résultats du test des êta-carrés (η^2).

8.1 Distribution des mesures

Nous disposons d'un total de 21 928 segments de parole sur l'ensemble du CEFC, tous corpus confondus (apprentissage, test natifs et test non-natifs), provenant de 1777 locuteurs différents[1]. 20 627 segments proviennent de locuteurs natifs et 432 de locuteurs non-natifs. Les 869 autres segments appartiennent à des locuteurs dont le statut du français est inconnu, et sont donc ignorés. Du côté du corpus des apprenants japonais, nous disposons de 96 segments provenant de 29 locuteurs.

Chaque segment se présente sous la forme de 16 mesures – une pour chaque paramètre rythmique. Dans cette partie, nous proposons de comparer la distribution des mesures pour chaque paramètre, en fonction de trois types de locuteurs : les francophones natifs du CEFC, les non-natifs de langues maternelles diverses du CEFC, et les non-natifs japonophones du corpus d'apprenants japonais.

Chaque échantillons sera constitué de 96 segments, soit le nombre de segments dont nous disposons pour le corpus japonais[2].

1. 2 435 locuteurs, auxquels ont été enlevés 658 locuteurs ne parlant pas suffisamment (durée de parole inférieure à 30 secondes) pour atteindre un segment.

2. Précisons que les segments sont sélectionnés aléatoirement, on se retrouve donc avec une grande diversité de locuteurs, en particulier pour les francophones natifs (la probabilité que 2 segments appar-

Les figures 8.1 à 8.6 présentent la distribution des mesures pour chaque échantillon, sous la forme de boxplots pour faciliter la comparaison.

D'une manière générale, on constate que les mesures des non-natifs du CEFC sont plus proches de celles des natifs, que de celles des apprenants japonais. Les différences les plus flagrantes sont celles du débit de parole $(SpeechRate)$ et du pourcentage de voisement ($pcdV$, *cf.* figure 8.1). On constate sans surprise que les locuteurs du corpus japonais parlent beaucoup plus lentement que les autres locuteurs (1,3 syllabes par seconde en moyenne, contre 3,9 pour les non-natifs du CEFC et 4,0 pour les natifs). Leur pourcentage de voisement est également beaucoup plus faible, à cause de longs silences qui parsèment leurs segments de parole (15% pour les japonais, 58% pour les non-natifs et 57% pour les natifs).

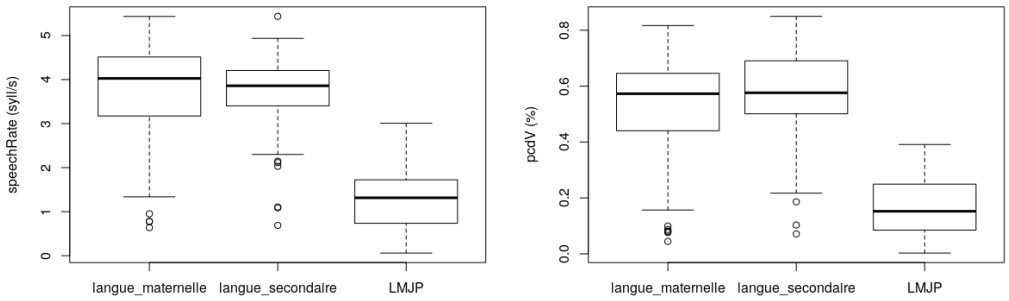


FIG. 8.1: *Distributions des mesures de débit de parole (à gauche) et de pourcentage de voisement (à droite) sur 96 segments par échantillon (langue_maternelle : natifs du CEFC ; langue_secondaire : non-natifs du CEFC ; LMJP : apprenants japonophones)*

On constate que la durée moyenne des intervalles voisés (μdV) est également beaucoup plus faible pour les apprenants japonais : 0,13 seconde, contre 0,21 s. pour les non-natifs du CEFC et 0,19 s. pour les natifs (*cf.* figure 8.2). L'écart type des durées (σdV) suit le même schéma : 0,09 s. pour les apprenants japonais, 0,19 s. pour les non-natifs du CEFC et 0,18 s. pour les natifs. On constate ici que pour les 2 paramètres, les non-natifs du CEFC se situent à l'opposé des apprenant japonais, par rapport aux natifs. Si l'on effectue un nouvel échantillonnage, on constate toutefois que ce n'est plus le cas, mais la différence entre natifs et non-natifs est infime (0,22 s. contre 0,20 s. pour la moyenne et 0,21 s. contre 0,20 s. pour l'écart type,

tiennent au même locuteur est très faible). Il a initialement été testé un rééchantillonnage au niveau des locuteurs, afin d'avoir 29 locuteurs pour chaque échantillon, mais l'échantillon de natifs était constamment beaucoup plus important que celui des non-natifs, qui ont tendance à avoir moins de segments de parole.

respectivement). La différence entre les apprenants japonais et les locuteurs du CEFC est réduite après normalisation par rapport au débit de parole, avec le coefficient de variation ($VarcoV$), mais reste toujours importante (0,70 contre 0,90 et 0,93).

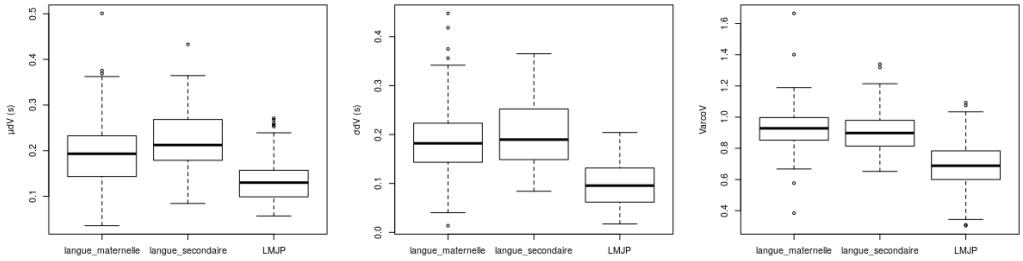


Fig. 8.2: *Distributions des mesures de la moyenne (gauche), de l'écart type (milieu) et du coefficient de variation des durées des intervalles voisés sur 96 segments par échantillon (langue_maternelle : natifs du CEFC ; langue_secondaire : non-natifs du CEFC ; LMJP : apprenants japonophones)*

En ce qui concerne les métriques faisant intervenir les durées d'intervalles non-voisés (figures 8.3 et 8.4), on s'aperçoit que quelques durées moyennes sont étrangement longues sur des segments du corpus japonais, notamment une montant à 15 secondes. Il s'agit du segment de `j5_Kaho-F`, pour qui très peu d'intervalles de voisement ont été détectées (*cf.* chapitre précédent, section 7.4). On constate que la médiane des moyennes de durée des intervalles non-voisés est identique pour les natifs et les non-natifs du CEFC (0,15 s.) mais beaucoup plus élevée chez les apprenants japonais (0,58 s.). Même pattern pour la médiane des écarts type de durée : 0,19 s. pour le CEFC et 0,73 s. pour les apprenants japonais. Cette différence disparaît si l'on normalise par la moyenne des segments ($VarcoU$) : 1,22 pour les Japonais, 1,27 pour les non-natifs du CEFC et 1,21 pour les natifs. On constate les mêmes différences avec les paires d'intervalles voisé+non-voisé.

Les comparaisons de paires d'intervalles voisés ($r/nPVI\_dV$) présentent un pattern similaire à celui des moyennes, des écarts type ou des coefficients normalisés de durée des intervalles seuls (*cf.* figure 8.5). La différence entre natifs et non-natifs du CEFC est infime (ici, respectivement 0,17 et 0,18 pour l'indice brut et 78,54 et 79,14 pour l'indice normalisé avec le débit de parole), et un indice généralement plus faible chez les apprenants japonais (0,09 et 64,57).

Enfin, si les médianes des moyennes, écarts type et coefficients de variation des deltas intersyllabiques sont assez proches entre les trois échantillons, on peut voir que la diversité des mesures au sein des 96 segments du corpus japonais est plus grande que pour les autres échantillons. La moyenne des deltas ($\mu\Delta t$) peut notamment être

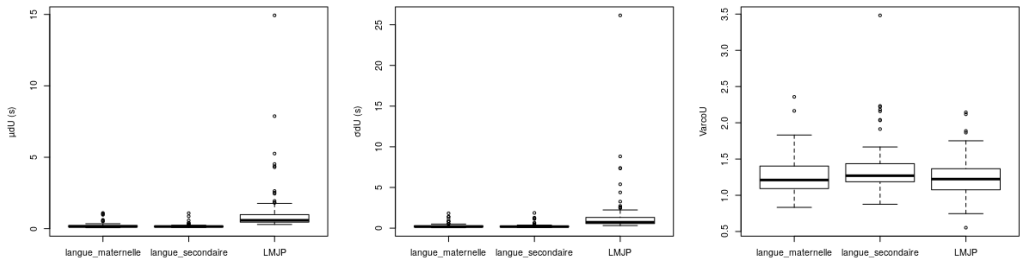


Fig. *8.3:* *Distributions des mesures de la moyenne (gauche), de l'écart type (milieu) et du coefficient de variation des durées des intervalles non-voisés sur 96 segments par échantillon (langue_maternelle : natifs du CEFC ; langue_secondaire : non-natifs du CEFC ; LMJP : apprenants japonophones)*

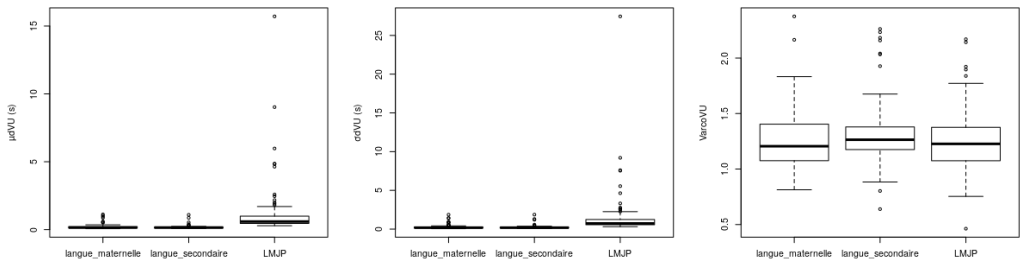


Fig. *8.4:* *Distributions des mesures de la moyenne (gauche), de l'écart type (milieu) et du coefficient de variation des durées des paires d'intervalles voisé et non-voisé sur 96 segments par échantillon (langue_maternelle : natifs du CEFC ; langue_secondaire : non-natifs du CEFC ; LMJP : apprenants japonophones)*

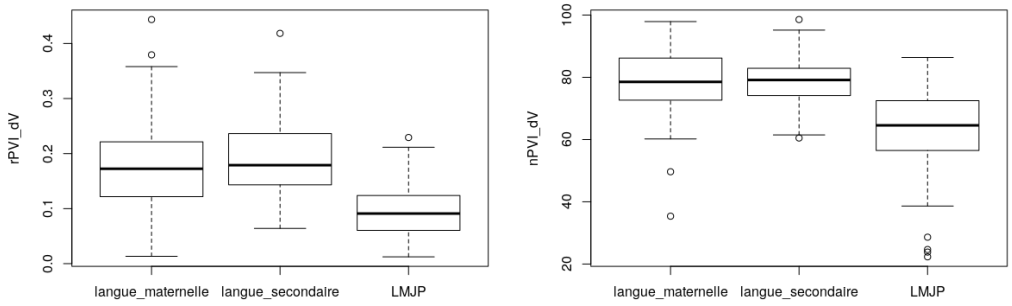


Fig. *8.5:* *Distributions des mesures du PVI brut (gauche) et normalisé au débit de parole (droite) sur 96 segments par échantillon (langue_maternelle : natifs du CEFC ; langue_secondaire : non-natifs du CEFC ; LMJP : apprenants japonophones)*

plus élevée que celle des locuteurs du CEFC, au moins 30ms de plus que le maximum observé sur les 2 autres échantillons, sans considérer les *outliers*, qui feraient passer la différence à 55ms). Dans le cas de l'écart type des deltas intersyllabiques ($\sigma\Delta t$), on trouve parfois des valeurs très réduites, donc un enchaînement de syllabes espacées dans le temps de manière très homogène sur un même segment de parole, là où les natifs font plus varier cet écart (valeur de l'écart type plus élevé). La grande variabilité dans ces mesures et le nombre important d'*outliers* chez les apprenants japonais, nous montre que ces locuteurs ont un rythme beaucoup moins fixé que les locuteurs natifs, tantôt avec des écarts intersyllabiques plus grands, tantôt avec des écarts plus petits.

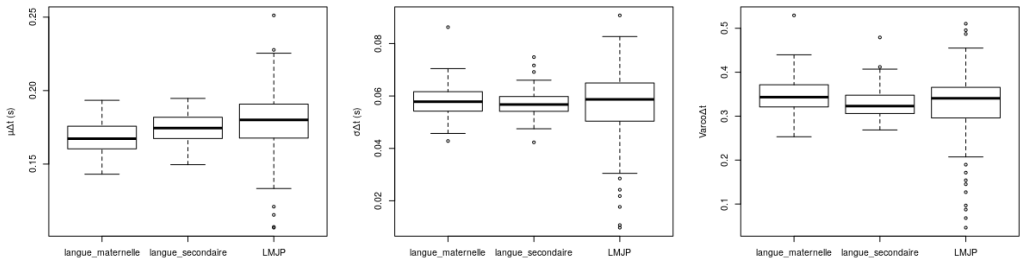


Fig. *8.6:* *Distributions des mesures de la moyenne (gauche), de l'écart type (milieu) et du coefficient de variation des deltas intersyllabiques sur 96 segments par échantillon (langue_maternelle : natifs du CEFC ; langue_secondaire : non-natifs du CEFC ; LMJP : apprenants japonophones)*

8.2 Efficacité des paramètres

L'évaluation de l'efficacité des paramètres dans la distinction rythmique entre natifs et non-natifs s'est faite avec les locuteurs non-natifs du CEFC d'une part, et avec les apprenants japonais d'autre part.

Étant donné le nombre réduit de segments dont nous disposons pour les locuteurs non-natifs (432 pour le CEFC et 96 pour les apprenants japonais), nous avons rééchantillonné les natifs à 432 segments pour la comparaison avec les non-natifs du CEFC, et 96 segments pour la comparaison avec les apprenants japonophones. Afin d'éviter tout biais de rééchantillonnage, nous avons réitéré 3 fois l'expérience pour mener le test sur des échantillons natifs différents.

Les tableaux 8.1a et 8.1b présentent résultats de la première itération. Les paramètres sont ordonnés en fonction de leur η^2. La colonne P-value indique la significativité du paramètre seul dans la différence natifs/non-natifs. Plus la p-value est faible, plus il est certain que le paramètre discrimine les locuteurs natifs et non-natifs. Plus l'η^2 est élevé, plus le fait que le locuteur soit natif ou non-natif explique une part importante de la variance globale de ce paramètre.

Sur les trois itérations avec les locuteurs du CEFC, ce sont toujours les trois mêmes paramètres qui arrivent en tête : la moyenne des durées inter-syllabiques $\mu\Delta t$ ($\eta^2 = 0,11; 0,12; 0,13$), suivi par son coefficient de variation $Varco\Delta t = \frac{\sigma\Delta t}{\mu\Delta t}$ ($\eta^2 = 0,07; 0,07; 0,06$) et le débit de parole SR ($\eta^2 = 0,05; 0,03; 0,05$). Vient ensuite avec un η^2 de $0,01$ le coefficient de variation des intervalles voisés $VarcoV$ pour la première et la deuxième itération, et $nPVI\_dV$ pour la troisième itération. Les paramètres suivants ont un η^2 inférieur à 1%.

Concernant l'efficacité des paramètres dans la distinction des natifs du CEFC et des apprenants japonais, les résultats sont très différents. Tout d'abord, seuls 4 paramètres ne sont pas significatifs dans au moins une itération sur trois ($VarcoU$, $VarcoVU$, $\sigma\Delta t$ et $Varco\Delta t$), les 12 autres ont une p-value inférieure à 0,0001 sur les trois itérations. L'ordre des paramètres en fonction de leur η^2 est identique jusqu'au dixième d'entre eux. Le débit de parole arrive en tête ($\eta^2 = 0,75; 0,67; 0,70$), suivi de près par le pourcentage de voisement $pcdV$($\eta^2 = 0,67; 0,60; 0,67$). On trouve ensuite les métriques impliquant les durées de segments voisés ($VarcoV$, σdV et μdV) ainsi que les PVI_dV, expliquant tous entre 40 et 30% de la variance globale. S'en suivent les métriques impliquant les durées de segments non-voisés et les intervalles syllabiques.

Les figures 8.7 et 8.8 sont une représentation graphique des η^2 du tableau 8.1.

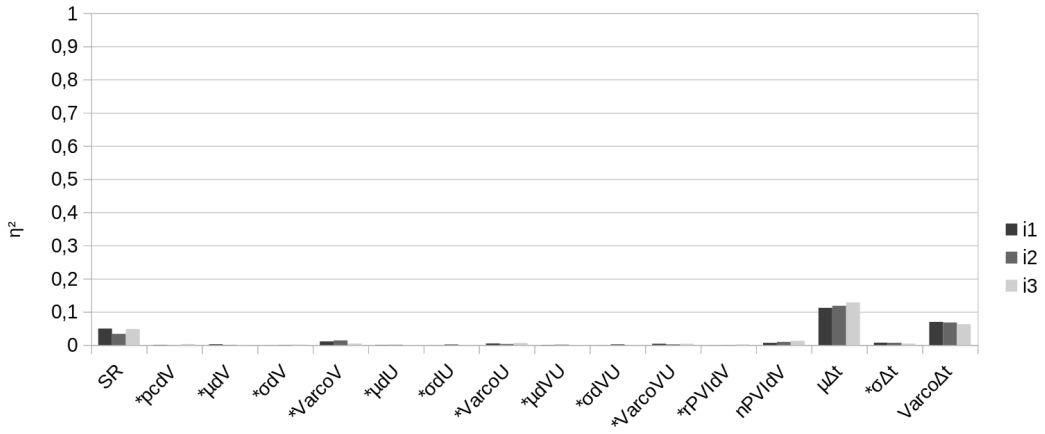


Fig. 8.7: *Valeur des η^2 entre natifs et non-natifs du CEFC, sur les trois itérations (la différence natifs/non-natifs n'est pas significative (p>0,05) pour les paramètres avec un astérisque dans au moins une itération)*

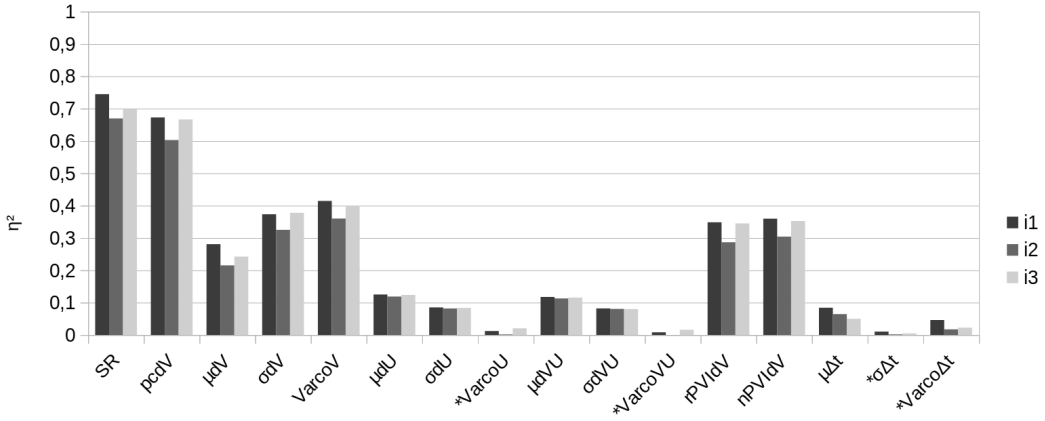


Fig. 8.8: *Valeur des η^2 entre natifs du CEFC et apprenants japonais, sur les trois itérations (la différence natifs/non-natifs n'est pas significative (p>0,05) pour les paramètres avec un astérisque dans au moins une itération)*

Paramètre	η^2	P-value	
$\mu\Delta t$	0,1121	$4{,}53e^{-24}$	<0,001
$Varco\Delta t$	0,0698	$2{,}92e^{-15}$	<0,001
SR	0,0498	$3{,}32e^{-11}$	<0,001
$VarcoV$	0,0111	0,002	<0,01
$\sigma\Delta t$	0,0073	0,012	<0,025
$nPVIdV$	0,0068	0,015	<0,025
$VarcoU$	0,0051	0,035	<0,05
$VarcoVU$	0,004	0,062	>0,05
μdV	0,0024	0,15	>0,05
$pcdV$	0,0007	0,43	>0,05
μdU	0,0007	0,439	>0,05
μdVU	0,0006	0,457	>0,05
$rPVIdV$	$8{,}93e^{-05}$	0,781	>0,05
σdU	$9{,}84e^{-06}$	0,927	>0,05
σdVU	$8{,}73e^{-06}$	0,931	>0,05
σdV	$3{,}26e^{-06}$	0,958	>0,05

(a) *η^2 entre natifs et non-natifs du CEFC (432 segments chacun)*

Paramètre	η^2	P-value	
SR	0,745	$3{,}07e^{-58}$	<0,0001
$pcdV$	0,673	$5{,}63e^{-48}$	<0,0001
$VarcoV$	0,415	$7{,}23e^{-24}$	<0,0001
σdV	0,373	$4{,}79e^{-21}$	<0,0001
$nPVIdV$	0,360	$3{,}81e^{-20}$	<0,0001
$rPVIdV$	0,349	$1{,}99e^{-19}$	<0,0001
μdV	0,281	$2{,}64e^{-15}$	<0,0001
μdU	0,125	$4{,}78e^{-07}$	<0,0001
μdVU	0,118	$1{,}11e^{-06}$	<0,0001
σdU	0,085	$3{,}94e^{-05}$	<0,0001
$\mu\Delta t$	0,084	$4{,}37e^{-05}$	<0,0001
σdVU	0,082	$5{,}46e^{-05}$	<0,0001
$Varco\Delta t$	0,046	0,003	<0,01
$VarcoU$	0,012	0,124	>0,05
$\sigma\Delta t$	0,011	0,157	>0,05
$VarcoVU$	0,008	0,206	>0,05

(b) *η^2 entre natifs du CEFC et apprenants japonais (96 segments chacun)*

TAB. *8.1:* *Résultats des η^2 sur la première itération (les segments natifs sont rééchantillonnés en fonction du nombre de segments non-natifs disponibles, les paramètres sont triés en fonction de leur pertinence dans le modèle)*

8.3 Discussion

Ces analyses ont permis de mettre en évidence le fait qu'il est difficile d'identifier les paramètres rythmiques les plus efficaces sur une population aussi hétérogène que les non-natifs du CEFC. Les niveaux de compétence en français et les langues maternelles sont trop diversifiées, et plus de la moitié des paramètres ne distinguent pas les locuteurs de manière significative, à cause de l'écart général peu important entre les natifs et les non-natifs. On retiendra malgré tout que l'appartenance à un groupe aussi hétérogène que les non-natifs du CEFC versus les natifs du CEFC se retrouve partiellement dans des paramètres rythmiques tels que la moyenne des deltas intersyllabiques (avec 10% de la variance expliquée par le groupe), leur coefficient de variation (seulement 7% de variance expliquée) et enfin le débit de parole (5% de la variance expliquée).

Dans le cas des apprenants japonais, la différence rythmique avec les natifs est beaucoup plus importante. La majorité des paramètres sont très significatifs (p<0,0001) : l'appartenance au groupe des natifs ou des japonophones explique jusqu'à 75% de la variance du débit de parole, 67% de celle du pourcentage de voisement, et plus de 25% de la variance des moyennes, écarts types et coefficients de variation des intervalles

voisés. Moins de 8% de la variance des mesures impliquant les deltas intersyllabiques est expliquée par le groupe.

Chapitre 9

Score rythmique et niveau de compétence en langue

Le score rythmique nous indique la distance qui existe entre les mesures effectuées sur un échantillon de parole et le modèle rythmique du français que nous avons mis au point. Dans ce chapitre, nous nous intéressons à la corrélation entre ce score et le niveau de compétence en langue du locuteur non-natif.

Pour chacun des locuteurs du corpus d'apprenants japonais, nous disposons de trois notes. La première est la note globale à l'examen de fin de semestre sur 100 points, qui est en réalité la somme de 4 notes : celle de la compréhension orale, celles de la compréhension écrite, de la production orale et enfin celle de la production écrite. La deuxième note est celle de la production orale de ce même examen (sur 25 points). La dernière note correspond aux 5 points attribués à l'aisance à l'oral (clarté et fluidité), constituant $\frac{1}{5}$ de la note de PO. Ces deux dernières notes ont été données par l'enseignant à partir des productions orales du corpus.

Nous avons calculé la corrélation entre le score rythmique des apprenants et leur note globale d'une part, ainsi qu'avec leur note de production orale d'autre part.

Pour l'ensemble des calculs de cette section, nous avons ignoré les locuteurs dont le score est inférieur à -50. En-dessous de ce seuil, nous considérons que les mesures acoustiques sont incohérentes, à causes de voix trop faibles ne permettant pas une bonne détection des intervalles de voisement et des noyaux syllabiques (*cf.* chapitre 7 section 7.3). Six locuteurs ne sont donc pas pris en compte ; leur score va de -462,91 à -12 544,14. Le corpus est donc réduit à 23 locuteurs.

9.1 Score rythmique & niveau global

La figure 9.1 présente les scores rythmiques en fonction de la note globale à l'examen de fin de semestre. On constate que les valeurs sont généralement corrélées : plus la note globale est élevée, plus le score rythmique est élevé aussi. Quelques locuteurs comme `l3_Aika-F` ont toutefois un score rythmique élevé (11,29) alors que son niveau global est relativement bas par rapport aux autres apprenants (58,5). Le cas inverse semble moins présent : tous les locuteurs ayant plus de 70 points à l'examen ont un score supérieur à 0, excepté `j8_Moeno-F` (rythme : -4,35 ; score global : 74).

Nous avons calculé plusieurs coefficients de corrélation : celui de Pearson est à 0,598, le coefficient de détermination à 0,358, et le ρ de Spearman à 0,478. Le test de Pearson et le coefficient de détermination sont significatifs à 1% et celui de Spearman à 5%. La colonne de gauche du tableau 9.1 indique la valeur de chaque coefficient et la p-value associée.

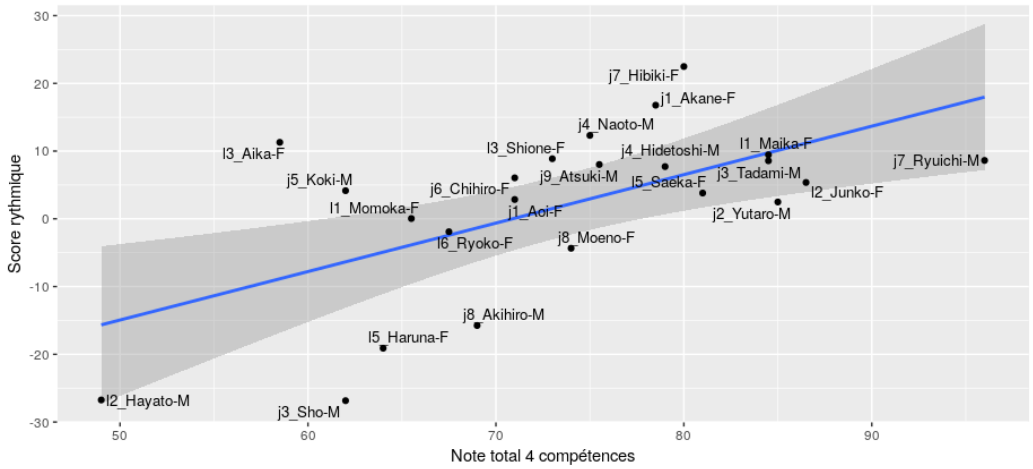


Fig. 3.1: *Score rythmique en fonction du niveau global (note sur 4 compétences : production orale et écrite, expression orale et écrite)*

	4 comp. / rythme	PO / rythme	PO-aisance / ryhtme
Pearson (r)	0,598 (p-val. 0,003)	0,257 (p-val. 0,237)	0,410 (p-val. 0,052)
coef. détermination (r^2)	0,358 (p-val. 0,003)	0,066 (p-val. 0,237)	0,168 (p-val. 0,052)
Spearman (ρ)	0,478 (p-val. 0,021)	0,315 (p-val. 0,144)	0,228 (p-val. 0,295)

Tab. 3.1: *Résultats des tests de corrélation entre le score rythmique et la note globale à l'examen (à gauche), la note de production orale (au milieu) ou à la note d'aisance de parole (à droite)*

9.2 Score rythmique & note de production orale

La corrélation entre le score rythmique et la note de production orale n'est significative pour aucun test. Comme on peut le voir sur la figure 9.2, certains apprenants comme `l2_Hayato-M` ont un score rythmique assez bas (-26,74) et pourtant une note de 21/25 en production orale. D'autres, comme `l3_Aika-F`, ont un score rythmique plutôt élevé (11,29) pour une note de PO plus faible que la plupart des apprenants (18/25). Les autres tests de corrélation se sont également révélés non-significatifs, mais les deux coefficients sont positifs (0,257 pour Pearson et 0,315 pour Spearman, *cf.* tableau 9.1).

La note de PO ne dépend pas seulement du niveau d'accent étranger, mais également de la faculté à parler de soi, ou à restituer des informations (*cf.* chapitre 4 section 4.2.4), cela peut expliquer le fait que la note ne soit pas directement corrélée avec le score rythmique. Toutefois, la non-significativité des tests est probablement due à la fourchette de note très restreinte (entre 17 et 24 sur 25).

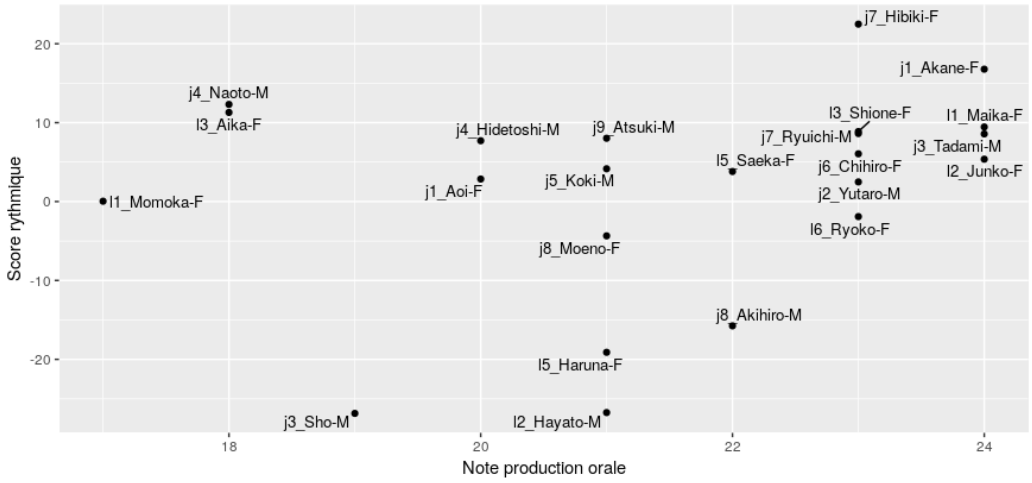


Fig. 3.2: *Score rythmique en fonction du niveau de production orale*

9.3 Score rythmique & note d'aisance à l'oral

Nous avons également examiné la corrélation entre le score rythmique et la note d'aisance à l'oral. Malheureusement aucun test ne s'est révélé significatif ici non-plus. La figure 9.3 présente les scores rythmiques en fonction de cette note sur 5 points. On voit que 20 locuteurs sur 23 ont soit 4, soit 5 points. La locutrice `l3_Aika-F` a un score rythmique peut cohérent avec la note attribuée par l'enseignant. Cette

note devrait pourtant être légitimement corrélée avec le score rythmique, puisqu'elle évalue la clarté et la fluidité de la parole de l'apprenant. On peut émettre l'hypothèse d'une évaluation trop rapide, mais d'autres facteurs viennent peut-être influencer le score rythmique. De manière générale, le peu de locuteurs dont nous disposons et la fourchette de notation extrêmement réduite (entre 3 et 5) nous permet difficilement de faire des statistiques significatives.

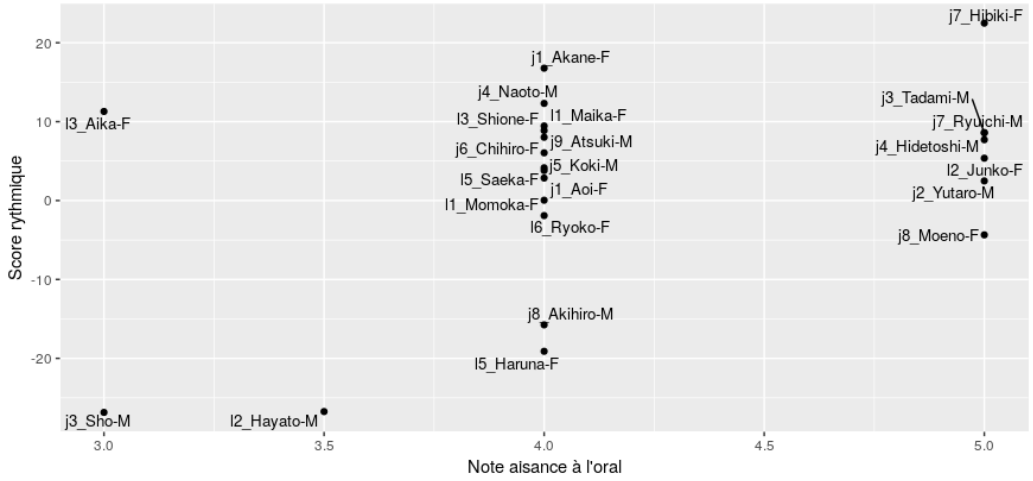


Fig. *3.3:* *Score rythmique en fonction du niveau d'aisance en production orale*

9.4 Discussion

Notre expérience a montré que le score rythmique est globalement corrélé avec le niveau de compétence général des apprenants ($r = 0,598; p < 0,01$). Cependant il n'a pas été possible d'évaluer précisément la corrélation avec la note de production orale et celle d'aisance à l'oral. Nous devons réitérer l'expérience avec plus de locuteurs, et peut-être avec une grille d'évaluation plus fine pour la production orale.

En effet, le nombre limité d'observations et les fourchettes réduites d'évaluation de l'oral permettent difficilement de faire des tests de corrélation significatifs. On constate en effet que beaucoup d'apprenants se sont vus attribuer 23 points en PO (6 sur 29 étudiants), et la majorité des notes se concentrent entre 21 et 24 (23 sur 29). Dans le cas de l'aisance à l'oral, toutes les notes varient entre 3 et 5.

Conclusion & perspectives

Nous nous sommes intéressés dans ce mémoire à la place du rythme dans la perception de l'accent étranger, et à sa modélisation informatique.

Les récentes études sur la perception de l'accent étranger ont mis en évidence l'importance de la prosodie, et notamment de l'intonation et des durées de segments. Ces paramètres dépendent d'une part de la langue maternelle du locuteur – chaque langue disposant d'un schéma rythmique qui lui est propre –, mais également du locuteur lui-même ainsi que de la situation d'énonciation. Les étudier implique donc de travailler sur de gros corpus diversifiés.

L'évaluation automatique de la prononciation en langue étrangère a tantôt recourt aux systèmes de reconnaissance automatique de la parole, en comparant ce qui est produit avec une référence ; tantôt à la modélisation de phonèmes ou de syllabes cibles, sélectionnés en fonction des difficultés observées chez les apprenants.

Nous avons proposé de modéliser le rythme du français dans sa globalité, et dans sa variation. Cette modélisation s'est faite à travers 16 dimensions, chacune représentant une mesure de durée de segment. Ces mesures sont inspirées de la littérature, tant dans le domaine de la classification rythmique des langues, que dans l'évaluation automatique de la fluence. Elles ont été choisies de manière à ne pas nécessiter de transcription ni de système d'ASR, nous nous sommes donc basés essentiellement sur les durées de voisement et les écarts intersyllabiques, qui peuvent tous être détectés automatiquement et uniquement à partir du signal de parole. La modélisation s'est faite par l'apprentissage d'un mélange gaussien décrivant les lois de densité de probabilité de ces 16 mesures, sur les enregistrements de 1 340 locuteurs natifs issus du CEFC, de différentes régions francophones et dans diverses situations d'énonciation.

Ces 16 paramètres rythmiques ont permis de distinguer les locuteurs natifs et non-natifs. Dans le cas des locuteurs non-natifs du CEFC, aux niveaux et aux langues maternelles hétérogènes, il ne s'agit que d'une tendance ($p = 0{,}067$), mais la différence est très significative avec les apprenants du corpus japonais, tous japonophones de niveau A2 ($p < 0{,}0001$).

En ce qui concerne l'efficacité des paramètres, les tests effectués sur les locuteurs du CEFC ont permis de mettre en évidence l'importance des durées intersyllabiques moyennes ($\eta^2 = 11\%$), de leur coefficient de variation (7%), et du débit de parole (5%) ; chacun de ces paramètres étant très significatifs ($p < 0{,}0001$). Les 13 autres paramètres expliquent 1% ou moins de la variation, et beaucoup ne sont pas significatifs.

Lorsque l'on calcule l'efficacité des paramètres avec les apprenants japonais, les

paramètres sont nettement plus significatifs. 13 d'entre eux expliquent entre 4 et 75% de la variation, avec en tête le débit de parole (75%), le pourcentage de voisement (67%), le coefficient de variation des durées de segments voisés (42%), l'écart type de leur durée (37%), ou encore les comparaisons de paires successives d'intervalles voisés normalisées ou non avec le débit de parole (36% et 35%). Seuls 3 paramètres sur 16 ne sont pas significatifs, il s'agit des coefficients de variation des durées de segments non-voisés, et des paires voisé+non-voisé, ainsi que de l'écart type des durées intersyllabiques.

Les théories sur le rythme des langues nous mènent à penser que les paramètres pertinents pour évaluer le rythme dépendent de la langue maternelle des apprenants. Le peu de locuteurs non-natifs par langue maternelle dans le CEFC ne nous permet malheureusement pas de mettre en évidence les paramètres pertinents en fonction des langues. Toutefois, nous avons là une cartographie des difficultés rythmiques observées sur nos 29 apprenants japonophones du français. Les figures 9.4 et 9.5 proposent une représentation graphique de ces difficultés.

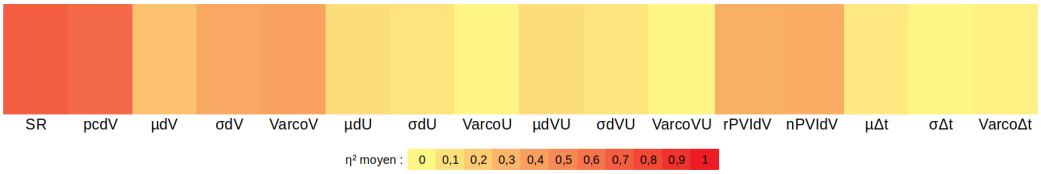


Fig. *3.4:* *Cartographie des difficultés rythmiques des apprenants japonophones du français (heatmap)*

Il serait maintenant intéressant de constituer cette cartographie pour des apprenants d'autres langues maternelles, et de comparer ces cartographies entre elles. Peut-être trouvera-t-on des similitudes entre celles des langues maternelles rythmiquement proches les unes des autres.

Il reste encore nécessaire de faire un pont entre ces difficultés et des outils pédagogiques adaptés à leur remédiation. Nous pourrions imaginer une génération de *feedbacks* pédagogiques à partir d'un score rythmique, et adaptés à la langue maternelle de l'utilisateur.

Toujours en fonction de la langue maternelle, il serait également intéressant de voir dans quelle mesure le score rythmique s'aligne sur les descripteurs de niveau du CECRL.

Une autre perspective de recherche serait d'intégrer de nouveaux paramètres au modèle, comme le pPVI de Ringeval et al. (2012), qui remplace les durées de segments par leur coefficient de variation, ou encore la mesure d'irrégularité rythmique de Scott et al. (1986). Il pourrait être intéressant d'intégrer également des mesures

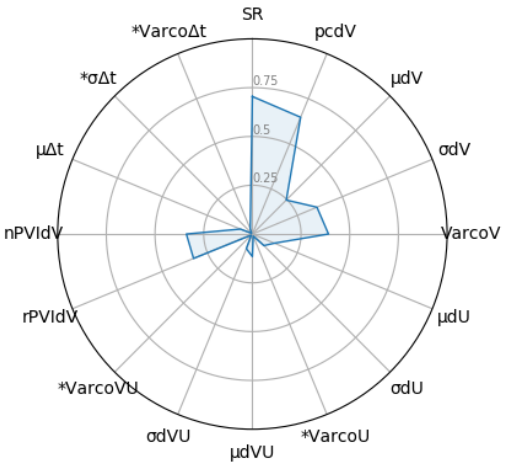


Fig. *3.5:* *Cartographie des difficultés rythmiques des apprenants japonophones du français (radar-chart)*

de F_0, et calculer le ΔF_0 entre deux noyaux syllabiques, la moyenne ou encore l'écart type de ces valeurs. Cela permettrait une cartographie plus large des difficultés prosodiques des apprenants.

Bibliographie

Abercrombie, David (fév. 1949). «Teaching Pronunciation». In : *eltj.* III.5, p. 113-122.

Ajili, Moez (2017). «Reliability of voice comparison for forensic applications». Thèse de doctorat dirigée par Bonastre, Jean-François et Rossato, Solange Informatique Avignon 2017. Thèse de doct.

Alazard, Charlotte (2013). «Rôle de la prosodie dans la fluence en lecture oralisée chez des apprenants de Français Langue Étrangère». Thèse de doctorat dirigée par Billières, Michel et Astesano, Corine Sciences du langage Toulouse 2 2013. Thèse de doct.

André-Obrecht, Régine (1986). *A new statistical approach for the automatic segmentation of continuous speech signals*. Research Report RR-0511. INRIA.

Arvaniti, Amalia (2009). «Rhythm, Timing and the Timing of Rhythm». In : *Phonetica* 66 1-2, p. 46-63.

— (2012). «Rhythm Classes and Speech Perception». In : *Understanding Prosody : The Role of Context, Function and Communication*. Sous la dir. d'O. Niebuhr. Walter de Gruyter, p. 75-92.

Arvaniti, Amalia et Tristie Ross (jan. 2010). «Rhythm classes and speech perception». In : *Speech Prosody 2010*.

Astésano, Corine (2016). «Prosodic characteristics of Reference French». In : Detey, Sylvain et al. *Varieties of Spoken French*. Oxford Scholarship.

Benzitoun, Christophe et al. (2016). «Le projet ORFÉO : un corpus d'études pour le français contemporain». In : *Corpus* 15, p. 91-114.

Bertinetto, Pier (1989). «Reflections on the dichotomy 'stress' vs. 'syllable-timing'». In : *Revue de Phonétique Appliquée* 91, p. 99-130.

Bertinetto, Pier et Chara Bertini (2008). «On modeling the rhythm of natural languages». In : *Speech Prosody 2008*, p. 427-430.

Best, Catherine T. et al. (fév. 2001). «Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system». In : *The Journal of the Acoustical Society of America* 109.2, p. 775-794.

Bhat, Suma et al. (jan. 2010). «Automatic Fluency Assessment by Signal-Level Measurement of Spontaneous Speech». In : *Second Language Studies : Acquisition, Learning, Education and Technology*.

Bloch, Bernard (jan. 1950). «Studies in Colloquial Japanese IV Phonemics». In : *Language* 26.1, p. 86-125.

Boersma, Paul et David Weenink (2019). *Praat : doing phonetics by computer [Computer program]*. Version 6.0.37, téléchargée en mars 2019 depuis http://www.praat.org/.

Bonastre, Jean-François et al. (2005). «ALIZE, a free toolkit for speaker recognition». In : *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* IEEE.

Boula de Mareüil, Philippe et Bianca Vieru-Dimulescu (déc. 2006). «The contribution of prosody to the perception of foreign accent». In : *Phonetica* 63.4, p. 247-267.

Buntine, Wray (oct. 1995). «Learning Classification Trees». In : *Statistics and Computing* 2.

Chen, Jin-Yu et Lan Wang (2010). «Automatic lexical stress detection for Chinese learners' of English». In : *2010 7th International Symposium on Chinese Spoken Language Processing*, p. 407-411.

Chen, Liang-Yu et Jyh-Shing Jang (déc. 2012). «Stress Detection of English Words for a CAPT System Using Word-Length Dependent GMM-Based Bayesian Classifiers». In : *Interdisciplinary Information Sciences* 18, p. 65-70.

Chen, Nancy et Haizhou Li (déc. 2016). «Computer-assisted pronunciation training : From pronunciation scoring towards spoken language learning». In : *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, p. 1-7.

Cohen, Jacob (1988). *Statistical Power Analysis for the Behavioral Sciences (2nd Edition)*. Routledge.

Coulange, Sylvain (juil. 2016). «Remédiation phonétique et phonologique en FLE par une approche multimodale chez les apprenants japonophones». Mémoire de master dirigé par Colletta, Jean-Marc et Rossato, Solange en master de Sciences du langage parcours de didactique du français langue étrangère. Mém. de mast., p. 160.

Cucchiarini, Catia et al. (2000). «Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology». In : *The Journal of the Acoustical Society of America* 2.107, p. 989-99.

Dauer, Richard (1987). «Phonetic and Phonological Components of Language Rhythm». In : *Proceedings of the 11th International Congress of Phonetic Sciences*. T. 5, p. 445-450.

De Meo, Anna (2012). «How credible is a non-native speaker ? Prosody and surroundings». In : *Busà M. G., Stella A. (Eds.) Methodological Perspectives on Second Language Prosody*, p. 3-9.

De Meo, Anna et al. (2012). «Comunicare in una lingua seconda. Il ruolo dell'intonazione nella percezione dell'interlingua di apprendenti cinesi di italiano». In : *La voce nelle applicazioni. Proceedings of the 7th Congress of Italian Association of Speech Sciences AISV*, p. 117-129.

Dehak, Najim et al. (avr. 2011). «Front-End Factor Analysis for Speaker Verification». In : *IEEE Transactions on Audio, Speech, and Language Processing* 19.4, p. 788-798.

Dellwo, Volker (2006). «Rhythm and Speech Rate : A Variation Coefficient for deltaC». In : *Language and language-processing*. Sous la dir. de P Karnowski et I Szigeti. Frankfurt/Main : Peter Lang, p. 231-241.

Dellwo, Volker et Adrian Fourcin (2013). «Rhythmic characteristics of voice between and within languages». In : *Revue Tranel (Travaux neuchâtelois de linguistique)* 59, p. 87-107. url : http://doc.rero.ch/record/233391.

Dellwo, Volker, Adrian Leeman et al. (2015). «Rhytmic variability between speakers : Articulatory, prosodic and linguistic factors». In : *The Journal of the Acoustical Society of America* 137.3.

Derwing, Tracey et Murray Munro (juil. 2015). *Pronunciation Fundamentals : Evidence-based Perspectives for L2 Teaching and Research*.

Deshmukh, Om et Ashish Verma (déc. 2009). «Nucleus-level clustering for word-independent syllable stress classification». In : *Speech Communication* 51, p. 1224-1233.

Detey, Sylvain (2007). «Transcription, translittération et didactique de l'oral en FLE au Japon : katakana, rômaji et orthographe française». In : *Revue japonaise de didactique du français* 2.1, p. 19-36.

— (2011). *Projet CLIJAF : corpus longitudinal interphonologique de Japonais apprenants de français*. Projets Kakenhi (B) n°23320121 et n°15H03227. Japanese Society for the Promotion of Science.

Detey, Sylvain et al. (jan. 2016). «Traitement de la prononciation en langue étrangère : approches didactiques, méthodes automatiques et enjeux pour l'apprentissage». In : *TAL Traitement Automatique des Langues* 57, p. 15-39.

Di Cristo, A et D.J Hirst (1997). «L'accentuation non emphatique en français : stratégies et paramètres». In : *Polyphonie pour Ivan Fónagy*. L'Harmattan, p. 71-101.

Eyben, Florian et al. (2013). «Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor». In : *Proceedings of the 21st ACM International Conference on Multimedia*. MM '13. Barcelona, Spain : ACM, p. 835-838.

Fagyal, Zsuzsanna et Mary-Annick Morel (1996). «Phonostylistique : étude du style dans la parole». In : *L'information grammaticale* 70.1, p. 16-20.

Ferrer, Luciana et al. (mai 2015). «Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems». In : *Speech Communication* 69.C, p. 31-45.

Flege, James (juil. 1988). «Factors affecting degree of perceived foreign accent in English sentences». In : *The Journal of the Acoustical Society of America* 84.1, p. 70-79.

— (jan. 1995). «Second language speech learning : Theory, findings and problems». In : W. Strange (Ed.), p. 229-273.

— (2003). «Assessing constraints on second-language segmental production and perception». In : *In A. Meyer & N. Schiller (Eds.), Phonetics and phonology in language comprehension and production : Differences and similarities*, p. 319-355.

— (2005). «Origins and development of the Speech Learning Model». Keynote lecture presented at the 1st ASA Workshop on L2 Speech Learning, Simon Fraser Univ., Vancouver.

Flege, James, E.M. Frieda et al. (1997). «Amount of native-language (L1) use affects the pronunciation of an L2». In : *Journal of Phonetics* 25.2, p. 169-186.

Flege, James, Murray Munro et al. (1995). «Factors affecting degree of perceived foreign accent in a second language». In : *Journal of the acoustical society of America* 97, p. 3125-3134.

Flege, James, Grace Yeni-Komshian et al. (1999). «Age Constraints on Second-Language Acquisition». In : *Journal of Memory and Language* 41.1, p. 78-104.

Fontan, Lionel et al. (sept. 2018). «Automatically Measuring L2 Speech Fluency without the Need of ASR : A Proof-of-concept Study with Japanese Learners of French». In : *Interspeech 2018*, p. 2544-2548.

Fourcin, Adrian et Volker Dellwo (juil. 2013). «Rhythmic classification of languages based on voice timing». In : *Tranel Review*, p. 87-107.

Franco, Horacio et al. (mai 1997). «Automatic pronunciation scoring for language instruction». In : *Acoustics, Speech, and Signal Processing*. T. 2, p. 1471-1474.

Furui, S. (avr. 1981). «Cepstral analysis technique for automatic speaker verification». In : *IEEE Transactions on Acoustics, Speech, and Signal Processing* 29.2, p. 254-272.

Garnerin, Mahault (avr. 2018). «Répartition hommes/femmes dans les systèmes d'IA : une étude pilote sur les grands corpus pour la transcription automatique de la parole». Mémoire de master dirigé par Rossato, Solange. Mém. de mast., p. 76.

Garofolo, John et al. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*. Philadelphia : Linguistic Data Consortium.

Gauvain, Jean-Luc et Chin-Hui Lee (avr. 1994). «Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains». In : *IEEE Transactions on Speech and Audio Processing* 2.2, p. 291-298.

Gibbon, Dafydd et Ulrike Gut (jan. 2001). «Measuring speech rhythm». In : *EUROSPEECH 2001*, p. 95-98.

Goldman, Jean-Philippe (jan. 2011). «EasyAlign : An Automatic Phonetic Alignment Tool Under Praat.» In : *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, p. 3233-3236.

Goldman, Jean-Philippe et al. (août 2007). «A Methodology for the Automatic Detection of Perceived Prominent Syllables in Spoken French». In : *A Methodologyfor the Automatic Detection of Perceived Prominent Syllables in Spoken French*.

Grabe, Esther et Ee Ling Low (jan. 2002). «Durational variability in speech and the rhythm class hypothesis». In : t. Vol. 7, p. 515-546.

Griffen, Toby D. (1991). «A Nonsegmental Approach to the Teaching of Pronunciation». In : *Teaching English Pronunciation*. Sous la dir. d'A. Brown, p. 178-190.

Guimbretière, Elisabeth (avr. 2012). *Prosodie et didactique*. accès en ligne le 21 mai 2019. URL : http://www.francparler-oif.org/images/stories/dossiers/phonetique_guimbretiere.htm.

Harrison, Alissa et al. (2009). «Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training». In : *SLaTE*, p. 45-48.

Haton, Jean-Paul et al. (2006). *Reconnaissance Automatique de la Parole Du signal à son interprétation*. UniverSciences (Paris) - ISSN 1635-625X. DUNOD, p. 392.

Huang, Xuedong et al. (avr. 1993). «The SPHINX-II speech recognition system : an overview». In : *Comput. Speech Lang.* 7.2, p. 137-148. ISSN : 0885-2308.

Jong, Nivja H. de et Ton Wempe (avr. 2009). «Praat script to detect syllable nuclei and measure speech rate automatically». In : *Behavior Research Methods* 41.2, p. 385-390.

Kahn, Juliette (2011). «Parole de locuteur : performance et confiance en identification biométrique vocale». Thèse de doctorat dirigée par Bonastre, Jean-François et Rossato, Solange Informatique Avignon 2011. Thèse de doct.

Kim, Yoon et al. (1997). «Automatic pronunciation scoring of specific phone segments for language instruction». In : *EUROSPEECH*.

Kuhl, P. K. (oct. 2000). «A new view of language acquisition». In : *Proceedings of the National Academy of Sciences* 97.22, p. 11850-11857.

Laird, Nan (1993). «14 The EM algorithm». In : *Handbook of Statistics*. Elsevier, p. 509-520.

Larcher, Anthony et al. (juin 2010). «LIA NIST-SRE'10 systems». In : *NIST-SRE'10*. Brno, Czech Republic.

Lenneberg, Eric H. (1967). «The Biological Foundations of Language». In : *Hospital Practice* 2.12, p. 59-67.

Léon, Pierre (1993). *Précis de phonostylistique : Parole et expressivité*. Nathan Université, Série « linguistique ». 335 p.

Li, Chaolei et al. (fév. 2007). «English sentence stress detection system based on HMM framework». In : *Appl. Math. Comput.* 185.2, p. 759-768. ISSN : 0096-3003.

Liss, Julie M. et al. (oct. 2009). «Quantifying Speech Rhythm Abnormalities in the Dysarthrias». In : *Journal of Speech, Language, and Hearing Research* 52.5, p. 1334-1352.

Missaglia, Federica (1999). «Contrastive prosody in SLA – An empirical study with adult Italian learners of German». In : *Proceedings of the XIV ICPhS*.

Morel, Michel et al. (2006). «Vous avez dit proéminence ?» In : AFCP, p. 183-186.

Mostow, Jack et al. (jan. 1993). «Towards a Reading Coach that Listens : Automated Detection of Oral Reading Errors». In : *Proceedings of the 11th National Conference on Artificial Intelligence*, p. 392-397.

Nazzi, Thierry et al. (1998). «Language discrimination by newborns : towards an understanding of the role of rythm». In : *Journal of Experimental Psychology : Human Perception and Performance* 3.24, p. 756-766.

Otake, Takashi et al. (avr. 1993). «Mora or Syllable ? Speech Segmentation in Japanese». In : *Journal of Memory and Language* 32.2, p. 258-278. issn : 0749-596X.

Oyama, Susan (1976). «A sensitive period for the acquisition of a nonnative phonological system». In : *Journal of Psycholinguistic Research* 5.3, p. 261-283.

Pellegrino, Elisa (2012). «The perception of foreign accented speech. Segmental and suprasegmental features affecting degree of foreign accent in Italian L2». In : *Mello H. et al. (Eds.) Proceeding of the 8 GSCP Conference*, p. 261-267.

Pettorino, Massimo, Anna De Meo et al. (2012). «Transplanting Credibility into a Foreign Voice. An Experiment on Synthesized L2 Italian». In : *Mello Heliana, Pettorino Massimo, Raso Tommaso (eds.) Speech and Corpora. Proceedings of the 7th GSCP International Conference*.

Pettorino, Massimo, Marta Maffia et al. (2013). *VtoV : a perceptual cue for rhythm identification*. University of Leuven (KU Leuven).

Piske, Thorsten et al. (2001). «Factors affecting degree of foreign accent in an L2 : a review». In : *Journal of Phonetics* 29.2, p. 191-215.

Poiré, François (2006). «La perception des proéminences et le codage prosodique». In : *Bulletin PFC* 6, p. 69-79.

Ramus, Franck et al. (déc. 1999). «Correlates of Linguistic Rhythm in the Speech Signal». In : *Cognition* 73, p. 265-292.

Ringeval, Fabien et al. (2012). «Novel Metrics of Speech Rhythm for the Assessment of Emotion». In : *INTERSPEECH*.

Rognoni, Luca et Maria Grazia Busà (jan. 2014). «Testing the effects of segmental and suprasegmental phonetic cues in foreign accent rating : An experiment using prosody transplantation». In : *Proceeding of the International Symposium on the Acquisition of Second Language Speech, Concordia Working Papers in Applied Linguistics* 5, p. 547-560.

Rossato, Solange et al. (2018). «Suivre le rythme de tes paroles». In : *Proc. XXXIIe Journées d'Études sur la Parole*, p. 37-45.

Schoonmaker-Gates, Elena (sept. 2012). «Foreign accent perception in L2 Spanish : the role of proficiency and L2 experience». In : *J. Levis & K. LeVelle (Eds.). Proceedings of the 3rd Pronunciation in Second Language Learning and Teaching Conference*, p. 84-92.

Scott, D.R. et al. (1986). «On the measurement of rhythmic irregularity : a reply to Benguerel». In : *Journal of Phonetics*, p. 327-330.

Shahin, Mostafa Ali et al. (2016). «Automatic Classification of Lexical Stress in English and Arabic Languages Using Deep Learning». In : *INTERSPEECH*.

Shobaki, Khaldoun et al. (jan. 2000). «The OGI kids' speech corpus and recognizers». In : *Sixth International Conference on Spoken Language Processing, ICSLP 2000 / INTERSPEECH 2000*, p. 258-261.

Simon, Anne C. et Anne Lacheret (2016). «Approaching variation in PFC - The prosodic level». In : Detey, Sylvain et al. *Varieties of Spoken French*. Oxford Scholarship.

Tepperman, Joseph et Shrikanth Narayanan (mar. 2005). «Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners». In : *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* T. 1, p. 937-940.

Tortel, Anne (déc. 2009). «Évaluation qualitative de la prosodie d'apprenants français : apport de paramétrisation prosodiques». Thèse de doctorat dirigée par Hirst, Daniel. Theses. Université de Provence - Aix-Marseille I.

Trubetzkoy, Nikolai S. (1939). *Principes de phonologie (Grundzüge der Phonologie)*. Klincksieck.

Truong, Khiet et al. (2004). «Automatic pronunciation error detection : an acoustic-phonetic approach». In : *Proceedings of the InSTIL/ICALL Symposium*, p. 135-138.

Variani, Ehsan et al. (mai 2014). «Deep neural networks for small footprint text-dependent speaker verification». In : *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Vitale, Marilisa et al. (jan. 2014). «An acoustic-perceptual approach to the prosody of Chinese and native speakers of Italian based on yes/no questions». In : *Proceedings of the International Conference on Speech Prosody*.

Wenk, Brian et François Wioland (1982). «Is French really syllable-timed ?» In : *Journal of Phonetics* 2.10, p. 193-216.

White, Laurence et Sven Mattys (2007a). «Calibrating rhythm : First language and second language studies». In : *J. Phonetics* 35, p. 501-522.

— (2007b). «Rhythmic typology and variation in first and second languages». In : *Segmental and prosodic issues in Romance phonology*. John Benjamins Publishing Company, p. 237-257.

Wiget, Lukas et al. (mar. 2010). «How stable are acoustic metrics of contrastive speech rhythm ?» In : *The Journal of the Acoustical Society of America* 127.3, p. 1559-1569.

Yoon, Kyuchul (jan. 2007). «Imposing native speakers' prosody on non-native speakers' utterances : The technique of cloning prosody». In : *Journal of the Modern British & American Language & Literature* 25, p. 197-215.

Zhao, Junhong et al. (2011). «Automatic Lexical Stress Detection Using Acoustic Features for Computer-Assisted Language Learning». In : *APSIPA ASC*.

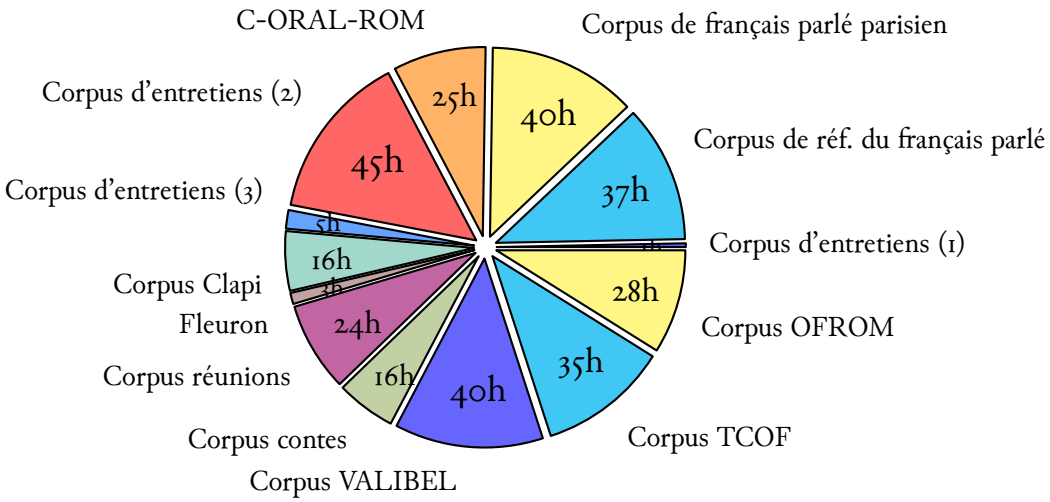
Table des figures

Liste des tableaux

Annexes

- **A. Les 13 sous-corpus du CEFC oral** - Détail des sous-corpus et de leur nombre de mots ;
- **B. Les coefficients cepstraux** - Brève présentation des MFCC et de la façon dont ils sont calculés ;
- **C. Aperçus de fichiers du CEFC** - 1. extrait d'un fichier de transcription (.orfeo), 2. extrait d'un fichier de métadonnées (.xml)

A. Les 13 sous-corpus du CEFC oral



ID Corpus	Taille (mots)	Durée
Corpus d'entretiens (1)	13 000	1h
Corpus de référence du français parlé	440 000	37h
Corpus de français parlé parisien	500 000	40h
C-ORAL-ROM	300 000	25h
Corpus d'entretiens (2, Y. Kawaguchi)	728 000	45h
Corpus d'entretiens (3)	62 000	5h
Corpus Clapi	210 000	16h
Corpus domaine académique (Fleuron)	40 000	3h
Corpus réunions	200 000	24h
Corpus contes (French Oral Narrative)	140 000	16h
Corpus VALIBEL	450 000	40h
Corpus TCOF	400 000	35h
Corpus OFROM	330 000	28h
TOTAL	**3 813 000**	**315h**

Fig. *6:* *Corpus sources du CEFC oral*

B. Les coefficients cepstraux

Les sons que nous produisons sont émis par une source (les cordes vocales) puis modelés par le conduit vocal (la langue, le conduit nasal, les dents...). Si l'on décompose le signal vocal, on retrouve des pics d'énergie sur différentes fréquences. Le premier pic, celui dont la fréquence est la plus basse, est le fondamental (F0), et représente la contribution de la source. Les autres pics sont appelés les fréquences formantiques, et rendent compte de la convolution de la source par le conduit vocal (Haton et al. 2006). Analyser la répartition de l'énergie dans l'espace fréquentiel permet de caractériser les sons du signal et la voix du locuteur.

Du moins jusqu'au début de cette dernière décennie, la plupart des systèmes de reconnaissance de la parole et d'identification du locuteur ont recourt à des méthodes spectrales fréquentielles (Kahn (2011), Haton et al. (2006)). Le spectre fréquentiel représente l'amplitude du signal (son énergie) en fonction de sa fréquence, et il est obtenu par transformation de Fourier du signal numérisé. En juxtaposant plusieurs spectres consécutifs, on obtient un spectrogramme, qui représente l'évolution du spectre dans le temps (Haton et al. 2006). En analyse de la parole il est courant de calculer des spectres de Fourier à court terme, qui correspondent au spectre calculé sur une trame de signal d'une vingtaine de millisecondes.

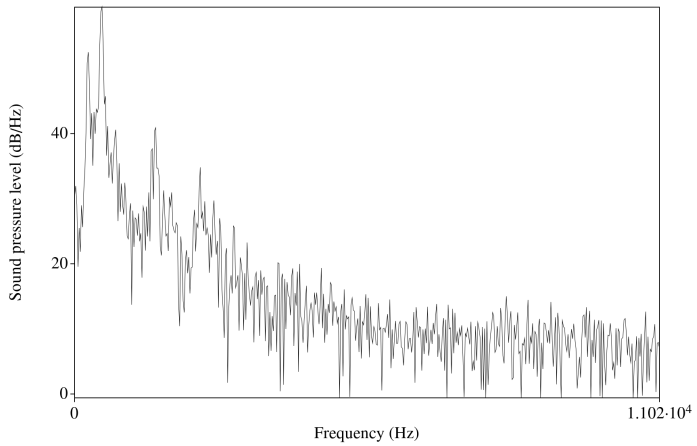


Fig. *7: Spectre d'une trame de 25ms de parole*

La figure 7 montre une transformée de Fourier d'une trame de signal de 25ms sur un [ɔ] de l'un des enregistrements du CEFC. On y distingue des piques d'intensité (en dB) en fonction de la fréquence (en Hz). La figure 8 présente la concaténation de ce spectre avec plusieurs autres avant et après, pour obtenir le spectrogramme d'un segment de 2 secondes de parole « j'aime beaucoup de choses ».

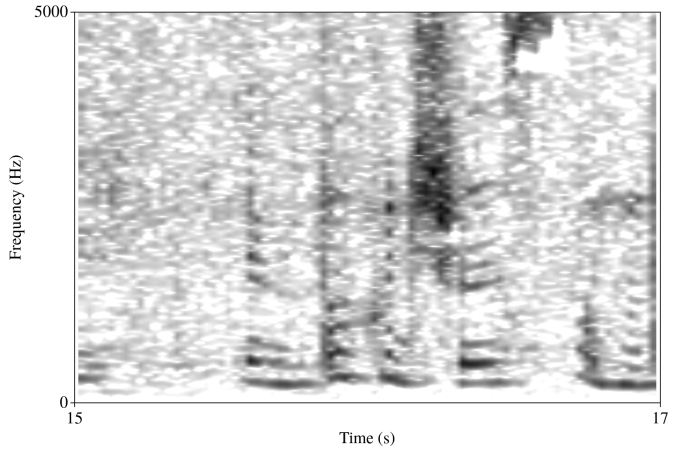


Fig. *8:* *Spectrogramme de 2 secondes de signal, trames et pas à 25ms*

Toutefois, il reste difficile de séparer la contribution de la source, et celle du conduit à partir de ce type de spectre. Pour ce faire, on a tendance à prendre le logarithme de ce spectre et y appliquer ensuite une transformation de Fourier inverse. On obtient alors un cepstre. Les MFCC sont des coefficients cepstraux pour lesquels le spectre est modifié par un banc de filtres Mels pour prendre en compte la psycho-acoustique de l'oreille humaine[1]. Pour chaque trame de signal, on calcule les coefficients comme suit :

$$c_i = \sum_{j=0}^{M} S(j) cos(i(j - \frac{1}{2})\frac{\pi}{N_f}) \quad (1)$$

avec M le nombre de coefficients (en général 10 à 15), i un itérateur sur le nombre de coefficients, $S(j)$ la transformée de Fourier de la trame et N_f le nombre de filtres (Haton et al. 2006).

À partir des coefficient, on peut calculer des dérivées temporelles. Les dérivées temporelles premières et secondes (delta et delta-delta) prennent en compte la variation immédiate des coefficients. Les premières rendent compte de la vitesse de variation, les secondes de son accélération.

1. En effet, l'oreille ne perçoit pas toutes les fréquences avec la même sensibilité, les bancs de filtres permettent alors de se focaliser sur certaines fréquences, et les échelles logarithmiques Mel ou Bark s'inspirent de l'audition humaine pour répartir les filtres selon les fréquences (Haton et al. 2006).

C. Aperçus de fichiers du CEFC

```
# sent_id = cefc-tcof-Acc_kom_07-2
# text = pour toi l'accent naturel c'est le tien
1       pour    pour    PRE     PRE     _       7       periph  _       _       2.770000        2.850000        L1
2       toi     moi     PRO     PRO     _       1       dep     _       _       2.860000        2.940000        L1
3       l'      le      DET     DET     _       4       spe     _       _       2.950000        2.970000        L1
4       accent  accent  NOM     NOM     _       7       periph  _       _       2.980000        3.150000        L1
5       naturel naturel ADJ     ADJ     _       4       dep     _       _       3.160000        3.650000        L1
6       c'      ce      CLS     CLS     _       7       subj    _       _       3.660000        3.720000        L1
7       est     être    VRB     VRB     _       0       root    _       _       3.730000        3.800000        L1
8       le tien le mien PRO     PRO     _       7       dep     _       _       3.810000        4.050000        L1
```

Fig. 3: *Aperçu d'un fichier de transcription ORFEO*

```
<langUsage>
  <language ident="fr">
   Français
  </language>
 </langUsage>
 <textClass>
  <catRef corresp="type" target="conversation"/>
  <catRef corresp="secteur" target="privé"/>
  <catRef corresp="milieu" target="amical"/>
  <catRef corresp="channel" target="face_à_face"/>
  <catRef corresp="modality" target="oral"/>
  <catRef corresp="nbLocuteurs" target="2+"/>
 </textClass>
 <settingDesc>
  <place>
   <p>
    France, Lorraine
   </p>
  </place>
 </settingDesc>
 <particDesc>
  <listPerson>
   <person xml:id="L1">
    <sex>
     M
    </sex>
    <age>
     16-20
    </age>
    <education>
     études supérieures
    </education>
    <birth>
     <placeName>
      France, Lorraine
     </placeName>
    </birth>
    <occupation>
     étudiant
    </occupation>
    <langKnowledge>
     <langKnown level="langue_maternelle">
      français
     </langKnown>
    </langKnowledge>
   </person>

```

Fig. 10: *Aperçu d'un fichier XML de métadonnées*

Résumé

Le présent mémoire de recherche s'intéresse à la place du rythme dans la perception de l'accent étranger, et à sa modélisation informatique.

Nous avons proposé de modéliser le rythme du français dans sa globalité, et dans sa variation, grâce au Corpus pour l'Étude du Français Contemporain (CEFC). Ce corpus a l'avantage de proposer une grande variété de parole, autant au niveau des situations d'énonciation que des profils de locuteur. La modélisation s'est faite à travers 16 dimensions, chacune représentant une mesure de durée de segment. Ces mesures ont été choisies de manière à ne pas nécessiter de transcription ni de système d'ASR, nous nous sommes donc basés essentiellement sur les durées de voisement et les écarts intersyllabiques, qui peuvent tous être détectés automatiquement et uniquement à partir du signal de parole. La modélisation s'est faite par l'apprentissage d'un mélange gaussien décrivant les lois de densité de probabilité de ces 16 mesures, sur les enregistrements de 1 340 locuteurs natifs, de différentes régions francophones et dans diverses situations d'énonciation.

Ces 16 paramètres rythmiques ont permis de distinguer les locuteurs natifs du corpus et les locuteurs non-natifs issus du même corpus et d'un corpus indépendant d'apprenants japonophones. Dans le cas des locuteurs non-natifs du CEFC, aux niveaux et aux langues maternelles hétérogènes, il ne s'agit que d'une tendance ($p = 0{,}067$), mais la différence est très significative avec les apprenants du corpus japonais, tous japonophones de niveau A2 ($p < 0{,}0001$).

Les tests effectués pour calculer l'efficacité des paramètres révèlent l'importance des durées intersyllabiques moyennes, de leur coefficient de variation, et du débit de parole dans le cas des locuteurs non-natifs du CEFC ; et du débit de parole et du pourcentage de voisement pour les locuteurs japonophones.