Reconnaissance automatique de sons de human beatbox

Solène Evain

Reconnaissance automatique de sons de *human beatbox*

EVAIN
Solène

Sous la direction de B. Lecouteux, D. Schwab et N. Henrich Bernardoni

Tuteur universitaire : T. Lebarbé

Laboratoires : LIG & GIPSA-lab
UFR LLASIC
Département Sciences Du Langage

Mémoire de master 2 mention Sciences du Langage - 20 crédits

Parcours : Industries de la Langue

Année universitaire 2018-2019

Reconnaissance automatique de sons de *human beatbox*

EVAIN
Solène

Sous la direction de B. Lecouteux, D. Schwab et N. Henrich Bernardoni

Tuteur universitaire : T. Lebarbé

Laboratoires : LIG & GIPSA-lab
UFR LLASIC
Département Sciences Du Langage

Mémoire de master 2 mention Sciences du Langage - 20 crédits

Parcours : Industries de la Langue – orientation professionnelle

Année universitaire 2018-2019

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 : EVAIN

PRENOM : Solène

DATE : Juin 2019

Remerciements

Je tiens tout d'abord à remercier mes trois encadrants Benjamin, Didier et Nathalie pour ce sujet hors du commun et véritablement passionnant. Je croise les doigts pour que l'aventure continue !

J'aimerais également remercier ma promotion de master toute entière ainsi que les professeurs qui nous ont accompagné pendant ces deux ans pour leur gentillesse, leur patience et la passion de leur travail qu'ils ont su nous transmettre. Parmi eux, je remercie tout spécialement Thomas Lebarbé pour avoir accepté de m'accompagner sur ce stage et Solange Rossato pour avoir accepté de faire partie de mon jury et pour le projet professionnel qui m'a fait faire mes premiers pas sur Kaldi. Je remercie également Claude Ponton et Olivier Kraif pour les projets très intéressants en programmation qui ont su me donner de bonnes bases pour être autonome dans la constitution de mes scripts.

Je remercie grandement les occupants de la 325 : W., M., E. et tout par-ticulièrement prof. M. qui m'aura permis de ne plus avoir peur des formules mathématiques et des probabilités. Merci pour votre bonne humeur et votre humour qui fait que l'on se sent tout de suite comme à la maison dans ce bureau. Je souhaite également remercier J. de ne pas m'avoir attendue le matin.

Pour finir, et même s'ils ne le verront probablement jamais, je remercie grandement mes parents pour leur soutien et leur confiance, malgré le fait qu'ils soient totalement étrangers au système universitaire.

Résumé

Le *human beatbox* est un art vocal qui consiste à utiliser les organes de la parole afin de reproduire des sons d'instruments, notamment des sons percussifs.
Nous avons travaillé lors de ce projet à l'élaboration d'un système de reconnaissance automatique de sons de *beatbox* à l'aide de la boîte à outils *Kaldi*, utilisée en reconnaissance automatique de la parole. Nous disposons de deux corpus de sons de *beatbox* contenant des répétitions de sons isolés (séparés par une pause) ainsi que des séquences rythmiques. La transcription est assurée par deux types d'écriture : un système de notation phonétique inspiré de l'alphabet de Stowell et un système d'écriture morphographique modulaire appelé *Vocal Grammatics*.
Ce travail de recherche est utilisé dans le cadre d'un projet Art Sciences proposant de faire découvrir au grand public et aux *beatboxeurs* professionnels les mécanismes mis en place pour la production de ces sons caractéristiques, mais également l'écriture *Vocal Grammatics* qui se base sur des aspects articulatoires.

Les principales contributions de ce mémoire sont les suivantes :
— état de l'art sur le *human beatbox* ;
— mise en forme d'un corpus de sons de *human beatbox* ;
— développement d'un système de reconnaissance automatique du *beatbox*.

Mots-clés : Reconnaissance vocale, *human beatbox*, Kaldi, musique, phonétique articulatoire

Abstract

Human beatboxing is a vocal art, using the speech organs to recreate instruments sounds and especially percussive sounds.
We worked on the development of an automatic recognition system for beatbox sounds with the help of the *Kaldi* toolbox used for automatic speech recognition. We have two corpora of beatbox sounds which both contain isolated sounds (seperated by a pause) and sound sequences. For transcription needs, we use two different types of writings : one inspired by Dan Stowell's alphabet based on the International Phonetic Alphabet and a modular morphographic system named *Vocal Grammatics*. This work has been conducted in the context of an "Art and Sciences" project offering an audiance of professional and inexperienced beatboxers the opportunity to learn what sort of articular mechanisms are used in the production of beatbox sounds and discover the *Vocal Grammatics* writing inspired by those mechanisms.

The mains contributions of this research project are :

— state of the art on human beatbox ;

— shapping a corpus of beatbox sounds ;

— development of an automatic recognition system for beatbox sounds.

Keywords : Speech recognition, human beatbox, Kaldi, music recognition, articulatory phonetics

Table des matières

Introduction

Ce travail de recherche s'inscrit dans le cadre d'un projet Art Sciences financé par la Structure Fédérative de Recherche Création (FED 4269) de Grenoble et par le pôle Grenoble Cognition. Il est porté par Nathalie Henrich Bernardoni, chercheure au GIPSA-lab.

L'équipe du projet *beatbox* se compose des personnes suivantes :

— Nathalie Henrich Bernardoni est directrice de recherche au CNRS et responsable du Département Parole et Cognition de GIPSA-lab. Elle apporte au projet son expertise de la parole et du chant. Elle dispose de plusieurs corpus de *human beatbox* enregistrés au cours d'études précédentes.

— Benjamin Lecouteux et Didier Schwab sont maîtres de conférences au Laboratoire d'Informatique de Grenoble (LIG). Ils apportent leur expertise sur les aspects informatique et reconnaissance de la parole.

— A. C. est designer graphique et *beatboxeur* amateur. Il est à l'origine d'un système d'écriture morphographique modulaire des sons du *human beatbox*[1] élaborée au cours de son projet de fin d'études pour l'obtention du Diplôme National Supérieur d'Expression Plastique de l'ESAD d'Amiens en 2015.

— A. P. est *beatboxeur* professionnel (nom de scène : "Andro"), ancien Champion de France en duo en 2015, participant aux championnats du monde en 2018 et classé dans le top 16 en solo en 2018. Il donne des cours de *beatbox* en milieu hospitalier pour aider à la rééducation langagière d'enfants atteints de troubles du langage mais aussi chez des particuliers qui veulent s'initier à cet art ou s'améliorer. Il travaille avec A. C. sur l'évolution de la grammaire pictographique *Vocal Grammatics*.

1. http ://www.vocalgrammatics.fr/

— AP. est doctorante au laboratoire GIPSA-lab et orthophoniste.
Elle travaille sur les aspects acoustiques, articulatoires et physiologiques de la production de sons de *human beatbox*.

— J. V. est docteur en physique et ingénieur de recherche en Arts Numériques. Il apporte au projet ses connaissances en technologies du web et s'implique sur les aspects artistiques et ludiques du projet.

— James Leonard est ingénieur de recherche au GIPSA-Lab et issu d'une formation en informatique, musique et ingénierie du son. Il travaille dans le cadre de ce projet sur la projection visuelle dynamique des sons *beatboxés* et, conjointement avec J. V. sur le développement d'une application web mobile de bio-feedback pour l'artiste.

— C. S. est ingénieur de recherche au CNRS, responsable du service plateformes du laboratoire GIPSA-lab et en charge de la plateforme expérimentale BEDEI du GIPSA-lab dédiée à l'acquisition de données multimodales en production de la parole.

— Dr. C. F. est interne en chirurgie ORL et chercheur invité au GIPSA-lab. Dr. I. A. est chirurgien ORL et phoniatre du CHU Grenoble Alpes. Leur contribution porte sur des mesures endoscopiques laryngées lors de la production de *human beatbox*.

Ce projet est un projet artistique et scientifique qui demande une installation, un public et une performance. Il est voué à être présenté dans divers festival et compétitions de *beatbox*. Le public, pouvant tout aussi bien être composé de néophytes que de *beatboxeurs* professionnels, sera invité à produire des sons de *beatbox* isolés, c'est-à-dire séparés par une pause, ou une rythmique. Un système de reconnaissance automatique de sons de *human beatbox* interviendra, permettant de transcrire les sons en écriture *Vocal Grammatics*. La transcription en signes sera présentée sous une forme dynamique et artistique. Une tablette viendra compléter cette écriture et servira de bio-feedback au public afin de les informer sur les mécanismes articulatoires mis en jeu lors de la production de différents sons. Certains réglages, comme le choix du microphone ou encore la distance entre le performeur et le microphone, restent encore à déterminer et dépendront notamment des résultats obtenus lors de ce travail de recherche.

Le développement de notre système est innovant étant donné qu'à ce jour aucun système de reconnaissance automatique du *beatbox* n'existe, bien que quelques recherches aient été faites sur le sujet [Picart et al., 2015], [Sinyor et al., 2005]. Nous présenterons dans un premier temps le travail d'annotation de corpus de sons de *beatbox* grâce à l'écriture *Vocal Grammatics*. Dans un deuxième temps, nous exposerons le développement d'un système de reconnaissance de sons de *beatbox* s'appuyant sur la boîte à outils état de l'art *Kaldi* [Povey et al., 2011] utilisé en reconnaissance de la parole. Enfin, nous estimerons la faisabilité d'un tel système en abordant les réussites et les perspectives de la reconnaissance automatique de sons de *beatbox*.

Chapitre 1

État de l'art

1.1 Le *human beatbox*

1.1.1 Qu'est-ce que le *human beatbox* ?

Le "*human beatbox*" souvent raccourci en "*beatbox*" est un art vocal né dans les années 1980 à New-York. Il consiste à utiliser les organes de la parole afin de produire des sons d'instruments, notamment des sons percussifs. Le *beatbox* (littéralement "boîte à rythme") est né du monde du hip-hop et était utilisé initialement comme une aide rythmique pour les rappeurs de rue, comme l'expliquent A. C. et A. P. dans leur TedX *L'écriture du beatbox* réalisé à Reims en 2017 [1]. En effet, le prix du matériel professionnel ne permettant pas à tous les rappeurs de s'équiper, la solution a été trouvée en reproduisant les sons des boîtes à rythme disponibles sur le marché comme le TR-808 Rythm Composer (Figure 1.1)[Lederer, 2005].

Initialement simple boîte à rythme humaine, les techniques de production de sons de *beatbox* vont évoluer au fur et à mesure des années pour devenir une discipline à part entière, accompagnée de championnats aux niveaux national et mondial. Doug E. Fresh et Darren Robinson sont considérés comme pionniers de la discipline [Hess, 2007]. Rahzel, artiste des années 90 connu sous le nom de *Godfather of Noyze*, s'est, lui, fait connaître grâce à son titre *If your mother only knew* et l'impression de polyphonie qui en émerge. Cette technique consiste à combiner une ligne de basse

1. www.youtube.com/watchvN̄SDnE2iGe8g&t4̄1s

Figure 1.1 – TR-808 Rythm Composer de chez Roland Corporation, 1980

Source: www.flickr.com/photos/polytropia/6944667788

avec un chant [Stowell and Plumbley, 2008] [Proctor et al., 2013]. La complexité des sons augmente année après année avec l'apparition des nouvelles générations de *beatboxeurs* sur scène. Chaque artiste essaie de se démarquer des autres en produisant des sons compliqués voire impossibles à reproduire pour d'autres. Pour cela, il n'est pas rare de les voir s'entraîner plusieurs mois à la réalisation d'un son afin de le maîtriser parfaitement. Les heures de travail sont importantes car les *beatboxeurs* cherchent à gommer au maximum le fait que les sons sont produits par l'humain, et essaient de se rapprocher au maximum du son d'un instrument.

Les compétitions

Les performances peuvent être *a capella* ou amplifiées. Les *beatboxeurs* tiennent leur microphone à un ou deux centimètres de leur bouche, ce qui accentue les basses. L'encapsulement est une méthode très régulièrement utilisée qui consiste à recouvrir le microphone d'une ou deux mains afin de créer une caisse de résonance qui va permettre de moduler la réponse acoustique [Stowell and Plumbley, 2008]. L'utilisation du microphone au niveau du nez ou de la gorge est possible et vient créer un effet de filtre passe-bas laissant passer les basses fréquences et atténuant les hautes fréquences.

Les compétitions de *beatbox* ne sont pas sans rappeler les battles de rap, avec

un système de "question-réponse". Le principe est simple : assurer sa supériorité sur son concurrent. La capacité à couper la parole en commençant sa séquence avec un son très fort en basses fréquences, la rapidité d'exécution de son tour de parole avec un maximum de sons produits en un temps donné et enfin la complexité des sons sont autant de paramètres d'évaluation utilisés par un jury pour déterminer le meilleur *beatboxeur*. Ces paramètres d'évaluation peuvent être indépendants les uns des autres étant donné que l'enchaînement de sons très complexes peut parfois être difficile. Là encore, un réel entraînement est nécessaire afin de trouver quels sons enchaîner pour être le plus rapide ou encore quelles stratégies d'articulation utiliser pour faciliter le glissement d'un son à un autre (co-articulation) et éviter de se fatiguer trop vite en produisant des sons trop complexes qui peuvent essouffler.

Une réelle maîtrise des organes de la parole

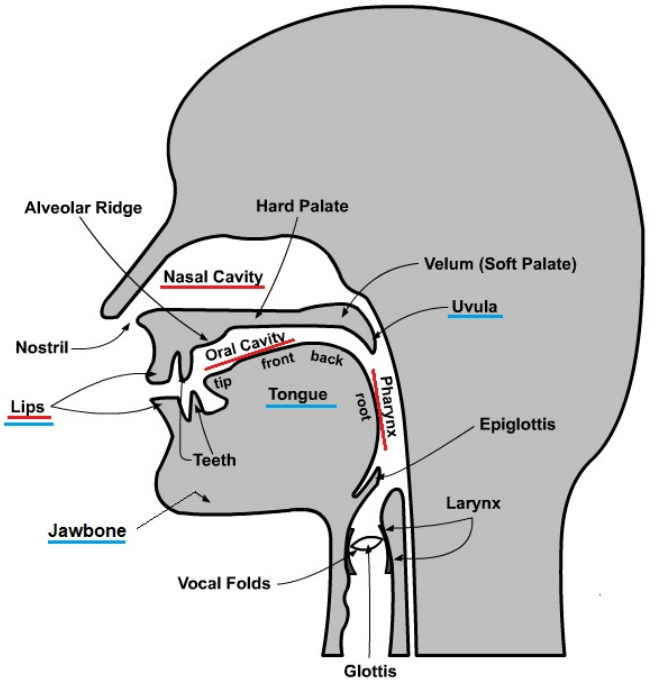


FIGURE 1.2 – Articulateurs et résonateurs de la parole.

Source: Inspiré de www.madbeppo.com/french-language/the-organs-of-speech-from-the-neck-up/

Comme nous avons pu le voir, un bon niveau en *beatbox* va de paire avec de nombreuses heures d'entraînement. Les *beatboxeurs* développent au cours de leur

apprentissage une maîtrise fine des articulateurs utilisés pour la parole (Uvule, larynx, machoire, lèvres, langue. Voir Figure 1.2) mais aussi de leur utilisation de l'air et des résonateurs (Larynx, cavités nasales et buccale, pharynx. Voir Figure 1.2). Les poumons, les plis vocaux, les articulateurs et résonateurs jouent tous un rôle dans la production des différents sons, auxquels s'ajoute un très bon sens du rythme. Cette maîtrise est attestée dans plusieurs travaux dont [Proctor et al., 2013] :

> *"The subject appears to be highly sensitive to ways in which fine differences in articulation and duration can be exploited for musical effect. Although broad classes of sound effects were all produced with the same basic articulatory mechanisms, subtle differences in production were observed between tokens, consistent with the artist's description of these as variant forms."*
>
> "Le sujet présente une grande sensibilité aux utilisations qui peuvent être faites de petites différences dans l'articulation et la durée pour la création d'un effet musical. Bien que les grandes classes de sons aient été produites sur une base de mécanismes articulatoires similaires, des différences subtiles dans la production ont été observées, en adéquation avec la description de ces sons comme des variantes par l'artiste."

et Annalisa Paroni dans son mémoire d'orthophonie [Paroni, 2014] :

> *"Our subject demonstrated to have extra fine control over his articulatory structures, being able to perform tiny adjustments in order to obtain very peculiar acoustic effects ; to reproduce very accurately his beatbox gestures ; to perform extremely fast and precise movements ; to master articulatory gestures not belonging to his native language and also not attested in any human language so far."*
>
> "Notre sujet a manifesté un contrôle très fin sur ses articulateurs en étant capable de réaliser des ajustements ciblés dans l'optique d'obtenir des effets acoustiques particuliers ; en étant capable de reproduire avec précision les mouvements réalisés [pour la production d'un son] ; en pouvant réaliser des mouvements rapides et précis ; en produisant efficacement des gestes articulatoires n'appartenant pas à sa langue d'origine ou à aucune langue connue."

De ces travaux, nous retenons que les *beatboxeurs* parviennent à apporter des ajustements précis à une articulation de base, ce qui aura pour effet d'en créer des variantes acoustiques profitables puisqu'elles agrandissent leur vocabulaire. Ces variations peuvent les aider à trouver le meilleur enchaînement possible de deux sons.

Se détacher de la parole

Une des difficultés du *beatbox* réside dans le fait que certains sons se rapprochent de la langue parlée par les *beatboxeurs* (les plosives /p/, /t/, /k/ pour le français par exemple). Il leur faut apprendre à passer outre ce qu'ils ont appris afin de prendre une distance avec la langue et de renforcer l'illusion que les sons proviennent d'un instrument. Pour ce faire, ils utilisent diverses techniques vocales qui leur permettent de reproduire différents timbres, le timbre étant ce qui permet de différencier les instruments. Ils essaient également de supprimer les indices linguistiques qui permettent de reconnaître la source des sons, en atténuant certaines fréquences présentes dans la parole, en supprimant les pauses présentes dans la parole pour respirer [Stowell and Plumbley, 2008] ou encore en augmentant le rythme articulatoire. Afin de pouvoir reprendre leur souffle, certains sons peuvent être réalisés de deux façons : en inspirant et en expirant. Cela leur permet de faire tourner une boucle de sons aussi longtemps qu'ils le veulent sans s'essouffler [Stowell and Plumbley, 2008] [Lederer, 2005].

1.1.2 Vers une écriture du *beatbox* inspirée de la phonétique articulatoire

L'utilité d'une écriture du *beatbox* est de permettre le partage et l'apprentissage de cet art ainsi que sa pérénnité. Dans les années 80-90, l'apprentissage se faisait en observant les autres *beatboxeurs* ou en inventant ses propres sons. L'apparition d'Internet et de plateformes de visionnage de vidéos telle que Youtube ont permis d'avoir un accès plus facile à des vidéos de compétitions de *beatbox*. De nombreux jeunes se sont donc formés en essayant de reproduire à l'oreille, avec plus ou moins une aide visuelle, ce qu'ils entendaient. Ils étaient néanmoins limités par le fait que la production de sons de *beatbox* n'est qu'à moitié observable et que l'utilisation qui est faite du velum ou de la langue par exemple est non visible.

Aujourd'hui, de nombreux tutoriels sont disponibles sur Internet mais les youtubeurs sont confrontés à la difficulté de ne pouvoir décrire avec précision ce qu'ils font. Une écriture ou, dans un premier temps, une façon de décrire tous ces méchanismes complexes et précis s'est donc avérée nécéssaire. L'apport subsidiaire d'une écriture étant qu'elle est à l'épreuve du temps et permet la conservation d'une langue et dans notre cas, du *beatbox*.

La Notation Standard du *beatbox* (SBN)

Beatbox Sounds With Standard Beatbox Notation

Sound	Standard Beatbox Notation
Kick drum	[b]
Nasal growl bass	[bgn]
Closed high hat	[t]
Open high hat	[ts]
Crash	[tsh]
Classic snare	[p]
Inward K snares	[k]

Figure 1.3 – Exemple de la SBN

Source: [Sapthavee et al., 2014]

Une première écriture du *beatbox* a été proposée par Mark Splinter et Revd Gavin Tyte entre 2006 et 2012 [Tyte and Splinter, 2014]. Cette écriture est appelée SBN pour Standard *beatbox* Notation. Elle utilise les lettres de l'anglais pour décrire l'acoustique des sons (voir Figure 1.3). Une première limite de cette écriture est venue du fait que les sons produits par les *beatboxeurs* sont parfois complexes et qu'un même son peut être produit de différentes manières sans que cela puisse être précisé dans cette écriture. [Tyte, 2019].

L'alphabet pour le *beatbox* de Dan Stowell

Une deuxième écriture a été développée par Dan Stowell. Elle vient compléter la SBN en introduisant des notions de phonétique grâce à l'Alphabet Phonétique International (API). L'API et ses diacritiques ont apporté des symboles supplémentaires permettant de préciser par exemple si un son est nasalisé, lateralisé, long...

Description	Phonetic
Simple kickdrum	b bo bm bʍ
Classic snare	bʃ pʃ
Simple closed hihats	t͡t t͡s ^t͡t ^t͡s
Simple open hihat	t͡s: ^t͡s:
606 snare	tˡ kˡ
808 snare	t͡ʃ ˇ
909 snare	ʘyʃ
Reverse snare	^ʃp̃
Reverse hat	^ʃt̃
808 rimshot	k ka
Inhaling handclap	^ʇˡ
Daysoftheweek kickdrum	ðw
Reverse kick	fd͡ vd͡
808 kick / techno swallow	ʘ
Inhaling lip-kick	^ɓ ^ʃɓ

Description	Phonetic
Maracas or other "palatal trills"	«
Clicks	ǃ ǁ ǂ (For the differences between these see the IPA)
Click roll	ǁǁ ǂǁǁ
Tongue pop	§
Kissy-kissy hihat	※
Cough snare	ʘa̰
Closed-hihat type "tutting"	t’
Fast hats	tkt
808 snare roll	r̰
"Fading-in" snare roll	r̰r̰
Quick unvocalised “tschowi” scratch	tʃʍʂ
Vocal tap	ɾ́
Reverse reverb	^ʂ ^ɕ

FIGURE 1.4 – Alphabet pour le *beatbox* inventé par Stowell

Source: http ://mcld.co.uk/beatboxalphabet/

Mais cette écriture comporte encore quelques limites comme le fait qu'un *simple closed hi-hat* noté [ts] n'est pas différencié de l'écriture phonétique de la fin d'un mot comme cats [k a t s] alors que le son est différent, ou le fait que certains sons sont tout simplement beaucoup trop éloignés des langues existantes pour être codés par l'API [Lederer, 2005] ,[Proctor et al., 2013], [Paroni, 2016]. De plus, la notion de rythme n'est pas transcrite. Cette proposition inclut donc plus de possibilités d'écritures que la précédente mais s'écarte du principe du *beatbox* de s'éloigner de la parole. Encore une fois, la multiplicité des sons et de leurs réalisations fait qu'ils ne sont pas toujours transcriptibles comme le concluent [Stowell and Plumbley, 2008] et [Paroni, 2014] dans leurs travaux. Enfin, cette transcription peut très vite devenir complexe autant à écrire qu'à lire et n'est donc pas optimale.

Vocal Grammatics

Utiliser l'IPA revient à se concentrer sur l'acoustique du son et non pas sur l'articulation. Une tentative d'écriture du *beatbox* par la graphiste Léa Chapon est développée en 2006. Son travail s'appuie également sur l'API qu'elle retravaille afin de créer une typographie propre au *beatbox*. L'écriture est rendue plus lisible mais le même problème ressort : la transcription des sons très éloignés des sons de l'API

est impossible. Le besoin se fait sentir de pouvoir représenter le *beatbox* en se détachant des sons produits dans les langues et de pouvoir annoter toutes les nouvelles techniques qui voient le jour avec les nouvelles générations.

Une nouvelle écriture a été inventée par A. C. graphiste, lors de son travail de fin d'études à l'ENSAD d'Amiens. Son idée est de prendre le problème à l'envers et de ne plus se concentrer sur l'acoustique mais sur la réalisation des sons. Il souhaitait pouvoir offrir aux utilisateurs la possibilité de décrire les techniques d'articulation utilisées en production de beatbox, avec un niveau de précision assez haut pour pouvoir noter les fines différences de réalisation à l'origine des variantes dont nous avons parlé en partie 1.1.1. Cette écriture s'appuie sur la représentation mentale du *beatboxeur* de l'utilisation qu'il fait de ses articulateurs et des zones de résonance lors de la production d'un son. Les caractères créés prennent la forme de glyphes construits avec différents signes, un peu à la manière des hiéroglypes égyptiens. Un premier signe indique quelles parties du conduit vocal sont utilisées (lèvres, dents, larynx...)(figure 1.5). Pour les décrire, les lieux d'articulation issus de la phonétique articulatoire sont utilisés.

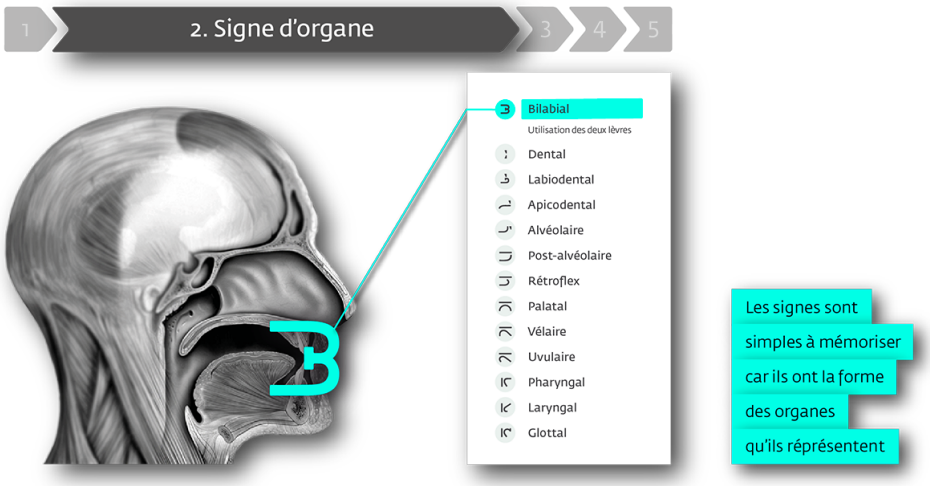


Figure 1.5 – Parties du conduit vocal utilisées

Source: www.vocalgrammatics.fr

Ensuite, un deuxième signe apporte une information sur le mode de production des sons (figure 1.6) est ajoutée : plosif, fricatif... à nouveau ces termes sont tirés de la phonétique.

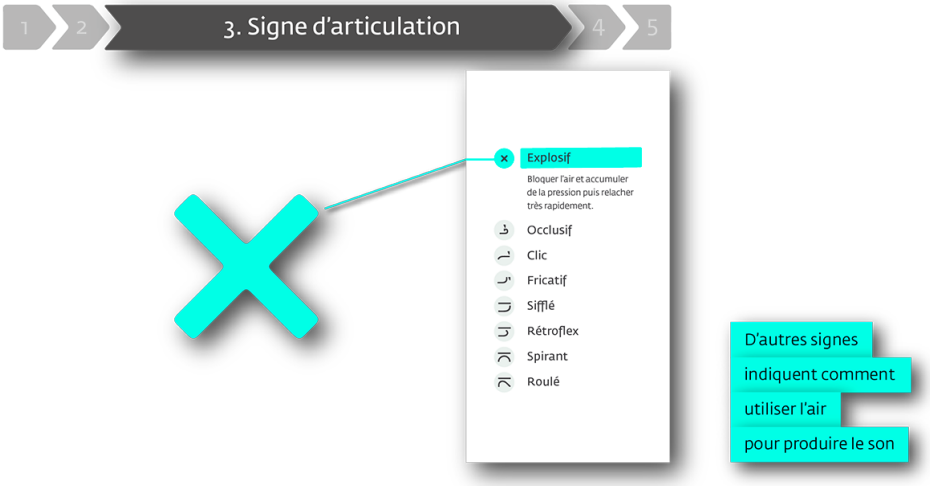


FIGURE 1.6 – Mode de production du son

Source: www.vocalgrammatics.fr

Les signes utilisés sont simples à lire puisqu'ils représentent la forme de "l'organe" utilisé. Une fois les signes combinés en glyphes, il est possible de les positionner sur une tablature, (figure 1.7) permettant alors d'avoir des informations sur la rythmique.

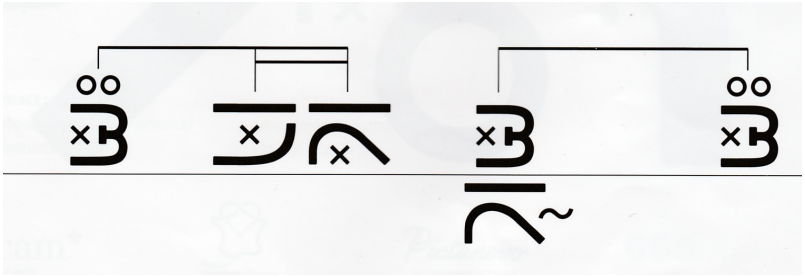


FIGURE 1.7 – Utilisation de l'écriture *Vocal Grammatics* en tablature

Source: www.vocalgrammatics.fr

Cette écriture déjà bien élaborée est encore en cours d'évolution. Comme nous l'avons vu, son principe est d'associer des informations sur le point d'articulation et le mode d'articulation. Le point d'articulation est le "point de rapprochement ou de contact entre la partie mobile (langue ou lèvre inférieure) et la partie fixe (lèvre supérieur, dent, palais...) de l'instrument vocal humain au cours de l'émission sonore" [2]. Il peut être bilabial, alvéolaire ... Le mode d'articulation est "la manière selon laquelle le courant d'air qui vient des poumons se dirige vers l'extérieur" [3] c'est-à-dire explosif, fricatif etc. L'écriture *Vocal Grammatics* est une représentation visuelle des lieux et modes d'articulation, inspirée de la phonétique articulatoire.

Nous pouvons observer plusieurs avantages à l'utilisation de cette écriture. Premièrement, elle se comprend très vite. Elle a été testée auprès d'enfants pour leur apprendre le *beatbox* comme nous pouvons le voir dans une vidéo [4] publiée sur Youtube via la chaîne *Vocal Grammatics*. Deuxièmement, elle est facile à retenir. Troisièmement, en comparaison aux écritures citées précédemment, elle permet de décrire des sons complexes et constitue une écriture à part entière, inventée spécialement pour le *beatbox*. Enfin, le rythme et les variations assez fines d'un son peuvent être transcrites. Néanmoins, une limite réside dans le fait qu'elle se base sur une estimation proprioceptive de la production d'un son qui peut être biaisée, d'autant plus que les sons sont produits assez rapidement. Il peut être difficile pour les *beatboxeurs* de faire le chemin inverse, et d'essayer de retrouver une articulation à partir de l'enregistrement d'un autre *beatboxeur*.

1.2 La reconnaissance vocale

1.2.1 Un rapide historique

La reconnaissance automatique de la parole est née dans les laboratoires Bell dans les années 1950. Le système était alors capable de reconnaître 9 chiffres prononcés par une personne.

2. https ://master-fldl.parisnanterre.fr/medias/fichier/phone-utique-1_1532524252625-pdf
3. https ://master-fldl.parisnanterre.fr/medias/fichier/phone-utique-1_1532524252625-pdf
4. https ://www.youtube.com/watch ?v=SlVae0sNGBU

Dans les années 1970, le département de la défense américain avec la *Defense Advanced Research Projects Agency* (DARPA), chargée de la recherche et développement des nouvelles technologies destinées à un usage militaire a énormément investit dans la reconnaissance vocale. Un système développé et nommé *Harpy* est capable de comprendre un vocabulaire d'environ 1000 mots. Au même moment, les laboratoires Bell améliorent leur système qui est alors capable de reconnaître la voix de plusieurs locuteurs.

Dans les années 1980, les modèles de Markov cachés (HMM - Hidden Markov Model) permettent à la reconnaissance vocale de faire un nouveau bond en avant et de pouvoir reconnaître plusieurs milliers de mots. Ces années marquent également l'apparition des réseaux de neurones.

Dans les années 1990, l'augmentation du nombre de foyers et d'entreprises à faire l'acquisition d'un ordinateur va donner un élan aux logiciels de reconnaissance vocale comme le logiciel Dragon pour la dictée.

Les années 2000 marquent de grandes avancées notamment avec l'apparition de la commande de recherche vocale développée par Google et leurs corpus d'apprentissage qui augmentent au fur et à mesure que les requêtes des utilisateurs sont collectées.

Aujourd'hui, les systèmes de reconnaissance vocale ont une place toujours plus importante, que ce soit dans la domotique [Le Grand, 2012], dans le monde professionnel ou encore dans les maisons avec Siri, Google Home ou encore Alexa (Amazon). La recherche de systèmes toujours plus efficaces amène les chercheurs à augmenter la taille des corpus d'apprentissage mais également à utiliser différentes techniques : les *Subspace Gaussian Mixture Models* (SGMM) [Povey et al., 2011], les réseaux de neurones [Liu, 2015], les systèmes hybrides [Yu and Deng, 2015]... En parallèle, la recherche continue pour les langues que l'on appelle "peu dotées", c'est-à-dire pour lesquelles les ressources sont largement inférieures à celles des langues parlées partout dans le monde. La Figure 1.8 de [Juang and Rabiner, 2004], article dont s'inspire cet historique, présente un résumé des évolutions connues par les systèmes de reconnaissance automatique de la parole jusqu'au début des années 2000.

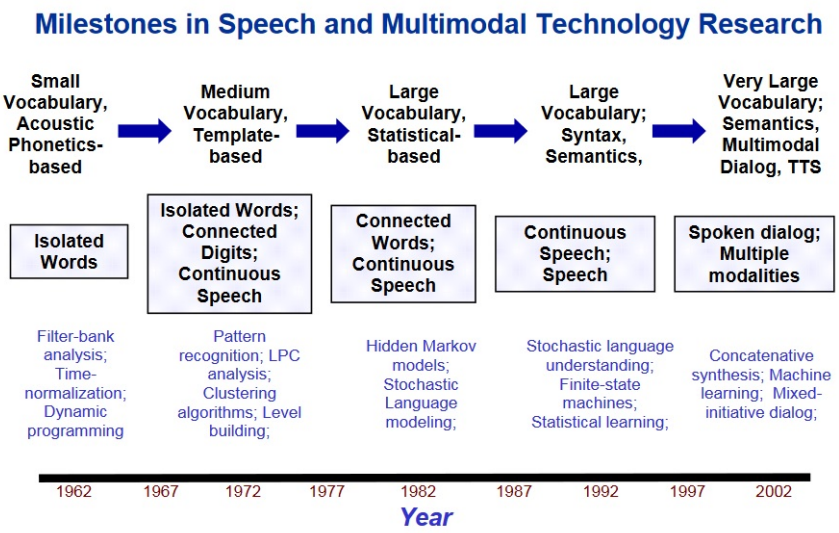


FIGURE 1.8 – Évolutions de la recherche en reconnaissance automatique de la parole

Source: [Juang and Rabiner, 2004]

1.2.2 Fonctionnement général d'un système de reconnaissance vocale

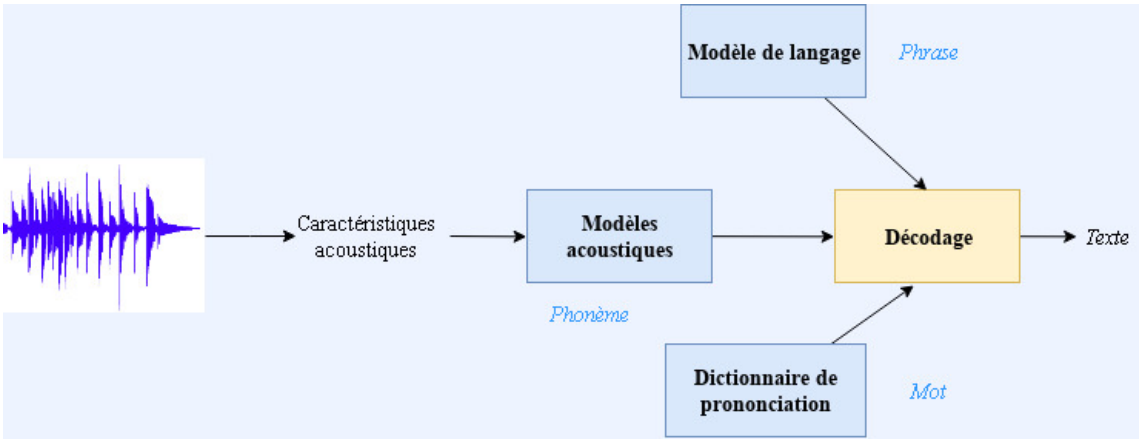


FIGURE 1.9 – Fonctionnement d'un système de reconnaissance vocale

Un système de reconnaissance vocale suit les étapes suivantes (Figure 1.9) : On part d'un signal, sur lequel sont calculés des paramètres acoustiques. Ces paramètres servent à extraire l'information nécessaire à la reconnaissance de la parole. Le fait de ne pas prendre le signal acoustique en entier permet d'alléger le traitement et d'enlever des sons qui ne sont pas nécessaires pour la reconnaissance. Ils sont calculés sur des fenêtres glissantes (généralement aux alentours de 20ms) et stockés dans des

vecteurs.

Ensuite, des modèles acoustiques estimés sur des paramètres, afin de modéliser une corrélation entre leur distribution et une unité acoustique (phonème, syllabe etc. selon le système), vont émettre l'hypothèse des probabilités d'émission pour chaque unité acoustique (ces modèles ont préalablement été estimés pour chaque unité acoustique sur des corpus de taille conséquente, afin d'observer le maximum de variabilité). On obtient ainsi une probabilité d'émission d'unités acoustiques correspondant au signal d'entrée.

Un dictionnaire de prononciation entre alors en jeu. Ce dictionnaire, fait correspondre à chaque mot du lexique sa (ou ses) transcription(s) en unités acoustiques. Il va permettre au système de poser des frontières et de construire des mots en concaténant les unités.

Enfin, le modèle de langue, permet de définir la probabilité d'émettre un mot en fonction des précédents afin de donner une cohérence linguistique au texte final. Toutes ces étapes constituent le décodage et donnent la transcription du signal en sortie.

D'un point de vue plus formel, l'équation d'un système de reconnaissance automatique de la parole se définit ainsi :

$$W^* = argmax_W P(W|A) = argmax_W \frac{P(W).P(A|W)}{P(A)}$$

Où W^* est la séquence optimale à trouver en fonction du modèle de langage W et du signal acoustique A. La probabilité du signal acoustique étant inutile dans le calcul d'argmax, la solution peut être simplifiée ainsi :

$$W^* = argmax_W P(W).P(A|W)$$

Le système aura donc à trouver la solution maximisant la probabilité linguistique avec la probabilité acoustique sachant une séquence donnée.

Nous pouvons donc voir que les modèles acoustiques et linguistiques doivent être de bonne qualité pour reconnaître correctement un signal. Généralement, plus la

quantité de données d'apprentissage pour ces deux modèles est importante, meilleure sera la qualité du système. Par ailleurs, il est nécessaire que les données d'entraînement soient similaires à celles qui seront rencontrées dans la "réalité". Par exemple, au niveau des modèles acoustiques, des modèles entraînés sur des voix d'adultes ne pourront pas servir à décoder de la parole d'enfants [Russell and D'Arcy, 2007]. En effet, les caractéristiques acoustiques des deux tranches d'âges sont différentes, à cause des différences physiologiques (voir partie 1.2.3). Il y a également une différence au niveau de la maîtrise de la langue et des sujets abordés qui entrent en jeu pour le modèle de langue.

1.2.3 Détail des grandes étapes d'un système de reconnaissance vocale

Nous détaillerons ici plus en détails les différentes étapes d'un système de reconnaissance vocale, à savoir l'extraction de paramètres acoustiques, le modèle de langue, le dictionnaire de prononciation et les modèles acoustiques.

Extraction de paramètres acoustiques

L'extraction de paramètres acoustiques va permettre d'aller chercher dans le son ce qui peut aider à l'identification des différents phonèmes et mettre de côté le reste (le bruit par exemple, les fréquences non utilisées pour la parole etc.). Plusieurs types de paramètres acoustiques peuvent être extraits : les LPC (*Linear Predictive Coding*), qui servent à estimer les formants, la méthode de filtrage RASTA (*RelAtive SpecTrAl*) qui sert à décoder des sons bruités en augmentant le signal de parole ou encore les MFCC (*Mel Frequency Cepstral Coefficients*) [Narang and Gupta, 2015] [Dave, 2013] dont nous allons parler ici. Les MFCC sont les paramètres les plus utilisés en reconnaissance de la parole. Ils servent à "imiter le comportement observé par le système auditif de l'Homme" et "maximiser la performance de la reconnaissance" [Picone, 1993]. Leur extraction permet de retrouver la forme qu'a pris le tractus vocal pour prononcer la phrase en cours d'analyse et permet donc de pouvoir déterminer quel phonème a été prononcé. En effet, chaque forme que prend le tractus vocal, à laquelle on ajoute le placement des articulateurs, donne un son. Pour chaque son, différentes fréquences sont amplifiées. C'est ainsi qu'en partant d'un son et en

analysant les fréquences qui le composent, on peut retrouver la forme du tractus vocal à l'émission de ce son ([Lyons, 2019].

Voici les étapes de calcul des coefficients MFCC [Dao et al., 2016]

— 1. Pré-emphase

C'est une étape où l'on augmente l'énergie en hautes fréquences.

— 2. Découpage en fenêtres de temps

Il consiste en un découpage en fenêtres de 20 à 40 ms car bien qu'un signal ne soit pas stable, il est considéré comme tel sur une courte période. Le signal est divisé en fenêtres de 256 échantillons. Deux fenêtres consécutives se recouvrent avec un décalage de 10 ms.

— 3. Fenêtre de Hamming

Chaque fenêtre est multipliée par une "fenêtre de hamming », afin de garder la continuité du signal et éviter la distorsion spectrale (effets de bord). [Pellegrini and Duée, 2003]

— 4. Transformée rapide de Fourier

La transformée rapide de Fourier va permettre de calculer le spectre de puissance de chacune des fenêtres ce qui permet d'avoir plus d'informations sur les fréquences présentes dans la fenêtre.

— 5. Analyse par banc de filtres

Ce qui est obtenu grâce à des FFT (Fast Fourier Transform) n'est pas exactement représentatif de ce que l'Homme peut percevoir. En effet, plus on monte en fréquences, moins on est capable de discerner deux sons proches. L'idée va donc être de prendre des "paquets" et de synthétiser la quantité d'énergie qui se trouve à différentes fréquences. C'est ce qu'on appelle une analyse par banc de filtres. Les filtres triangulaires se trouvent sur l'échelle Mel. Les bornes du banc de filtre peuvent être changées pour cibler une bande de fréquence particulière, c'est-à-dire enlever ou rajouter des hautes ou basses fréquences [Pellegrini and Duée, 2003]. Couramment, ce sont 24 filtres qui sont calculés, correspondant aux 24 bandes critiques de l'audition humaine sur l'échelle de Bark d'Harvey Fletcher.

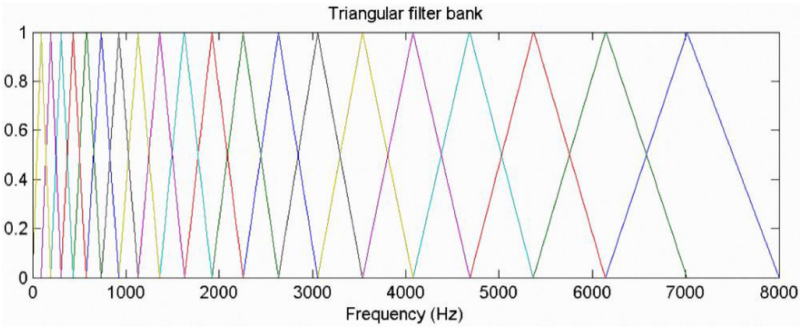


FIGURE 1.10 – Filtres triangulaires sur l'échelle Mel

Source: https ://dsp.stackexchange.com/questions/28898/mfcc-significance-of-number-of-features

Comme le montre la figure 2.8, plus on monte en fréquences, plus les filtres sont larges, ce qui est dû au fait que l'Homme distingue de moins en moins de fréquences proches au fur et à mesure que la fréquence augmente, phénomène que les coefficients MFCC permettent de représenter.

— 6. Transformation en cosinus discrète (*discrete cosine transform* ou DTC)

L'échelle de notre audition n'est pas linéaire mais logarithmique. On prend donc le logarithme des énergies des filtres triangulaires. Ensuite, une conversion du spectre logarithmique de Mel en temps au moyen de la DTC est réalisée. Ceci permet de décorréler les énergies des filtres. La transformée de Fourier inverse donne alors les Frequency Cepstrum Coefficients (MFCC). L'ensemble des coefficients est un vecteur acoustique qui représente toutes les informations phonétiques importantes pour la reconnaissance vocale. Chaque mot de départ est transformé en une séquence de vecteurs acoustiques. Ces vecteurs peuvent être utilisés pour représenter et reconnaître les caractéristiques de la voix du locuteur. On garde les 13 premiers coefficients qui représentent l'enveloppe du spectre (les coefficients suivants donnent des détails sur le spectre), le premier étant l'énergie.

— 7. Calcul des delta et delta-delta

Le calcul de la dérivée temporelle première augmente le nombre de paramètres et informe sur l'évolution de l'énergie entre deux vecteurs MFCC : stationnaire, en augmentation ou en diminution. La dérivée temporelle seconde apporte quant à elle des informations sur l'accélération, c'est-à-dire si

l'augmentation ou la diminution observée avec les delta est forte ou non.

Modèle de langue

Le modèle de langue représente la connaissance de la langue qu'utilise un individu lorsqu'il parle. Il faut donc, pour l'entraîner, un grand corpus textuel, qui ne soit pas trop éloigné de la parole qui doit être décodée. Par exemple, un modèle de langue appris sur des journaux télévisés, parlant de guerre, de politique ou encore de finance, ne conviendra pas pour de la reconnaissance de voix d'enfants surtout si ceux-ci parlent de contes. Il existe plusieurs types de modèles de langue : les modèles à base de grammaires formelles qui se basent sur des représentations acoustiques de phrases stockées dans le système et que le signal inconnu doit venir matcher, les modèles de langue à base de réseaux de neurones et les modèles stochastiques, basés sur les probabilités. Nous détaillerons pas les deux premiers car il n'y avait pas nécessité de les utiliser dans nos travaux.

Un modèle de langage statistique facile à estimer et extrêmement performant en dépit de sa simplicité est le modèle n-gram ([Jurafsky and Martin, 2018]). Il permet d'estimer la probabilité pour un mot d'apparaître en fonction des n mots précédents. Il se base sur des statistiques calculées sur des corpus textuels et permet de donner du sens à une transcription automatique en trouvant une suite phonétique de mots cohérente d'un point de vue langagier.

Le calcul de la probabilité d'apparition d'un mot, en regardant les n mots précédents [5], est ce qu'on appelle l'hypothèse de Markov. Cette hypothèse émet l'idée que l'on peut deviner la probabilité d'apparition d'un mot sans avoir à regarder toute la phrase. Une séquence de deux mots donne un bigramme, de trois mots, un trigramme *etc...* La vraisemblance d'une suite de mots de 1 à n, c'est-à-dire "la probabilité *a priori* pour qu'un locuteur l'énonce" ([Zitouni, 2000]), se calcule comme suit :

$$P(w_1, w_2, ..., w_n) = \prod_{i=1}^{n} P(w_i / w_1 ... w_{i-1})$$

5. rarement > à 4

où $P(w_i/w_1...w_{i-1})$ est la probabilité du i-ième mot de la séquence, sachant tous les mots précédemment émis.
Ce qui donnera pour le calcul de trigrammes où n=3 :

$$P(w_1).P(w_2/w_1).P(w_3/w_1,w_2).....P(w_n/w_{n-1},w_{n-2})$$

Un modèle de langue peut calculer la probabilité pour un seul mot d'apparaître, c'est ce qu'on appelle l'unigramme,

$$P(w) = \frac{w_n}{w_{tot}}$$

avec w_n le nombre d'apparitions du mot et w_{tot} le nombre total de mots dans le corpus.
Il est également possible de mettre une même probabilité pour tous les mots, que l'on appellera alors des zérogrammes. Cela permet de tester l'efficacité des modèles acoustiques, sans influence du modèle de langue.

Cette méthode est néanmoins confrontée à un problème lors qu'il s'agit d'attribuer une probabilité à des mots ou séquences de mots qui n'ont jamais été observés dans le corpus d'apprentissage du modèle de langue. Celle-ci sera nulle. Une technique de lissage va permettre d'attribuer une probabilité non nulle à ces événements pour leur laisser une chance d'apparaître et de baisser la probabilité des événements les plus forts.

Le principe d'un système de reconnaissance vocale est de trouver la plus forte probabilité (argmax) pour une suite de mots, d'avoir généré un signal, en multipliant la plus forte probabilité pour un signal de donner une phrase (les modèles acoustiques), avec la plus forte probabilité de cette phrase d'exister dans la langue (le modèle de langage) :

$$argmax(P(w|y)) = argmax(P(y|w).P(w))$$

avec $P(y|w)$ représentant le modèle acoustique et $P(w)$ le modèle de langue. Le poids donné par le modèle de langue pour la reconnaissance vocale est important, notamment du fait que la parole n'est pas stable et est sujette à diverses variations (voir partie 1.2.4) et que la reconnaissance se basant sur l'acoustique seule est peu fiable ([Zitouni, 2000]).

Dictionnaire de prononciation

Le dictionnaire de prononciation est un lexique associant à chaque mot sa décomposition en phonèmes et les différentes prononciations qu'il peut avoir. Il peut également décrire une phrase avec sa décomposition en mots. Pour plus d'efficacité, la transcription phonétique doit tenir compte des liaisons faites dans la parole, mais aussi des différences de prononciation dûes aux accents. Par exemple, le nombre "six" pourra se prononcer /si/, /sis/ et /siz/ selon les phrases dans lesquelles il est prononcé :
"Six filles." [si]
"Il y a six ans." [siz]
"Un taux de 10,6." [sis]

Ce dictionnaire doit être élaboré avec justesse car le système ne pourra reconnaître que les mots qui y sont compris. De plus, il sert de base à la construction des modèles acoustiques [Adda-Decker and Lamel, 2000].

Modèles acoustiques

Tout comme pour les modèles de langue, différents types de modèles acoustiques existent. Les modèles acoustiques à base de réseaux de neurones (DNN), les modèles hybrides DNN-HMM [6], ou encore les modèles HMM-GMM [7] que nous allons développer ici.

Les modèles acoustiques permettent d'estimer la probabilité qu'une séquence de paramètres acoustiques génère un phonème. Le découpage en phonèmes permet aux systèmes de reconnaissance vocale d'être capables de reconnaître un large vocabulaire puisqu'au lieu de créer un modèle acoustique par mot, il découpe le mot en unités acoustiques finies. Par ailleurs, de cette manière, il est plus aisé d'oberver les différentes variations au niveau phonétique.

Un HMM (*Hidden Markov Models*) est un automate à état fini, qui contient des états (généralement trois ou plus, dans le cadre de la reconnaissance automatique de la parole) et sert à estimer la distribution de paramètres correspondant à

6. HMM : Hidden Markov Model
7. GMM : Gaussian Mixture Model

des monophones (un seul phonème) ou des triphones (ensemble de trois phonèmes, contextuels). Les HMM de Bakis sont utilisés pour représenter les différents états acoustiques d'un mot au cours du temps sans retour en arrière possible entre deux états. À chaque état est associé la probabilité de rester sur ce même état, ou de passer à un autre état et ce dépendamment du ou des états précédents. En effet, sur la même idée que les n-grams calculés pour le modèle de langue, les modèles de Markov peuvent être de différents ordres n selon si l'on prend en compte le ou les n état(s) précédent(s) pour déterminer l'état qui suit. Un HMM repose sur deux types d'états : les états cachés (phonèmes) et les états observables (paramètres acoustiques) dont la distribution est représentée par les GMM et qui permettent de deviner la "valeur" des états cachés. À un HMM à trois états est associé un phonème. À chaque état sont associés des GMM. Ces GMM sont une représentation des paramètres acoustiques présents dans notre corpus pour un état HMM.

La figure 1.11 [Le Blouch, 2009] montre la topologie d'un monophone. Le mono-

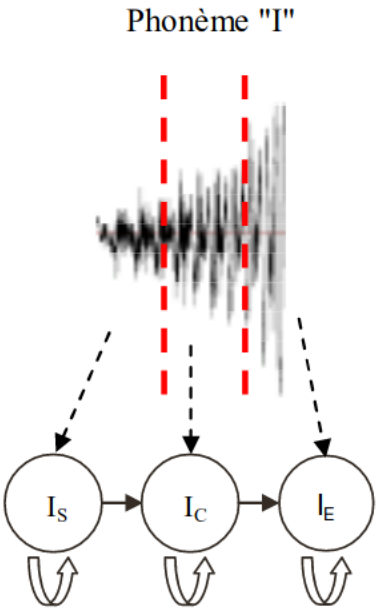


Figure 1.11 – Topologie d'un monophone.

Source: [Le Blouch, 2009]

phone est considéré comme stationnaire seulement en son centre. Le HMM permet de modéliser les trois parties qui le composent : le début, le centre et la fin. Le triphone va s'appuyer sur le même raisonnement à la différence qu'aux états de début

et de fin sont ajoutées des informations sur les phonèmes qui précèdent et qui suivent la réalisation du phonème en question. Par exemple, dans le mot "midi", le [m] et le [d] entourant le [i] vont venir influencer la réalisation du son [i].

L'apprentissage d'un modèle acoustique va consister à aligner un échantillon du signal à un phonème. Plusieurs itérations de cet alignement sont faites selon l'algorithme de Baum-Welch, (en réapprenant à chaque fois sur le dernier alignement effectué), au cours duquel l'algorithme d'expectation-maximization va aider à trouver l'estimation la plus probable des états cachés étant données les observations possibles. Une fois les paramètres optimaux obtenus ou le nombre d'itérations maximum voulu atteint, on obtient un modèle final.

1.2.4 Métriques d'évaluation

Lors d'un décodage, un système émet une hypothèse sur ce qui a été prononcé. Afin d'évaluer la performance d'un système de reconnaissance vocale, on compare l'hypothèse à la transcription, manuelle ou automatique, de référence. La métrique la plus utilisée est le WER pour Word Error Rate (Taux d'erreur mot). Il se calcule comme suit :

$$\frac{I + S + D}{W_{ref}}$$

avec I repésentant les insertions, S les substitutions et D les délétions, le tout divisé par le nombre de mots dans la référence. Une insertion correspond à l'ajout d'un mot dans l'hypothèse :
Ref : il *** mange
Hyp : il la mange
Une substitution est comptabilisée lorsqu'un mot est reconnu à la place d'un autre :
Ref : il mange
Hyp : île mange
Enfin, une délétion est l'inverse de l'insertion et correspond à la non détection d'un mot alors qu'il devrait y en avoir un :
Ref : il le mange
Hyp : il *** mange
Le WER est un pourcentage qui doit être le plus bas possible.

Il est également possible de calculer le SER - sentence error rate - selon le même principe, le PER - phoneme error rate - ou encore le CER - character error rate. Enfin, il est également possible de regarder le taux ou le nombre de mots correctement reconnus sur le nombre de mots total dans la référence ou encore la précision c'est-à-dire le taux de proximité entre la référence et l'hypothèse qui, à l'inverse du WER, doivent être les plus élevés possible.

Á ce jour, des systèmes comme Siri d'Apple ou Alexa d'Amazon peuvent obtenir une précision de plus de 90% pour des langues comme le français ou l'anglais, dans des conditions bien définies. Par ailleurs, dans le cadre de parole semi-préparée, les meilleurs systèmes actuels sont capables d'atteindre des WER aux alentours des 5%.

1.2.5 Les difficultés en reconnaissance automatique de la parole

Comme nous avons pu le voir, nous parvenons aujourd'hui à avoir des systèmes performants. Quelques difficultés ou contraintes persistent tout de même :

Les langues peu dotées ne profitent pas de systèmes aussi performants que ceux de langues très utilisées comme l'anglais, le français etc. qui sont parlées à travers le monde et qui permettent de disposer très vite de grands corpus écrits et oraux. Le manque de données fait que les nouveaux systèmes performants comme les systèmes à base de réseaux de neurones ne peuvent être utilisés trivialement.

Plusieurs types de parole sont dinstingués en reconnaissance de la parole : les mots isolés, la lecture ou encore la parole spontanée. La reconnaissance de mots isolés a été l'objet des premiers systèmes de reconnaissance vocale. La pause présente entre chaque son, que ce soit un mot ou un phonème, facilite l'alignement du son avec sa transcription puisqu'il n'y a pas d'effet de coarticulation. Les systèmes se sont ensuite concentrés sur la lecture, qui est une parole que l'on peut qualifier de préparée et qui n'a rien à voir avec du spontané. Les disfluences telles que les reprises, les répétitions, ou les mots de remplissage comme "euh" ou "hum" sont moins présentes en lecture qu'en parole spontanée [Nakamura et al., 2008]. Par définition, la parole spontanée est une parole non préparée à l'avance. [Silverman et al., 1992] men-

tionnent dans leur article le fait que les pauses en lecture sont plus courtes et apparaissent de façon moins aléatoire qu'en spontané. Pour finir, [Nakamura et al., 2005], [Nakamura et al., 2008] mentionnent que plus la parole est spontanée, moins la distance entre les phonèmes est grande et plus la variance (la dispersion des valeurs) pour chaque phonème augmente.

Une autre source de difficultés est la variabilité inter et intra-locuteur. L'accent régional, les différences morphologiques dues à l'âge, à un handicap ou au sexe sont autant de sources de variabilités dans la parole. Chez les enfants notamment, le tractus vocal et les cordes vocales sont moins longs que chez les adultes, ce qui donne des caractéristiques acoustiques spécifiques comme une fréquence fondamentale et une fréquence de formants plus hautes [Elenius et al., 2004]. La reconnaissance de parole de voix d'enfants est d'autant plus compliquée qu'une variabilité importante peut exister dans une même tranche d'âge, du fait que l'enfant est en évolution constante et tous les enfants n'apprennent pas au même rythme. La même chose se produit chez les personnes âgés pour qui le contrôle des articulateurs devient de moins en moins précis au cours du temps, la respiration moins bonne, la voix moins forte [Le Grand, 2012]...

La variabilité intra-locuteur fait que l'on ne prononce jamais un même mot de la même façon. Il y a par exemple des différences au niveau de la durée des phonèmes, mais aussi des différences dues à l'endroit où l'on se trouve ou à notre état de santé qui peuvent nous faire parler plus ou moins fort [8], plus ou moins aigu ou encore le fait de forcer sur la voix *etc*...

Enfin, une dernière difficulté est le bruit compris dans le signal. Celui-ci peut être dû à de l'écho, une trop forte réverbération, de la musique, des personnes qui parlent autour... Ce terme de bruit rassemble tout ce qui vient altérer le signal de parole et le rendre moins clair.

8. forcer sur sa voix en contexte bruité est appelé *effet Lombard*

1.3 Vers une reconnaissance automatique d'un corpus grand vocabulaire de sons du *human beatbox*

Les algorithmes de classification sont utilisés en reconnaissance vocale pour sélectionner le pattern (phonème, syllabe...) représentant le mieux le signal. Pour cela, différentes méthodes de classification peuvent être utilisées [Nasereddin and Omari, 2017] :

- Les modèles de Markov cachés (HMM)
- La déformation temporelle dynamique (DTW)
- Les réseaux bayésiens dynamiques (DBN)
- Les séparateurs à vaste marge (SVM)
- Les réseaux de neurones artificiels (ANN)
- Les réseaux de neurones profonds (DNN)

Des travaux ont déjà été effectués sur de la classification de beatbox. Nous les aborderons dans la partie *Classification de sons de human beatbox*. Nous aborderons ensuite les questions que nous nous posons sur la reconnaissance automatique de beatbox ainsi que nos hypothèses.

1.3.1 Classification de sons de *human beatbox*

Classification de *Human beatbox* et *Music Information Retrieval*

La *Music Information Retrieval* (MRI) est un domaine qui consiste à rechercher ou classer des corpus audio de musique en fonction d'une requête [Klapuri, 2019]. Cette requête peut être sous forme d'un fredonnement pour retrouver une mélodie, d'un extrait sonore d'instrument pour retrouver tous les enregistrements de la base de données qui le contiennent ou encore d'un enregistrement à un certain rythme pour retrouver toutes les musiques qui suivent le même tempo.

[Kapur et al., 2004] ont développé deux systèmes permettant d'utiliser le *beatbox* comme requête à un système MRI, particulièrement pour de la récupération de séquences rythmiques. Ils veulent donner la possibilité aux DJs de rechercher de la musique en utilisant les sons et le rythme de séquences de *beatbox*. Ceci est particulièrement valable pour des types de musique comme le Drum & Bass ou la House

pour lesquels la dimension rythmique surplombe la mélodie. Leur premier système est appelé *Bionic Beatboxing Voice Processor*. Il découpe les séquences de beatbox en sons individuels, les analyse, et retrouve les sons de batterie correspondants dans une base de données afin de transformer la voix du beatboxeur en instrument. Le deuxième système, *Musescape*, analyse le rythme et le style de musique (Dub, Drum&Bass...) et recherche les sons correspondant au même type de musique dans la base de données. Pour ce travail, ils ont utilisés trois grandes classes de sons de beatbox : des sons de grosse caisse (bass drum), des sons de caisse claire (snare drum) et des sons de charleston (hi-hat).

Pour la classification des sons, les paramètres acoustiques testés sont les suivants : *ZeroCrossings*, *Spectral Centroid*, *Spectral Centroid Rolloff*, *LPC* (*Linear Predictive Coding)* et MFCC. Le classifieur utilisé est un classifieur à base de réseau de neurones artificiel avec rétropropagation. Ils arrivent pour chaque classe de son à une classification avec une précision allant de 89.3% à 97.3% selon le type de paramètres extraits, les ZeroCrossings donnant les meilleurs résultats et les MFCC les moins bons.

Classification de cinq classes de sons de *human beatbox*

[Sinyor et al., 2005] ont également testé la classification de sons de beatbox. Ils se sont concentrés sur cinq classes :

— kick drum ;
— closed hihat ;
— open hihat ;
— p-snare ;
— k-snare.

Plusieurs paramètres acoustiques ont été extraits, parmi lesquels les : *Spectral Centroid*, *root mean square* et *zero-crossing*. Un classifieur utilisant la méthode des k plus proches voisins a été utilisé. L'utilisation des paramètres acoustiques avec un classifieur à 1 plus proche voisin a donné une précision de 98.15%.

Adaptation d'un classifieur existant pour de la classification de sons de *human beatbox*

[Hipke et al., 2014] ont développé un système appelé *BeatBox*, qui est une interface représentant un pad [9] électronique. A chaque touche du pad est associée une classe de sons de beatbox et un enregistrement du son de batterie correspondant. Leur système permet à un utilisateur de définir autant de classes qu'il le souhaite, en enregistrant au moins une version beatboxée par classe. Ensuite, il peut utiliser le système en *live*. Chaque son beatboxé est alors transformé par le son de batterie correspondant à la classe reconnue. Ce système peut être utilisé pour des sons isolés et pour des séquences.

Les paramètres acoustiques utilisés sont les *spectral centroid* et les *Root Mean Square*. Ils ont utilisé un classifieur utilisant la méthode des k plus proches voisins. L'évaluation du système a été faite sur une échelle de Likert. Les sept participants ont reconnu l'efficacité du système en donnant une note moyenne au système de 2.85/7 selon l'échelle de Likert [10].

1.3.2 Reconnaissance automatique de sons de *human beatbox*

À notre connaissance, seule l'étude de [Picart et al., 2015] relate le développement d'un système de reconnaissance automatique de sons de *beatbox*. Leur corpus a été enregistré par deux *beatboxeurs*et contient 9000 enregistrements, dont 1835 sons percussifs et 1579 imitations d'instruments. Ils n'ont pas cherché à reconnaître les sons percussifs un à un mais à les cataloguer dans cinq grandes classes : cymbal, hihat, kick, rimshot and snare. Les instruments représentés sont : la guitare électrique (deux types de sons), la guitare basse, le saxophone, la trompette (3 types de sons), le didgeridoo et l'harmonica. Les sons percussifs et les imitations d'instruments ont fait l'objet de deux systèmes séparés.

Leurs systèmes ont été entraînés grâce à la boîte à outils HTK. Plusieurs para-

9. Un pad prend souvent la forme d'une tablette composée de plusieurs touches sur lesquelles ont vient taper avec une baguette pour reproduire les sons de batterie enregistrés.

10. 1 étant le meilleur et 7 le moins bon

métrages ont été testés :

- 18 à 22 MFCC + les dérivées premières et secondes ;
- une fenêtre temporelle de 10ms ;
- un décalage des fenêtres de 2 à 10ms ;
- de 3 à 21 états HMM ;
- l'extraction de MFCC, LPC (*Linear Predictive Coding)*, LPCC (*Linear Predictive Cepstral Coefficients)*, PARCOR *(PARtial CORrelation)* ou PLP *(Perceptual Linear Prediction* pour les paramètres acoustiques.

Leurs systèmes les plus performants utilisent un décalage des fenêtres de 2ms, 21 états HMM et 22 MFCC pour les sons percussifs contre 18 MFCC pour les instruments. Les WER obtenus sont de 9% pour la reconnaissance des cinq classes de sons percussifs et 41% pour la reconnaissance des neuf instruments.

1.3.3 Questionnements pour ce travail

Dans leur essai de système de reconnaissance automatique du *beatbox*, [Picart et al., 2015] possédaient 9000 événements musicaux mais ne cherchaient à reconnaître que cinq grandes catégories de sons percussifs (cymbal, hihat, kick, rimshot and snare) et neuf instruments. Ces deux ensembles avaient fait l'objet de tests séparés. De même, les différentes classifications de sons de beatbox effectuées par [Kapur et al., 2004] et [Sinyor et al., 2005] ne cherchaient à reconnaître que trois à cinq classes de sons de beatbox. Cependant, l'extraction de différents paramètres acoustiques ont été testés, parmi lesquels des paramètres acoustiques dont l'usage est possible en reconnaissance vocale. [Bezdel and Bridle, 1969] ont par exemple utilisé l'extraction de *zero-crossings* et [Wijoyo, 2011] l'extraction de LPC.

Au vu des résultats précédents, nous pouvons dès à présent nous demander si l'extraction de MFCC, étant les coefficients les plus utilisés en reconnaissance de la parole, nous permettra d'obtenir un système de reconnaissance de sons de beatbox efficace étant donné que ce ne sont pas toujours les paramètres acoustiques les plus performants pour les classifieurs de sons de beatbox dont nous avons parlé.

Nous nous interrogeons également sur la possibilité d'utiliser la boîte à outils Kaldi, développée pour de la reconnaissance de la parole et sur la reconnaissance

d'un corpus de sons de beatbox grand vocabulaire. De plus, l'annotation du corpus grand vocabulaire utilise une écriture décrivant très précisément les différentes articulations, ce qui fait que nous avons quelques sons très ressemblants à l'oreille. Nous nous demandons si le système sera capable de les discriminer.

Pour finir, nous pouvons considérer le beatbox comme une langue peu dotée, c'est-à-dire avec peu de ressources. Le manque de ressources complique la tâche de reconnaissance. Pour ce travail, nous sommes surtout concernés par le manque de données textuelles ne nous permettant pas d'entraîner un modèle de langue efficace. Nous nous demandons si notre système sera capable de fonctionner malgré cela.

Chapitre 2

Matériel et méthodes

Nous présenterons dans ce paragraphe les deux corpus avec lesquels nous avons développé notre outil. Un tableau récapitulatif se trouve en Figure 2.3. Nous présenterons également le déroulement de l'annotation des données ainsi que les outils utilisés pour la reconnaissance vocale.

2.1 Les corpus

2.1.1 Corpus petit vocabulaire

Cette base de données est la première enregistrée par A. C. et A. P. sur Grenoble (voir tableau récapitulatif en Figure 2.3). Elle a été enregis-trée à l'occasion de la découverte de l'existence de l'écriture *Vocal Grammatics* par une équipe du GIPSA-lab qui travaillait sur le *beatbox*. La scéance d'enregistrement s'est déroulée au GIPSA-lab en 2018 et avait pour but de faire une première ana-lyse physiologique des comportements des *beatboxeurs*, en lien avec la description articulatoire mise en avant dans l'écriture *Vocal Grammatics*.

Les sujets *beatboxeurs* sont deux hommes. A. C., 31 ans, est *beatboxeur* amateur. A. P., 25 ans, est *beatboxeur* professionnel. A. et A. étaient tous deux équipés de capteurs à l'intérieur de la bouche et sur les lèvres pour des mesures d'articulographie électromagnétique 3D (EMA3D). Des enregis-trements de contact glotique (EGG) et de pléthysmographie respiratoire (VisuResp) ont été effectués en même temps. Ces mesures ne seront pas utilisées ici mais sont

importantes à prendre en compte car les capteurs positionnées sur les articulateurs peuvent gêner les *beatboxeurs* pour la bonne prononciation des sons.

Le corpus est composé d'enregistrements de sons de *beatbox* isolés et répétés (son1, pause, son1, pause etc...) et des enregistrements de séquences répétées.

A. C. a enregistré 14 sons différents et A. P. 17 avec à chaque fois 3 à 4 enregistrements contenant chacun une dizaine de répétitions du son. Le vocabulaire et les sons enregistrés par *beatboxeur* sont à retrouver en annexe A.
Nous disposons d'environ 22 minutes d'enregistrement. Les sons ont été enregistrés avec une fréquence d'échantillonnage à 22050 Hz, en mono, 16 bits et sous le format wav.

2.1.2 Corpus grand vocabulaire

Le deuxième corpus dont nous disposons est une base de données enregistrée spécifiquement pour ce travail de recherche par les deux *beatboxeurs* A. C. et A. P.. Le vocabulaire a été étendu à 80 sons.

La session s'est déroulée dans un studio d'enregistrement à Césaré (centre de création musicale), à Reims. Les *beatboxeurs* disposaient de 6 microphones différents pour les enregistrements. Deux stands ont été mis en place que nous appellerons A et B.

Stand A	Stand B
Enregistrements non encapculés	Enregistrement encapsulé (voir Figure 2.1)
5 microphones	1 microphone

Le stand A contient 5 micros de qualités différentes et disposés différemment :

— Un Brauner VM1, microphone statique [1], avec filtre anti-pop [2], positionné à

1. Les microphones statiques et dynamiques ont des sensibilités différentes. Le microphone dynamique est hypersensible, capte beaucoup de choses et très souvent utilisé en studio. Le statique, à l'inverse, est beaucoup moins sensible et sert plutôt pour les conditions de live.

2. Un filtre anti-pop est simplement un filtre placé devant le microphone qui vient atténuer

10cm de la bouche du *beatboxeur*.

— un DPA 4006, microphone statique d'ambiance, positionné à 50cm de la bouche du *beatboxeur*

— un DPA 4060, microphone cravate, positionné à 10cm de la bouche du *beatboxeur*, au niveau du manubrium

— deux Shure SM58, microphones dynamiques, véritables références dans le *beatbox* live selon le site *humanbeatbox.com*[3], positionnés à 10cm et 15cm de la bouche du beatboxeur.

L'enregistrement de ces cinq microphones se fait simultanément.

Le stand B ne comprend qu'un microphone, un Shure Beta 58, microphone dynamique. La bouche du *beatboxeur* est située à environ 1cm.

FIGURE 2.1 – Encapsulement de microphone

Source: Photo de l'*Instituto Voz Cultura e Conhecimento* publiée sur Flickr

Nous disposons de 80 sons unitaires et de 52 séquences contenant uniquement des sons compris dans les enregistrements unitaires (La liste des sons est disponible en annexe). Pour chacun des sons unitaires nous avons les enregistrements des 6 microphones. A. C. n'étant pas professionnel, il n'a enregistré que 56 sons unitaires sur les 80. A. a enregistré les 80 sons unitaires, ainsi que les séquences. Pour chacun des sons nous disposons d'une dénomination dans la culture du *beatbox* (ou proposés par A. et A.), son écriture phonétique et son

(voire supprimer) les sons perçus comme des bruits (comme les plosives par exemple) dûs à un déplacement d'air plus fort que d'ordinaire.

3. www.humanbeatbox.com/articles/gearing-up-for-beatbox/

écriture phonétique abrégée (voir annexe B). Par exemple, un *Classic kick* est un *Bi-labial_Explosif* alors noté *BiLa_Exp*. Ces noms nous ont été fournis par les *beatboxeurs*. Une version pictographique de l'écriture *Vocal Grammatics* est en cours d'élaboration.

Les enregistrements d'Adrien contiennent une dizaine de répétitions de chaque son. Ils sont prononcés de façon normale ou avec des variations (de durée, de hauteur...). Les enregistrements d'Antoine comprennent trois blocs :

— Un bloc "normal", où le *beatboxeur* produit le son comme il le produit habituellement.

— Un bloc exécuté comme un débutant, où le *beatboxeur* essaie de reproduire ce qu'un débutant en *beatbox* pourrait produire face au son proposé.

— Un bloc composé de variations de durée ou de hauteur.

Il y a donc pour chaque *beatboxeur* une dizaine de répétitions d'un même son produits selon les différents modes "normal", "débutant" ou "avec variations". Les sons sont en 44100Hz, 16 bits, mono, wav.

La représentation de l'arborescence des fichiers résume la construction du corpus :

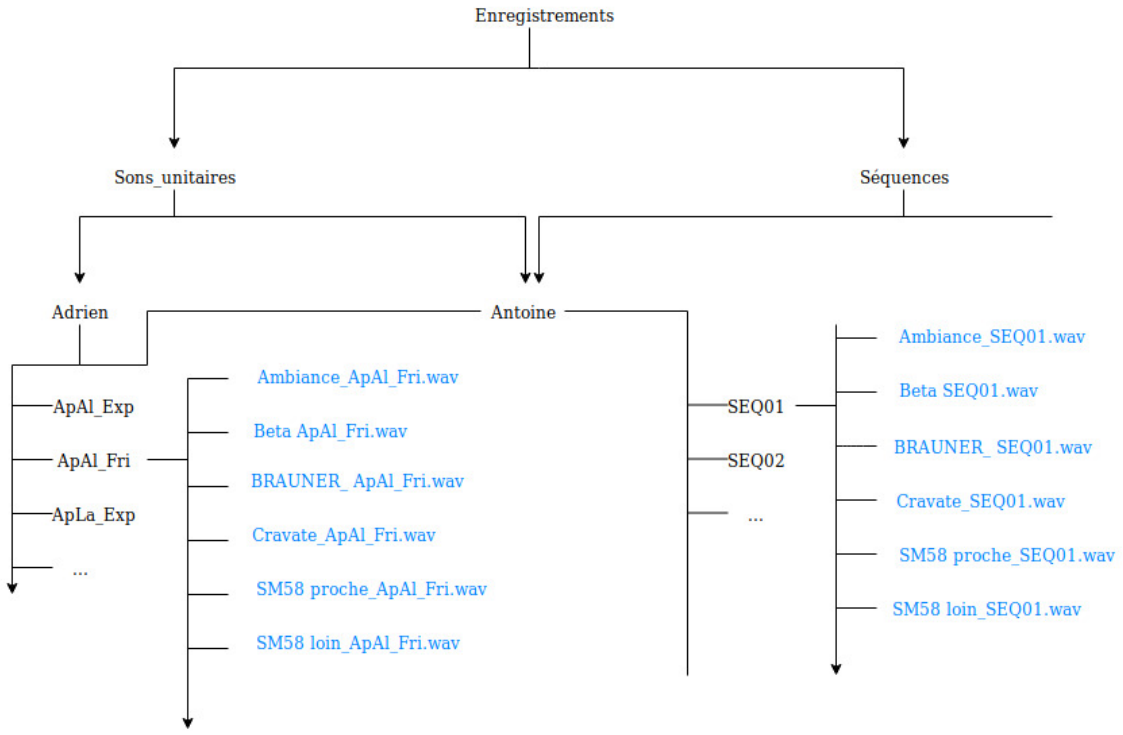


Figure 2.2 – Arborescence des fichiers du corpus

	Corpus petit vocabulaire		Corpus grand vocabulaire	
Date	2018		2019	
Durée d'enregistrement	22mn		3h25	
Participants	A. C.	A. P.	A. C.	A. P.
Niveau	Amateur	Professionnel	Amateur	Professionnel
Nombre sons unitaires	14/19	19/19	56/80	80/80
Nombre répétitions	environ 10		environ 10	
Nombre séquences	3/3	3/3	0/52	52/52
Annotation	Phonétique		Vocal grammatics	
Sons	22050 Hz 16 bits mono wav		44100 Hz 16 bits mono wav	
Microphones	1 microphone		6 microphones dont 1 encapsulé	
Détail			1 à 2 blocs Normal Variations	3 blocs Normal Débutant Variations

FIGURE 2.3 – Tableau récapitulatif de la constitution des corpus

2.2 Annotation des données

2.2.1 Selon l'API

Le premier corpus a été annoté à la main. Deux transcriptions ont été faites : une version avec le nom du son produit dans la culture *beatbox* et une deuxième inspirée de l'alphabet de Stowell. L'annotation du corpus a fait l'objet d'un stage précédent. Elle a été effectuée avec Praat. La transcription phonétique de la séquence *boots and cats* n'était pas présente et a donc été rajoutée selon le même principe.

2.2.2 Avec l'écriture *Vocal Grammatics*

Le premier corpus n'a pas été annoté selon l'écriture *Vocal Grammatics*.

Concernant le corpus grand vocabulaire, nous avons tout d'abord voulu le transcrire avec Praat, en nous inspirant de ce qui avait été fait pour le premier corpus. Nous avons été confronté à un problème avec les différents blocs qui composent chaque enregistrement. Si nous avions annoté chaque bloc selon les modes "normal", "débutant" et "avec variations" avec à l'intérieur le nombre de répétitions du son, nous n'aurions pu avoir un extrait de chaque bloc pour l'apprentissage et le développement de notre système. En effet, le découpage d'un enregistrement en deux segments, un pour l'apprentissage et un pour le test du système, ne peut se superposer. Les sons utilisés pour le développement doivent être inconnus du système et ne doivent pas être utilisés pour l'apprentissage.

Mélange aléatoire des boxèmes

Nous avons donc décidé dans un premier temps de découper chaque son selon le nombre de boxèmes qu'il comprend. Une fois un fichier son créé par boxème, il nous a été alors possible de mélanger aléatoirement les fichiers et de construire deux fichiers différents : un fichier "train.wav" et un fichier "dev.wav" pouvant chacun comprendre des boxèmes des différents blocs "normal", "débutant" ou "avec variations".

Ceci a pu être fait de façon semi-supervisée. Pour le découpage, nous avons utilisé le logiciel libre de d'édition de sons numériques *Audacity*. L'option *mod-script-pipe* pouvant être activé à la compilation du logiciel nous a permis de contrôler le logiciel via un script Python. Nous avons choisi ce logiciel pour sa fonction *soundfinder* permettant de découper un son en fonction de l'énergie. Nous devions préciser comme réglage le nombre de décibels en dessous duquel le son doit être considéré comme un silence et un temps avant et après le son détecté permettant de poser des frontières et créer un label. Nous pouvons observer sur la figure 2.4 que les frontières sont assez larges. Un numéro a été donné à chaque label.

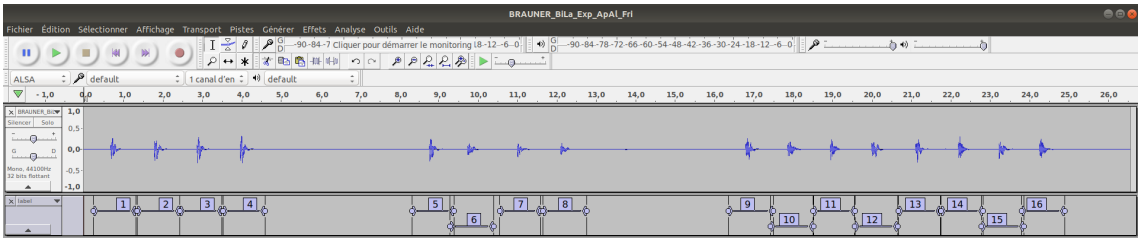


FIGURE 2.4 – Découpage d'un enregistrement avec la fonction SoundFinder d'Audacity

Les labels permettent ensuite de faire un export multiple et de découper l'enregistrement selon les temps de début et de fin de chaquel label. Ici, nous avions donc 16 fichiers wav en sortie. Le nom de chaque fichier wav était formé du nom du son (ici BiLa_Exp_ApAl_Fri) + le numéro du label. Les enregistrements du stand A étant effectués en même temps, le fichier de label a été créé sur un premier son et réutilisé pour les 4 autres.

Voici le déroulement de la division des sons pour les 5 microphones du stand A avec, en vert, ce qui est effectué automatiquement grâce au script et en bleu et gras ce qui est fait manuellement : Pour le premier microphone :

- 1. Ouverture de la fenêtre permettant de sélectionner le son.
- **2. Sélection du son voulu.**
- 3. Sélection du son et application de la fonction *soundfinder* selon les paramètres choisis.
- **4. Vérification et modification des frontières si besoin.**
- 5. Export multiple.

- **6. Nommer les fichiers créés selon le nom du son et les labels et les enregistrer dans un dossier au nom du microphone.**
- **7. Valider chacun des sons.**
- 8. Ouverture d'une fenêtre pour enregistrer le fichier de labels.
- 9. Enregistrer le fichier de labels.
- 10. Fermer la fenêtre.
- **11. Choisir d'enregistrer le projet ou non.**

Pour chacun des 4 microphones suivants :

- 12. Ouverture d'une fenêtre pour la sélection d'un nouveau son.
- **13. Sélection du son.**
- 14. Ouverture d'une fenêtre pour la sélection du fichier de labels créé à la deuxième étape.
- **15. Sélection du fichier de labels.**
- 16. Répétition des étapes 4 à 11 de la liste précédente.

Le temps de vérification de labellisation du son est chronométré. La création du fichier de labels a été faite à chaque fois sur l'enregistrement provenant du microphone Brauner. Pour le son du microphone encapsulé, seules les étapes 1 à 11 sont effectuées.

Le mélange aléatoire des fichiers obtenus et la création des fichiers *train.wav* et *dev.wav* a été effectué grâce à un script en Bash. Nous avons choisi de toujours avoir le même nombre de répétitions d'un son dans le fichier *train.wav* afin qu'un son ne soit pas sur-représenté à l'apprentissage. Le fichier *dev.wav* contient donc tous les autres sons, concaténés aléatoirement.

Mélange des boxèmes selon le mode normal, débutant ou avec variations

Un deuxième découpage a été effectué dans un second temps. Sur le premier découpage nous n'avions pas de fichier indiquant quel son se trouvait dans*train.wav* ou *dev.wav*. Nous avons donc ajouté une option qui créé un fichier "historique" permettant de savoir quel boxème se trouve à quel endroit. Ce deuxième découpage a été fait en réutilisant les fichiers de labels existants et en changeant les noms des labels par "normal1", "normal2", "deb3", "deb4", "varia5"... Nous avions donc cette fois l'information sur le mode utilisé pour la réalisation du boxème (normal, débu-

tant ou avec variations) en plus du numéro du label. Ceci nous a permi de générer par la suite un fichier *train* et un fichier *dev* pour chaque mode : *normal_train.wav*, *normal_dev.wav*, *deb_train.wav*, *deb_dev.wav*, *varia_train.wav*, *varia_dev.wav*.

2.2.3 Construction des fichiers pour Kaldi

Kaldi nécessite au moins 4 fichiers pour pouvoir fonctionner, en plus de la création d'un identifiant par utterance, par locuteur et par fichier son :

- un fichier "text" qui regroupe des informations telles que l'identifiant de l'utterance et sa transcription.
- un fichier "utt2spk" qui donne pour chaque identifiant d'utterance l'identifiant du locuteur qui l'a prononcée.
- un fichier "spk2utt" qui est l'inverse du fichier précédent et donne pour chaque locuteur toutes les phrases qu'il a prononcées.
- un fichier "wav.scp" qui donne, pour chaque identifiant de son, le chemin absolu
- un fichier "segments" (facultatif) donne pour chaque identifiant d'utterance le chemin relatif du son où se trouve cette utterance ainsi que le temps de début et de fin de l'utterance dans le son. (non obligatoire si un fichier son représente une utterance)
- un fichier "spk.txt" (facultatif) qui regroupe des informations complémentaires sur les locuteurs.

Nous avons réadapté un script Python permettant de créer ces fichiers automatiquement à partir d'un fichier *Excel*. Ce tableur contient toutes les informations utiles concernant le corpus. On y trouve le nom, prénom, âge et sexe du *beatboxeur* ainsi que leur niveau d'expertise en *beatbox* (débutant, amateur ou professionnel). Viennent ensuite le type de verbalisation (son unitaire ou séquence), le microphone utilisé, le son de *beatbox* en écriture phonétique raccourcie, la dénomination du son dans la culture *beatbox*, le nombre de répétitions du son, le style (normal, débutant ou avec variations), le chemin vers le fichier *wav*, la transcription et enfin les temps de début et de fin de l'utterance transcrite. Les dénominations des sons en phonétique raccourcie sont tirés des noms de dossiers. La transcription comprenant n répétitions d'un même mot pour les sons unitaire est écrite automatiquement sui-

vant le nombre de répétitions voulu et le nom en phonétique. Les temps de début et de fin de chaque son ont été calculés grâce au module "Audiosegment" de Python.

Kaldi demande également la construction de fichiers nécéssaires à la création d'un modèle de langue tels qu'un lexique ou encore la liste des phonèmes et la liste des "phonèmes représentant le silence", comme "PAUSE". Nous avons construit le lexique en mettant la transcription phonétique comme représentant du mot et le nom du son en *beatbox* à sa droite. Ce choix nous a permis de pouvoir voir l'articulation produite et de pouvoir voir si deux sons confondus ont la même base d'articulation.

2.3 Matériel pour la reconnaissance vocale

2.3.1 Division des données

Corpus petit vocabulaire

Un système de reconnaissance vocale repose sur la division des données pour l'apprentissage et pour le test. Pour le premier corpus, la partie apprentissage de notre corpus pour A. comprend trois à quatre enregistrements des 14 sons produits, répétés une dizaine de fois par enregistrement. La partie développement comprend un enregistrement de chacun des sons avec une dizaine de répétitions du son à l'intérieur. Pour ce qui est des séquences, nous avons 3 séquences différentes, avec au moins deux enregistrements de ces séquences. La partie test comprend un enregistrement de chacune des 3 séquences.
Pour A., la partie apprentissage comprend trois à quatre enregistrements des 17 sons, répétés une dizaine de fois par enregistrement. La partie test est similaire à celle d'Adrien. Pour ce qui est des séquences, nous avons 3 séquences différentes enregistrées et répétées avec au moins deux exemplaires de ces séquences. La partie test est encore une fois similaire à celle d'Adrien. Enfin, A. a également enregistré un son de trompette et de diphonie qui se trouvent dans l'apprentissage et dans le test.

Corpus grand vocabulaire

Les premiers tests sur cette base de données avaient pour but de voir si un microphone donnait de meilleurs résultats qu'un autre. Nous avons donc trié nos données pour avoir dans un premier temps un microphone pour l'apprentissage et le même microphone pour le décodage test.

Ensuite, nous avons testé la reconnaissance sur un microphone avec pour données d'apprentissage tous les autres microphones excepté le microphone encapsulé.

Nous avons également testé l'incidence des modes "normal", "débutant" et "avec variations" en faisant les mêmes opérations que précédemment. Dans ce cas, la partie apprentissage des données contient, pour un microphone, un fichier *xx_train.wav* (avec xx représentant "normal", "deb" ou "varia") et pour le test un fichier *xx_dev.wav*.

Nous avons également fait un apprentissage contenant tous les fichiers *xx_train.wav* de tous les micros et fait un décodage sur le fichier *xx_dev.wav* d'un microphone pour voir si l'un des modes de réalisation dégrade les performances du système. Nous y reviendrons en partie III.

2.3.2 Boîtes à outils utilisées

Nous nous sommes servi de la boîte à outils Kaldi, utilisée pour de la reconnaissance automatique de la parole, afin de créer le système de reconnaissance du *beatbox*. C'est un outil état de l'art qui a été développé en 2009 lors du workshop *Low Development Cost, High Quality Speech Recognition for New Languages and Domains* de l'université Johns Hopkins. Il est aujourd'hui maintenu et développé par Daniel Povey et nécessite d'être à l'aise avec les concepts relatifs à la reconnaissance vocale et au parcours de graphes. Cet outil est open-source et gratuit et utilisable sous Linux (de préférence) et Windows ([Povey et al., 2011]).

Kaldi nécessite certains fichiers en entrée pour pouvoir fonctionner. Nous verrons dans la partie "matériel pour la préparation des données" ci-dessous comment ils sont construits.

Le modèle de langage est créé grâce au toolkit SRI Language Modeling développé au SRI Speech Technology and Research Laboratory depuis 1995 [4] qui sert à calculer des modèles de langage statistiques.

4. http ://www.speech.sri.com/projects/srilm/

Chapitre 3

Résultats de la reconnaissance automatique

Les fichiers "bruts" des résultats indiquant le nombre de substitutions, insertions, délétions, le nombre de mots et le nombre de phrases sont disponibles en annexe. Nous utiliserons, pour une lecture plus facile, les abréviations suivantes :

- MA : Modèles acoustiques
- mono : monophones
- mono + LDA et MLLT : monophones avec Δ Δ ajoutés aux MFCC et adaptations LDA et MLLT
- tri2a : triphones avec Δ Δ ajoutés aux MFCC
- tri2b : triphones avec Δ Δ ajoutés aux MFCC et adaptations LDA et MLLT

Les réglages par défaut dans Kaldi sont les suivants :

- 13 MFCC (Nous avons utilisé l'énergie.) ;
- calcul des $\Delta\Delta$ pour les triphones ;
- 3 états HMM pour les phonèmes non silencieux et 5 états HMM pour les phonèmes représentant un silence.

3.1 Reconnaissance à base d'apprentissage sur petit vocabulaire

Au début de ce projet nous ne disposions que de la base de données avec un petit vocabulaire. Celle-ci a fait l'objet d'une annotation lors d'un stage précédent. Vous

pouvez retrouver le vocabulaire utilisé en annexe A.

Le corpus d'apprentissage du système contient entre 3 et 4 enregistrements composés d'une dizaine de répétitions pour chaque son unitaire de chaque locuteur et 3 séquences enregistrées par les deux locuteurs à un rythme plutôt lent (voir en annexe pour un tableau récapitulatif). Le corpus de test comporte un enregistrement d'une dizaine de répétitions pour chaque son de chaque locuteur et un enregistrement de chacune des trois séquences par les deux locuteurs. Les enregistrements d'Antoine étant bruités, nous avons utilisé la fonction *réduction de bruit* du logiciel *Audacity* afin de les corriger.

3.1.1 Résultats

	Nombre mots reconnus	Substitutions	Insertions	Délétions	WER
A.	164/188	3,19%	3,72%	9,57%	16,49%
A.	215/271	6,64%	4,80%	14,02%	25,46%
Total	379/459	5,23%	4,36%	12,20%	21,79%

Figure 3.1 – Résultats du décodage sur le corpus petit vocabulaire avec des MA mono

	Nombre mots reconnus	Substitutions	Insertions	Délétions	WER
A.	158/188	4,26%	2,66%	11,70%	18,62%
A.	220/271	4,80%	5,17%	14,02%	23,99%
Total	378/459	4,58%	4,14%	13,07%	21,79%

Figure 3.2 – Résultats du décodage sur le corpus petit vocabulaire avec des MA tri2a

Le WER individuel est plus élevé pour le *beatboxeur* professionnel (A.) que pour le *beatboxeur* amateur (A.) : +10 points. Le nombre global de mots correctement reconnus est sensiblement le même que l'on soit sur des mono ou des tri2a. Le WER, lui, est exactement le même : 21,79%. Le passage aux triphones a fait

baisser le taux de substitutions et d'insertions mais a fait augmenter le nombre de délétions, c'est-à-dire le nombre de boxèmes non détectés ou pas du tout reconnus.

Pour les monophones, les sons les plus confondus sont :

— la cymbale [ts] qui est confondue huit fois avec un classic kick hardware[P] ;
— la classic snare drum [k] qui est confondue quatre fois avec un classic kick hardware [P] ;
— le reversed classic kick [^P] qui est confondu trois fois avec un classic kick hardware [P].

Le son le plus confondu est le classic kick hardware puisqu'il est confondu 18 fois avec un autre son sur un total de 24 substitutions.

Pour les triphones :

— la cymbale est confondue 4 fois avec un classic kick hardware ;
— le classic kick est confondu 3 fois avec le classic snare drum ;
— le classic kick est confondu 2 fois avec un dry snare hardware [T !].

La cymbale, le classic kick et le classic snare drum sont également souvent non reconnus. Le nombre de délétions avec des modèles acoustiques monophones les concernant est de 20, 6 et 10, respectivement et de 23, 8 et 14 pour des modèles acoustiques triphones.

3.2 Impact du type de microphone sur la qualité de la reconnaissance

Une fois le corpus grand vocabulaire enregistré, nous nous sommes concentrés dessus. La grammaire *Vocal Grammatics* ayant servi à l'annotation est à retrouver en annexe. Nous avons souhaité voir dans cette partie si un des microphones rend le système plus efficace qu'un autre ou si, à l'inverse, l'un d'entre eux dégrade fortement les résultats.

Nous avons effectué un apprentissage par microphone. Le test du système a été effectué sur le reste des sons enregistrés avec le microphone ayant servi à l'apprentissage. Les données d'apprentissage contiennent toutes le même nombre de répétitions : 6. Le nombre de répétitions des sons dans le corpus de test varie

entre 1 et 12. Nous avons calculé des modèles acoustiques monophones et triphones tri2a et tri2b. Les triphones tri2a incluent des adaptations LDA-MLLT. La LDA, *Linear Disciminant Analysis*, est une matrice utilisée pour réduire la dimension du vecteur de paramètres acoustiques, en sélectionnant les principaux composants [Haeb-Umbach and Ney, 1992]. La MLLT, *Maximum Likelihood Linear Transform* est une méthode de décorrélation des paramètres acoustiques [Rath et al., 2013].

	Micro ambiance	Micro encapsulé	Micro Brauner	Micro cravate	Micro SM58 éloigné	Micro SM58 proche
WER global sur mono	53,93%	70,79%	67,47%	51,63%	50,68%	54,46%
WER global sur tri2a	53,93%	66,87%	62,22%	49,84%	46,59%	53,31%
WER global sur tri2b	51,21%	68,73%	55,09%	57,64%	81,74%	106,30%

Figure 3.3 – Résultats des tests sur chaque microphone du corpus grand vocabulaire avec des MA mono, tri2a et tri2b

Le classement des microphone est, du meilleur au plus mauvais :

— le microphone SM58 éloigné ;
— le microphone cravate ;
— le microphone d'ambiance ;
— le microphone SM58 proche ;
— le microphone Brauner ;
— le microphone encapsulé.

Les quatre premiers microphones ont plus ou moins les mêmes WER avec des MA mono. Avec les MA tri2a il y a une différence d'environ 7 points entre le SM58 éloigné et le microphone d'ambiance. Avec des modèles acoustiques tri2b, le WER est totalement différent de ce qu'il était sur les monophones. Il est parfois amélioré, parfois grandement dégradé, comme pour le microphone SM58 proche qui passe de 54,46% à 106,30% de WER.

Le microphone le moins bon est le microphone encapsulé, avec un WER entre

66,87% et 70,79% selon les modèles acoustiques utilisés.

Les taux de substitutions, insertions et délétions pour le microphone SM58 éloigné (le meilleur) et le microphone encapsulé (le moins bon) avec des MA mono sont visibles dans la Figure 3.4.

	Substitutions	Insertions	Délétions
SM58 éloigné	30,95%	13,33%	6,40%
microphone encapsulé	45,92%	14,65%	10,22%

FIGURE 3.4 – Substitutions, insertions et délétions pour le meilleur microphone et le moins bon

Nous pouvons observer ici que les substitutions, aussi bien que les insertions et délétions, sont plus nombreuses sur le microphone encapsulé que sur le microphone SM58 éloigné.

	Micro ambiance	Micro encapsulé	Micro Brauner	Micro cravate	Micro SM58 éloigné	Micro SM58 proche
A. (amateur)	53,57%	75,61%	62,70%	58,47%	52,78%	57,94%
A. (pro) (professionnel)	54,07%	69,16%	69,19%	49,22%	49,93%	53,21%

FIGURE 3.5 – WER par locuteur sur les différents microphones avec des modèles acoustiques monophones.

Le nombre de points de WER entre A. et A. s'élève à 9 points avec le microphone cravate. Nous pouvons voir que le *beatboxeur* le mieux reconnu n'est pas forcément toujours le *beatboxeur* professionnel.

3.3 Influence de la variabilité dans la production

Notre corpus grand vocabulaire est constitué d'enregistrements exécutés sous différents modes : normal, débutant et avec variations. Au vu des résultats précédents,

nous avons souhaité tester les performances d'un système entraîné sur chacun de ces modes, tout comme nous l'avons fait pour les différents types de microphones.

3.3.1 Tests avec le microphone SM58 éloigné

Nous nous sommes concentrés dans un premier temps sur le microphone SM58 éloigné car c'est le microphone qui a donné les meilleurs résultats en partie 3.2. Nous avons comme données d'apprentissage, trois sous-corpus : un sous-corpus de sons en mode débutant, un sous-corpus de sons en mode normal et un sous-corpus de sons en mode variations avec deux répétitions de chaque son par enregistrement. Les données de test sont constituées de trois sous-corpus : un débutant, un normal, et un avec variations également.

	Normal	Débutant	Variations
Wer mono	78,33%	63,98%	75,36%
Wer mono + Δ Δ et adaptations LDA MLLT	87,22%	55,90%	76,45%

Figure 3.6 – Résultats des décodages par mode pour le microphone SM58 éloigné

Le mode de production avec le meilleur WER est le mode débutant. Le mode normal est celui qui semble poser le plus de problèmes, autant sur monos que sur les monos+LDA et MLLT. Nous observons une dégradation des résultats avec les adaptations LDA et MLLT, excepté pour le mode débutant. Les trois modes donnent des décodages avec des WER plutôt élevés (>50%). Il y a une différence d'environ 14 points sur les monophones entre le mode débutant et le mode normal.

3.3.2 Tests avec le microphone SM58 proche

Les corpus d'apprentissage utilisés pour les décodages en Figure 3.6 ne contenaient que les sons enregistrés avec le microphone SM58 éloigné. Les systèmes de reconnaissance ont une meilleure performance lorsque le nombre de données d'ap-

prentissage augmente. Nous avons donc utilisé cette fois-ci tous les microphones [1] pour constituer les sous-corpus d'apprentissage. Par exemple, pour le mode normal, nous avons en apprentissage tous les enregistrements en mode normal de tous les microphones excepté le microphone encapsulé, à raison de 2 répétitions du son par enregistrement.

Les corpus de test ont été constitués de la même façon que pour le SM58 éloigné, c'est-à-dire avec tous les enregistrements test du microphone SM58 proche, répartis selon les trois modes de production.

	Normal	Débutant	Variations
Wer mono	34,32%	34,36%	35,91%
Wer mono + Δ Δ et adaptations LDA MLLT	31,73%	56,44%	36,27%

FIGURE 3.7 – Résultats des décodages par mode sur le microphone SM58 proche

Les WER ont été réduis de moitié par rapport aux décodages précédents, sur le microphone SM58 éloigné. L'écart entre les différents modes est réduit également. Il passe de 14 points à un peu plus d'un point. Cette fois-ci, les meilleurs résultats sont pour la production en mode normal. Encore une fois, il y a une dégradation des résultats avec des modèles acoustiques monophones avec adaptation LDA et MLLT.

3.4 Paramétrage du système

Nous avons souhaité tester certains réglages qui, selon nous, avaient une chance d'améliorer les performances du système. Certains de ces réglages ont été pensés directement pour la reconnaissance de sons isolés. Nous avons également voulu tester les réglages mentionnés par [Picart et al., 2015] dans leur article sur la reconnaissance du *beatbox*.

Nous avons testé les réglages suivants :

1. Excepté le microphone encapsulé.

— la probabilité d'apparition d'un silence ;
— l'ajout d'une "PAUSE" avant et après chaque boxème dans le lexique ;
— le nombre de gaussiennes calculées lors de l'apprentissage des MA mono et mono+LDA et MLLT ;
— le nombre de MFCC ;
— le nombre d'états des HMM.

Pour les données d'apprentissage, nous avons tous les enregistrements en mode normal de tous les micros excepté le microphone encapsulé. Nous nous sommes concentrés sur le microphone SM58 proche pour le décodage et sur le mode de production normal.

3.4.1 Variation de la probabilité d'apparition d'un silence

Voici les résultats obtenus :

Taux	0,5 (par défaut)	0,6	0,7	0,8	0,9	0,99
WER mono	34,32%	28,23%	28,78%	26,94%	25,83%	26,94%
Nbre mots corrects	407/542	411/542	403/542	417/542	412/542	423/542
WER mono + Δ Δ et adaptations LDA MLLT	31,73%	72,14%	31,55%	48,71%	25,46%	41,51%
Nbre mots corrects	449/542	456/542	440/542	463/542	452/542	447/542

Figure 3.8 – Variation de la probabilité d'apparition d'un silence sous Kaldi

Nous pouvons observer que la probabilité d'apparition du silence qui donne le meilleur WER et le meilleur nombre de mots reconnus en même temps est celle à 0,99 pour les MA monos et à 0,9 pour les MA monos+LDA et MLLT. Le paramétrage par défaut de cette probabilité donne les moins bons résultats. Les adaptations LDA et MLLT n'améliorent pas toujours le WER mais viennent toujours augmenter le

nombre de mots reconnus.

3.4.2 Ajout d'une pause dans le lexique en contextes gauche et droit

Taux + pause	0,6	0,7	0,8	0,9	0,99
WER mono	17,16%	17,71%	16,97%	17,71%	19,93%
Nbre mots corrects	449/542	446/542	451/542	446/542	434/542
WER mono + Δ Δ et adaptations LDA MLLT	14,94%	14,76%	24,91%	18,08%	22,69%
Nbre mots corrects	465/542	469/542	464/542	452/542	466/542

FIGURE 3.9 – Décodages avec différentes probabilités d'apparition d'un silence et l'ajout d'un contexte pause

L'ajout d'une pause améliore les résultats de façon générale puisque nous étions en Figure 3.8 à un WER de 27,3% en moyenne sur les cinq derniers résultats et maintenant à un WER à 17,8%. Le décodage obtenant le meilleur WER et le meilleur nombre de mots reconnus est celui avec la probabilité d'apparition du silence à 0,8 pour les MA monos et à 0,7 pour les MA monos+LDA et MLLT. Encore une fois, les adaptations LDA et MLLT ne baissent pas toujours le WER mais viennent à chaque fois augmenter le nombre de mots reconnus.

3.4.3 Réduction du nombre de gaussiennes pour le calcul des monophones

Taux + pause gaussiennes à 400	0,6	0,7	0,8
WER mono	29,15%	30,26%	30,44%
Nbre mots corrects	384/542	380/542	378/542
WER mono + Δ Δ et adaptations LDA MLLT	18,08%	13,65%	15,68%
Nbre mots corrects	450/542	468/542	460/542

FIGURE 3.10 – Décodages avec différentes probabilités d'apparition d'un silence et l'ajout d'un contexte pause et la diminution du nombre de gaussiennes à 400

Les meilleurs résultats sont donnés avec la probabilité d'apparition du silence à 0,6 pour les MA monos et à 0,7 pour les MA monos+LDA et MLLT. En comparaison avec les résultats précédents, les WER calculés sur les décodages avec des monophones sont presque doublés. Les adaptations LDA et MLLT viennent en revanche améliorer le WER à chaque fois et augmentent le nombre de mots reconnus.

3.4.4 Augmentation du nombre de coefficients MFCC

Nous avons voulu tester l'augmentation du nombre de MFCC comme proposé par [Picart et al., 2015]. Nous avons défini le nombre de MFCC à 22, 22 étant le paramétrage optimal présenté dans l'article.

Taux + pause + gauss 400 + 22 mfcc	0,7
WER mono	22,88%
Nbre mots corrects	420/542
WER mono + Δ Δ et adaptations LDA MLLT	17,71%
Nbre mots corrects	449/542

FIGURE 3.11 – Décodages avec différentes probabilités d'apparition d'un silence, l'ajout d'un contexte pause, un nombre de gaussiennes à 400 et 22 MFCC

Taux + pause + gauss par défaut + 22 mfcc	0,7
WER mono	14,58%
Nbre mots corrects	465/542
WER mono + Δ Δ et adaptations LDA MLLT	17,53%
Nbre mots corrects	452/542

FIGURE 3.12 – Décodages avec différentes probabilités d'apparition d'un silence, l'ajout d'un contexte pause, un nombre de gaussiennes par défaut et 22 MFCC

Le premier test a été fait avec des MA monos calculés avec un nombre de gaussiennes à 400. Le second a été calculé avec le nombre de gaussiennes par défaut dans Kaldi : 1000.

Nous remarquons que l'augmentation du nombre de MFCC avec un nombre de

gaussiennes par défaut améliore le WER sur les monophones ainsi que le nombre de mots bien reconnus. En revanche, sur les monophones avec adaptations LDA et MLLT, les résultats sont moins significatifs.

Nous pouvons comparer les résultats de la Figure 3.12 avec les résultats du taux à 0,7 de la Figure 3.7 puisque seul le nombre de MFCC a évolué entre ces deux décodages. Nous remarquons que l'augmentation du nombre de MFCC améliore les résultats sur les MA monos mais pas sur les MA monos+LDA et MLLT.

Enfin, nous remarquons que la tendance s'inverse entre le décodage avec les MA monos et le décodage avec les MA monos+LDA et MLLT. L'augmentation du nombre de MFCC fait que les adaptations LDA et MLLT ne permettent plus de voir une amélioration du nombre de mots correctement reconnus.

3.4.5 Augmentation du nombre d'états HMM

Cette fois encore, ce réglage se base sur les résultats de [Picart et al., 2015]. L'augmentation du nombre d'états HMM leur avait donné de meilleurs résultats. Nous avons donc testé le développement d'un système avec des HMM à 5 états.

Taux + pause 5 HMM	0,7
WER mono	15,50
Nbre mots corrects	458/542
WER mono + Δ Δ et adaptations LDA MLLT	33,58
Nbre mots corrects	440/542

Figure 3.13 – Décodages avec différentes probabilités d'apparition d'un silence, l'ajout d'un contexte pause, et 5 états HMM

Les résultats de cette Figure peuvent être comparés aux résultats correspondant au taux à 0,7 de la Figure 3.7. Seul le nombre d'états HMM a varié entre les deux.

Nous remarquons que l'augmentation du nombre d'états HMM améliore les résultats sur les monophones, que ce soit pour le WER ou pour le nombre de mots reconnus. Sur les monophones avec adaptations LDA et MLLT en revanche, les résultats sont dégradés.

3.5 Sélection des meilleurs systèmes et analyse des substitutions

3.5.1 Avec des modèles acoustiques monophones

Si l'on ne regarde que les monophones, le meilleur modèle est le système entraîné avec une probabilité d'apparition du silence à 0,7, deux contextes pause dans le lexique et des HMM à cinq états. Vous trouverez dans la Figure qui suit un exemple des substitutions effectuées par le système. La liste complète se trouve en annexe C.3.3.

L'évaluation de 0 à 2 pour la ressemblance s'explique comme ceci :

— 0 pour deux sons extrêmement différents ;

— 1 pour deux sons acoustiquement ou avec une articulation quasi similaires ;

— 2 pour deux sons dont la seule différence est le voisement.

Cette évaluation est faite sur la seule base acoustique, par nous-mêmes. Nous obtenons pour ce système les substitutions en Figure 3.14.

	Référence	Hypothèse	Nbre	Ress.
1	ApPosDe_Exp	ApAl_Exp	7	1
2	ApDe_Ins_Fri_BiLa_Pro	PosDoVe_Lat_Exp	3	0
3	DoPa_Exp	ApPa_Cli_PlSu_Pro	3	0
4	PreDoPrePa_Cli	DoPa_Cli	3	1
5	BiLa_Exp_Rou	BiLa_Exp_Osc	2	1
6	BiLa_Fri	BiLa_Exp_LaPh_Rou_Nas	2	0
7	BiLa_Fri	BiLa_Exp_Voi_Nas	2	0
8	DoPa_Cli	ApPa_Cli_PlSu_Pro	2	1
9	DoPa_Exp	PosDoVe_Lat_Exp	2	1
10	GlPh_Exp_PreDoPrePa_Fri_Voi	ApPa_Exp_ApAl_Rou_Fri	2	0
11	PreDoPrePa_Cli	ApPa_Cli_PlSu_Pro	2	1
12	PreDoPrePa_Cli	ApPosDe_Exp	2	0
13	PreDoPrePa_Fri	ApAl_Fri	2	1
14	ApAl_Exp	ApLa_Exp	1	0
15	ApAl_Osc	BiLa_Osc_Voi	1	0
16	ApPa_Exp_DoPa_Fri	BiLa_Exp_Nas_OrPh_Fri	1	0
17	ApPa_Exp_DoPa_Fri	PosDoVe_Lat_Exp	1	0
18	ApPa_Exp_PosDoVe_Fri	ApPosDe_Rou_LaDe_Fri	1	0
19	ApPa_Ins_NonPul_Rou	ApLa_Exp	1	0
20	ApPosDe_Rou_LaDe_Fri	ApPa_Exp_ApAl_Rou_Fri	1	0

Figure 3.14 – Exemple de substitutions pour le meilleur système avec des MA monos

Nous pouvons observer que les deux premiers sons confondus en ligne 1 sont des sons qui se ressemblent acoustiquement. Si nous devions les décrire phonétiquement, nous pourrions dire que le son ApPosDe_Exp donne [Ts] avec le son [s] très court, et le son ApAl_Exp donne [T]. Sept sons sur vingt sont assez ressemblants acoustiquement parlant ou dans leur production.

3.5.2 Avec des modèles acoustiques monophones + adaptations LDA et MLLT

Le meilleur modèle avec des monophones avec adaptations LDA et MLLT, est celui avec une probabilité d'apparition du silence à 0,7 et deux contextes pause dans le lexique. Vous trouverez en Figure 3.15 un exemple des substitutions effectuées par le système. La liste complète se trouve en annexe C.3.3.

	Référence	Hypothèse	Nbre	Ress.
1	ApDe_Ins_Fri_BiLa_Pro	ApPa_Exp	3	0
2	BiLa_Osc	BiLa_Osc_Voi	3	2
3	ApPosDe_Exp	ApAl_Exp	2	1
4	DoPa_Exp	ApPa_Cli_PlSu_Pro	2	0
5	LaDe_Exp	ApLa_Exp	2	0
6	PoDoUv_Rou	PoDoUv_Ins_Rou_Voi	2	2
7	ApDe_Ins_Fri_BiLa_Pro	LaDe_Exp	1	0
8	ApLa_Ins_Exp_BiLa_Pro_GlPh_Exp	BiLa_Exp_Voi_Nas	1	1
9	ApPa_Exp_PosDoVe_Fri	PosDoVe_Fri	1	0
10	ApPosDe_Rou_LaDe_Fri	BiLa_Exp_LaDe_Fri	1	0
11	BiLa_Fri	BiLa_Osc	1	0
12	BiLa_Fri_Sif	ApPa_Cli_PlSu_Pro	1	0
13	BiLa_Lat_Ins_Rou	BiLa_Osc_Voi	1	0
14	BiLa_Rou	BiLa_Exp_Osc	1	1
15	DoPa_Cli	DoPa_Exp	1	0
16	DoPa_Exp	ApAl_Exp	1	0
17	GlPh_Exp_PreDoPrePa_Fri_Voi	ApPa_Exp_ApAl_Fri	1	0
18	GlPh_Ins_Exp_NonPul	GlPh_Exp	1	1
19	LaDe_Exp	BiLa_Exp_Voi	1	1
20	PosDoVe_Ins_Fri_ApPosDe_Pro	ApAl_Exp	1	0

Figure 3.15 – Exemple de substitutions pour le meilleur système avec des MA mono+LDA et MLLT

Les sons confondus se ressemblent sept fois sur vingt, dont une fois où la seule différence entre les sons est le fait qu'il soit voisé ou non voisé.

Chapitre 4

Discussion

Pour commencer, il convient de noter que les résultats précédents prouvent la faisabilité d'un système de reconnaissance automatique du *beatbox*, que l'on utilise un petit ou un grand vocabulaire.

Les tests suivants ont été effectués avec des modèles de langue constitués de zérogrammes. La probabilité d'apparition est la même pour tous les mots ce qui fait que le modèle de langue n'a pas de poids dans les décodages effectués. Nous n'avons pas calculé de modèle de langue avec des bigrammes et trigrammes car nous ne disposons pas de données transcrites autres que nos deux corpus. Nous aurions donc été en sur-apprentissage si nous avions appris des statistiques sur nos données. De plus, pour que les données statistiques soient intéressantes et utilisables, il faut un corpus textuel assez conséquent. Enfin, comme nous l'avons dit précédemment, il aurait été impossible de demander à nos deux sujets de transcrire avec leur grammaire *Vocal Grammatics* des productions qu'ils n'ont pas effectuées car leur écriture se concentre sur l'articulation. Il est quasiment impossible de savoir avec certitude comment un son a été produit par quelqu'un d'autre.

Notre prototype étant voué à être utilisé par des débutants, il pourrait être intéressant par la suite de construire un modèle de langue afin d'augmenter la probabilité de reconnaissance des sons les plus faciles. Il nous sera indispensable d'en avoir un le jour où nous nous concentrerons sur les séquences. À ce jour, les tests effectués sur le corpus grand vocabulaire se concentrent uniquement sur des sons unitaires.

4.1 Reconnaissance sur petit vocabulaire

Les premiers résultats sur le corpus petit vocabulaire nous ont permis d'avoir un WER de 21,79%. Celui-ci a été légèrement faussé par le chant diphonique et la trompette qui sont deux sons particuliers par rapport aux autres sons plus percussifs. L'annotation de ces deux sons n'était pas bonne car nous ne savions pas si nous devions annoter comme un boxème tout ce qui se trouve entre deux pauses longues ou annoter un boxème à chaque changement de note. Il est probable que l'annotation d'un son pour chaque note soit plus efficace sachant que chaque note imitée est due à une propulsion d'air. Le chant diphonique, quant à lui, joue avec les fréquences. Sachant que les paramètres MFCC se basent sur l'énergie en fréquences, nous pouvons émettre l'hypothèse que le système se perd à cause de cela.

Avec un taux de mots bien reconnus à 82,5% et un WER à 21,79% nous avons la preuve de l'efficacité d'un système sur un petit vocabulaire. Le WER étant le même pour les monophones que pour les triphones, nous émettons quelques doutes sur la nécessité de calculer des modèles acoustiques triphones.

La cymbale [ts] et le classic snare drum [K] sont souvent confondus avec d'autres sons. Ils n'étaient présents dans l'apprentissage du système que dans les séquences. Nous n'avions pas d'enregistrement répétés individuels de ces sons. Les séquences n'étant qu'au nombre de six dans le corpus, nous pensons que le fait qu'ils soient confondus peut être dû à un manque de données rendant difficile pour le système d'apprendre des sons en contexte. Ils sont tout autant confondus avec des modèles acoustiques monophones que triphones.

4.2 Impact du type de microphone

Nous pouvons confirmer notre hypothèse précédente sur l'utilisation de modèles acoustiques triphones pour le décodage de sons unitaires. En effet, les triphones apportent soit une amélioration moindre, soit une dégradation conséquente, comme on peut le voir avec le microphone SM58 proche avec les modèles acoustiques tri2b. Ceci peut s'expliquer par le fait que nous ne travaillons pour le moment que sur des sons isolés et que les seuls contextes gauche et droit qui existent sont du silence. Il n'y

a donc pas d'intérêt réel à calculer des triphones. Concernant les adaptations LDA et MLLT, nous décidons pour la suite de les garder et de les calculer directement sur des monophones. Pour la suite de cette analyse, nous ne prendrons en compte que les monophones.

Nous pouvons observer que le *beatboxeur* professionnel et le *beatboxeur* amateur n'ont pas un WER significativement différent. Ceci peut peut-être s'expliquer par le fait qu'Adrien, le *beatboxeur* amateur, a eu l'intelligence de ne pas enregistrer des sons qu'il ne maîtrise pas afin de ne pas fausser les résultats en ayant des productions trop différentes.

Le microphone donnant les pires résultats est le microphone encapsulé. Les productions encapsulées sont très différentes acoustiquement des autres productions, dû au fait que le son est très renforcé dans les basses fréquences. Nous décidons pour la suite de mettre ce microphone de côté et de nous concentrer sur les autres microphones avec des enregistrements simultanés et plus proches acoustiquement.

Le microphone avec les meilleurs résultats est le microphone SM58 éloigné. Les résultats des tests avec le microphone SM58 proche, le microphone cravate et le microphone d'ambiance sont proches de ceux du SM58 éloigné. Nous pouvons en conclure que le type de microphone n'a pas d'effet sur la reconnaissance. Néanmoins, les résultats n'étant pas très bons (>50% WER), nous pensons que les différents modes de production (normal, débutant, avec variations) peuvent avoir un effet négatif sur l'apprentissage lorsqu'ils sont mélangés.

4.3 Influence de la variabilité dans la production

Nous pouvons tout d'abord observer que la division de chaque mode de production pour un microphone ne donne pas des résultats satisfaisants. Les résultats étaient meilleurs lorsque tous les modes de production étaient mélangés dans l'apprentissage (Figure 3.3). Néanmoins, le nombre de répétitions par enregistrement pour l'apprentissage a été considérablement réduit ici, puisqu'il n'était que de deux, contre six précédemment pour cinq fois plus d'enregistrements. Lorsque nous mélangeons toutes les productions normales, en mode débutant ou avec variations de

tous les microphones, nous pouvons voir (Figure 3.7) que les WER sont réduits de moitié. Ceci confirme notre hypothèse selon laquelle le nombre d'échantillons de son n'était pas assez conséquent dans le premier test (Figure 3.6). Les tests ont certes été faits sur un même microphone placé différemment, éloigné pour le premier et proche pour le deuxième, mais rappelons que la différence de WER en Figure 3.3 entre ces deux enregistrements n'était pas significative. Les résultats de nos deux tests sont donc comparables. Nous avons travaillé sur le SM58 éloigné pour les premiers tests car c'était le microphone avec les meilleurs résultats. Pour les deuxièmes tests, nous avons travaillé avec le SM58 proche car lors des représentations artistiques, le public sera invité à se placer assez près du microphone.

En ajoutant les échantillons de chaque microphone aux données d'apprentissage lors de la séparation des modes de production (Figure 3.7), nous obtenons de meilleurs résultats que lors des décodages où tous les types de production étaient mélangés (Figure 3.3). Ceci s'observe, bien que le nombre de répétitions de boxèmes pour les corpus d'apprentissage soit inférieur pour ces décodages que lors des décodages par microphone. Nous pouvons donc en conclure que le mélange des différents modes de production dégrade le système. Nous sommes optimistes quant à l'idée d'avoir un système encore plus efficace si nous augmentons encore le nombre de données d'apprentissage.

Les deuxièmes tests en Figure 3.7 nous montrent également qu'aucun mode de production n'est vraiment meilleur qu'un autre lorsqu'ils sont séparés. Le WER du mode débutant est plus bas, donc meilleur, mais l'apprentissage et les décodages sur ce mode n'ont pu être effectués que sur le *beatboxeur* professionnel étant donné que le *beatboxeur* amateur n'a aucune production de ce genre. C'est ce qui peut expliquer que les résultats semblent meilleurs. Nous émettons l'hypothèse que les trois types de sons étaient trop différents pour être mélangés et empêchait les différentes classes de boxèmes d'être correctement discriminées.

Enfin, nous remarquons que les adaptations LDA et MLLT ne sont pas constantes et n'apportent pas toujours une amélioration du WER. Nous pensons que cela peut provenir de la quantité de données qui est trop faible.

4.4 Paramétrage du système

Nous nous sommes intéressés sur cette partie au nombre de mots bien reconnus par le système. Celui-ci nous donne une évaluation qui n'est pas dépendante du nombre de substitutions, d'insertions et de délétions.

Augmenter la probabilité d'apparition d'un silence donne de meilleurs résultats que ce qui est proposé par défaut. Ceci s'explique par le fait que notre corpus est composé de nombreuses pauses. L'ajout en plus d'un contexte pause avant et après chaque mot dans le lexique, améliore encore les performances. Nous avons pu remarquer sur les résultats en Figure 3.8 et en Figure 3.9 que les adaptations LDA et MLLT, qui jusque là nous semblaient peu efficaces en regardant le WER, améliorent finalement le nombre de mots reconnus par le système à chaque fois .

Nous pensions que la réduction du nombre de gaussiennes améliorerait nos résultats étant donné que nous n'avons pas beaucoup de données. Finalement, cela n'est pas le cas. Nous avons observé l'effet inverse. Nous ne continuerons donc pas avec ce réglage.

L'augmentation du nombre de MFCC est conseillée par [Picart et al., 2015]. Nous avons vu nos résultats s'améliorer sur les monophones mais se dégrader sur les monophones avec adaptations LDA et MLLT. L'augmentation des MFCC nous a fait perdre le bénéfice des adaptations LDA et MLLT sur le nombre de mots reconnus.

Enfin, nous avons augmenté le nombre d'états HMM en le passant de trois à cinq. Nous supposions que cette augmentation pourrait bénéficier aux sons complexes composés de plusieurs boxèmes en permettant d'analyser plus finement l'évolution temporelle du signal. Ce changement du nombre d'états apporte une légère amélioration du WER et a permis de reconnaître douze mots de plus qu'en Figure 3.9 [1]. Néanmoins, il nous fait également perdre le bénéfice des adaptations LDA et MLLT. Nous pensons que combiner des HMM à trois états et des HMM à cinq états dans un système pourrait être efficace. Les HMM à trois états serviraient pour les sons composés d'un seul boxèmes et les HMM à cinq états serviraient à reconnaître les sons composés de plusieurs boxèmes.

1. Probabilité d'apparition du silence à 0,7 et pauses ajoutées dans le lexique

4.5 Sélection du meilleur système et analyse des substitutions

L'évaluation effectuée ici a été faite par nos soins et se base uniquement sur notre écoute. Nous souhaitions voir si le système confondait toujours des sons extrêmement différents comme pour le tout premier décodage sur le corpus petit vocabulaire.

4.5.1 Avec des modèles acoustiques monophones

Ce système confond des sons qui se ressemblent et des sons qui se ne ressemblent pas. Nous pouvons observer une régularité pour les deux premiers sons qui sont confondus sept fois sur neuf. Néanmoins, même si cela n'apparaît pas dans notre notation, certains sons sont confondus et, bien qu'ils ne se ressemblent pas totalement acoustiquement, ils peuvent avoir des caractéristiques communes comme pour les sons ApAl_Exp [t] et ApLa_Exp [p] qui sont tous les deux des sons brefs et les sons PreDoPrePa_Fri [s]et ApAl_Fri [ʃ] qui sont tous les deux des sons fricatifs.

4.5.2 Avec des modèles acoustiques monophones + adaptations LDA et MLLT

Tout comme le système entraîné avec cinq états HMM sur des modèles acoustiques monophones, il semble y avoir ici une tendance à confondre plus de sons totalement différents que de sons qui se ressemblent.

Nous observons d'une façon générale sur les substitutions de ces derniers tests que les sons confondus peuvent être des sons qui se ressemblent de part leur articulation ou acoustiquement, tout comme des sons très différents. Notons tout de même que l'évaluation de la ressemblance ayant été faite par les soins d'une seule personne, il serait intéressant de faire une vraie campagne d'évaluation composée d'évaluateurs *beatboxeurs* et non *beatboxeurs*. Nous pourrions la baser sur deux critères, la ressemblance acoustique et la ressemblance de la production. Celle-ci permettrait de voir si les humains obtiennent les mêmes résultats que les systèmes que nous avons développé et essayer d'analyser ce qui trompe le système.

Chapitre 5

Conclusion et perspectives

5.1 Conclusion

Dans ce mémoire, nous avons montré qu'il était possible de développer un système de reconnaissance automatique du *beatbox* à partir d'une boîte à outils faite pour de la reconnaissance automatique de la parole. Nous avons pu développer des systèmes fonctionnant sur des corpus petit vocabulaire et grand vocabulaire.

Les meilleurs résultats observés ont été obtenus avec des modèles acoustiques monophones et des modèles acoustiques monophones avec adaptations LDA et MLLT. Les meilleurs résultats avec des monophones ont été obtenus grâce à des HMM à cinq états, l'augmentation de la probabilité d'apparition d'un silence à 0,7 et l'ajout d'un contexte pause gauche et droit dans le lexique. Ces tests ont été effectués sur un mode de production normal, avec toutes les productions normales de tous les microphones [1] pour l'apprentissage, et le microphone SM58 positionné à 10cm de la bouche du *beatboxeur* pour les tests. Le WER était égal à 15,50% et le nombre de mots correctement reconnu à 458/542. Les meilleurs résultats avec des monophones + adaptations LDA et MLLT ont été obtenus avec une probabilité d'apparition d'un silence à 0,7 et l'ajout d'un contexte pause gauche et droit dans le lexique. Nous avons obtenu un WER de 14,76% et un nombre de mots correctement reconnus à 469/542. Ceci a pu être produit avec un modèle de langue constitué de zérogrammes.

Un des grands principes de la reconnaissance vocale a pu être observé ici : l'aug-

1. excepté le microphone encapsulé

mentation des données d'apprentissage améliore significativement les systèmes. Nous n'avons pas observé de différence significante entre les deux beatboxeurs et pourrons donc à l'avenir ouvrir l'agrandissement de notre base de données à des *beatboxeurs* amateurs.

Nous avons pu remarquer qu'il n'y a pas d'influence du type de microphone (cravate, éloigné, proche) sur le système. Néanmoins, les enregistrements ont été faits dans un studio d'enregistrement silencieux. Nous pouvons imaginer qu'un microphone d'ambiance tel que le DPA 4006 capterait beaucoup trop de sons inutiles dans une salle remplie de personnes qui parlent, toussent, bougent *etc.* Le microphone encapsulé semble se détacher négativement des autres. Cela n'est pas dû au microphone en lui-même mais plutôt à l'utilisation qui en est faite.

Les différents modes de production des sons ont eu une influence sur nos systèmes. Nos enregistrements en mode débutant ayant été réalisés par un beatboxeur professionnel essayant d'imiter des débutants, ceux-ci ont été faussés. En séparant les différents modes de production nous avons pu obtenir des WER 1,5 fois meilleurs par rapport à un système où tous les modes de production sont mélangés dans la phase d'apprentissage.

Concernant les sons substitués, nous n'avons pas encore d'explication sur le fait que certains sons substitués sont parfois ressemblants, parfois très différents. Nous pensons qu'une classification des sons en fonction de la durée et du nombre de boxèmes compris à l'intérieur de ces sons pourrait nous permettre de définir des HMM à différents états pour chaque type de son et d'améliorer le système.

5.2 Perspectives

Nous souhaiterions tester le système sur des données de locuteurs inconnus afin de voir si nous ne sommes pas en sur-apprentissage. Le système étant destiné à être testé par un public tout venant, nous souhaiterions tout particulièrement tester son efficacité sur des voix de femmes ou d'enfants, ainsi que sur des productions de débutants. Ceci demandera l'agrandissement de la base de données afin d'y inclure des voix de femmes et des voix d'enfants, ainsi que des productions de vrais débutants.

De même, nous souhaiterions avoir un système capable de décoder efficacement des sons encapsulés car c'est un type de production très utilisé par les professionnels. L'accent n'a pas été mis dessus lors de ce travail car ce projet et son aménagement n'ont pas été pensés pour de la production de sons encapsulés.

Nous pensons envisageable la transcription de sons complexes tel que le *808 snare roll* en unités phonétiques plus petites. En effet, un son complexe est composé de plusieurs boxèmes qui semblent avoir une réalisation temporelle et non simultanée, sur une très courte durée. Ceci nous permettrait de réduire le vocabulaire tout comme cela a été fait pour les langues et leur transcription en phonétique.

Afin d'améliorer notre système, nous souhaiterions élaborer un modèle de langue en récoltant un grand corpus de séquences de *beatbox* transcrits en écriture *Vocal Grammatics*. Ceci nous permettrait de pouvoir utiliser des modèles acoustiques triphones pour le décodage de séquence et donc de prendre en compte la coarticulation des sons. Enfin, il serait intéressant de tester notre système avec des paramètres acoustiques autres que les MFCC. Pour cela, nous nous baserons notamment sur les paramètres acoustiques utilisés par [Kapur et al., 2004] et [Sinyor et al., 2005] pour la classification de sons de *beatbox*.

Bibliographie

[Adda-Decker and Lamel, 2000] Adda-Decker, M. and Lamel, L. (2000). The Use of Lexica in Automatic Speech Recognition. In *Lexicon Development for Speech and Language Processing*, pages 235–266. Springer Netherlands, Dordrecht.

[Bezdel and Bridle, 1969] Bezdel, W. and Bridle, J. S. (1969). Speech recognition using zero-crossing measurements and sequence information. *Proceedings of the Institution of Electrical Engineers*, 116(4) :617–623.

[Dao et al., 2016] Dao, V.-L., Nguyen, V.-D., Nguyen, H.-D., and Hoang, V.-P. (2016). Hardware Implementation of MFCC Feature Extraction for Speech Recognition on FPGA. pages 248–254, Thai Nguyen city, Vietnam.

[Dave, 2013] Dave, N. (2013). Feature extraction methods LPC, PLP and MFCC in speech recognition. *International Journal For Advance Research in Engineering And Technology(ISSN 2320-6802)*, Volume 1(6).

[Elenius et al., 2004] Elenius, D., Blomberg, M., Stockholms universitet, and Humanistiska fakulteten (2004). Comparing speech recognition for adults and children. pages 156–159, Dept. of linguistics, Stockholm University. OCLC : 1026904002.

[Haeb-Umbach and Ney, 1992] Haeb-Umbach, R. and Ney, H. (1992). Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *[Proceedings] ICASSP-92 : 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 13–16 vol.1, San Francisco, CA, USA. IEEE.

[Hess, 2007] Hess, M. (2007). *Icons of hip hop : an encyclopedia of the movement, music, and culture*. Greenwood Press, Westport, Conn. OCLC : 85766109.

[Hipke et al., 2014] Hipke, K., Toomim, M., Fiebrink, R., and Fogarty, J. (2014). BeatBox : End-user Interactive Definition and Training of Recognizers for Percussive Vocalizations. pages 121–124, Como, Italy. ACM.

[Juang and Rabiner, 2004] Juang, B. H. and Rabiner, L. R. (2004). Automatic Speech Recognition – A Brief History of the Technology Development. page 24.

[Jurafsky and Martin, 2018] Jurafsky, D. and Martin, H. J. (2018). N-gram Language Models. In *Speech and Language Processing - Draft chapters*, page 28.

[Kapur et al., 2004] Kapur, A., Tzanetakis, G., and Benning, M. (2004). Query-by-Beat-Boxing : Music Retrieval For The DJ. Barcelona, Spain.

[Klapuri, 2019] Klapuri, A. (2019). Music Information Retrieval. page 5.

[Le Blouch, 2009] Le Blouch, O. (2009). *Décodage acoustico-phonétique et applications à l'indexation audio automatique*. thesis, Toulouse 3, Toulouse, France.

[Le Grand, 2012] Le Grand, J. (2012). *Amélioration des systèmes de reconnaissance de la parole des personnes âgées*. mémoire de master, Université Grenoble-Alpes, Grenoble, France.

[Lederer, 2005] Lederer, K. (2005). *The Phonetics of Beatboxing*. mémoire de master, Leeds university, Leeds, UK.

[Liu, 2015] Liu, L. (2015). *Acoustic Models for Speech Recognition Using Deep Neural Networks Based on Approximate Math*. mémoire de master, Massachusetts Institude of Technology, Massachusetts, USA.

[Lyons, 2019] Lyons, J. (2019). *Mel Frequency Cepstral Coefficient (MFCC) tutorial*. Page web, http ://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/, Consultée en dernier le : 2019-06-17 20 :29 :56.

[Nakamura et al., 2005] Nakamura, M., Iwano, K., and Furui, S. (2005). Analysis of Spectral Space Reduction in Spontaneous Speech and its Effects on Speech Recognition Performances. page 4.

[Nakamura et al., 2008] Nakamura, M., Koji, I., and Furui, S. (2008). Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech & Language*, 22 :171–184.

[Narang and Gupta, 2015] Narang, S. and Gupta, D. (2015). Speech Feature Extraction Techniques : A Review. *International Journal of Computer Science and Mobile Computing*, 4(3) :107–114.

[Nasereddin and Omari, 2017] Nasereddin, H. H. O. and Omari, A. (2017). Classification Techniques for Automatic Speech Recognition (ASR) Algorithms used with Real Time Speech Translation. London, UK.

[Paroni, 2014] Paroni, A. (2014). *Paroni Annalisa How do beatboxers play.* mémoire de master, Universita degli Studi di Padova, Padova.

[Paroni, 2016] Paroni, A. (2016). *Production de sons plosifs : Comparaison du beatbox et de la parole.* mémoire de master, Université Grenoble-Alpes, Grenoble, France.

[Pellegrini and Duée, 2003] Pellegrini, T. and Duée, R. (2003). Suivi de Voix Parlée grâce aux Modèles de Markov Cachés. Rapport de stage DEA ATIAM, IRCAM, Paris, France.

[Picart et al., 2015] Picart, B., Brognaux, S., and Dupont, S. (2015). Analysis and automatic recognition of Human BeatBox sounds : A comparative study. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4255–4259, Brisbane, QLD, Australia.

[Picone, 1993] Picone, J. (1993). Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, 81(9) :1215–1247.

[Povey et al., 2011] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlıcek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. page 4, Hawaii, US.

[Proctor et al., 2013] Proctor, M., Bresch, E., Byrd, D., Nayak, K., and Narayanan, S. (2013). Paralinguistic mechanisms of production in human "beatboxing" : A real-time magnetic resonance imaging study. *The Journal of the Acoustical Society of America*, 133(2) :1043–1054.

[Rath et al., 2013] Rath, S. P., Povey, D., Vesely, K., and Cernocky, J. (2013). Improved Feature Processing for Deep Neural Networks. page 5.

[Russell and D'Arcy, 2007] Russell, M. and D'Arcy, S. (2007). Challenges for computer recognition of children's speech. page 4, Farmington, PA, USA.

[Sapthavee et al., 2014] Sapthavee, A., Yi, P., and Sims, H. S. (2014). Functional Endoscopic Analysis of Beatbox Performers. *Journal of voice : official journal of the Voice Foundation*, 28(3) :328–331.

[Silverman et al., 1992] Silverman, K., Blaauw, E., Spitz, J., and Pitrelli, J. (1992). A prosodic comparison of spontaneous speech and read speech. Alberta, Canada.

[Sinyor et al., 2005] Sinyor, E., McKay, C., Fiebrink, R., McEnnis, D., and Fujinaga, I. (2005). Beatbox classification using ACE. page 4, London, UK.

[Stowell and Plumbley, 2008] Stowell, D. and Plumbley, M. (2008). Characteristics of the beatboxing vocal style. Technical report C4DM-TR-08-01, Centre for Digital Music Department of Electronic Engineering, Queen Mary, university of London, UK.

[Tyte and Splinter, 2014] Tyte and Splinter (2014). *Standard Beatbox Notation (SBN)*. Page web, Consultée en dernier le : 2019-06-15 20 :49 :12, https ://www.humanbeatbox.com/articles/standard-beatbox-notation-sbn/.

[Tyte, 2019] Tyte, G. (2019). *Beatboxology – Introduction*. Page web, https ://bzzktt.com/beatboxology-introduction/, Consultée en dernier le : 15/06/2019 à 21 :00.

[Wijoyo, 2011] Wijoyo, S. (2011). Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot. page 5, Bangkok, Thailand.

[Yu and Deng, 2015] Yu, D. and Deng, L. (2015). Deep Neural Network-Hidden Markov Model Hybrid Systems. In Yu, D. and Deng, L., editors, *Automatic Speech Recognition : A Deep Learning Approach*, Signals and Communication Technology, pages 99–116. Springer London, London.

[Zitouni, 2000] Zitouni, I. (2000). *Modélisation du langage pour les systèmes de reconnaissance de la parole destinés aux grands vocabulaires : application à MAUD*. PhD thesis, Université Henri Poincaré - Nancy 1, Nancy, France.

Annexes

Annexe A

Sons enregistrés pour chaque *beatboxeur* sur le corpus petit vocabulaire

Son	Écriture phonétique	A.	A.
Sons unitaires			
snare-drum	T	x	x
snare-roll	Tw	x	x
bongo-drum	ˆKng		x
brushed-cymbal	th	x	x
chant diphonique	diph		x
classic-bass	P°	x	x
classic-kick-hardware	P	x	x
classic-kick-humming	Pv	x	x
classic-snare-drum-hardware	Pf	x	x
classic-snare-drum-humming	Pfv	x	x
click	!	x	x
combination snare	ˆKCLh		x
dry-snare-hardware	T !	x	x
dry-snare-humming	T !v	x	x
inward-clap	ˆCL	x	x
inward-k-snare	ˆK		x
reversed-classic-kick	ˆP	x	x
reversed-snare-hat	ˆT	x	x
trumpet	trump		x
Séquences			
PuTiKa	P ts K	x	x
PuTiKa enchaîné	P ts K ts	x	x
Boots & cats enchaîné	B ts K ts	x	x

Figure A.1 – Sons enregistrés pour chaque *beatboxeur* sur le corpus petit vocabulaire

Annexe B

Dictionnaire *Vocal Grammatics*

B.1 Dictionnaire *Vocal Grammatics*

B.2 Tableau des abréviations

NOM BEATBOX	FORMULE PHONETIQUE	FORMULE PHONETIQUE ABREGEE
Classic kick	Bi-labial_Explosif	BiLa_Exp
Vocal kick	Bi-labial_Explosif_Voisé	BiLa_Exp_Voi
Roll kick	Bi-labial_Explosif_Roulé	BiLa_Exp_Rou
Oscilled kick	Bi-labial_Explosif_Oscillé	BiLa_Exp_Osc
Breath kick	Bi-labial_Explosif_Nasalisé	BiLa_Exp_Nas
Vocal nose kick	Bi-labial_Explosif_Voisé_Nasalisé	BiLa_Exp_Voi_Nas
Classic snare drum	Bi-labial_Explosif_Labio-dental_Fricatif	BiLa_Exp_LaDe_Fri
Snake snare	Bi-labial_Explosif_Apico-dental_Fricatif	BiLa_Exp_ApDe_Fri
Symbal kick	Bi-labial_Explosif_Prédorso-prépalatal_Fricatif	BiLa_Exp_PreDoPrePa_Fri
Techno snare	Bi-labial_Explosif_Apico-alvéolaire_Fricatif	BiLa_Exp_ApAl_Fri
Bass nose kick	Bi-labial_Explosif_Laryngo-pharyngal_Roulé_Nasalisé	BiLa_Exp_LaPh_Rou_Nas
Kick lip roll	Bi-labial_Latéralisé_Inspiré _Explosif_Roulé	BiLa_Lat_Ins_Exp_Rou
Kick vocal lip roll	Bi-labial_Latéralisé_Inspiré _Explosif_Roulé_Voisé	BiLa_Lat_Ins_Exp_Rou_Voi
None lip roll	Bi-labial_Latéralisé_Inspiré_Non-pulmonaire_Roulé	BiLa_Lat_Ins_NonPul_Rou
None kick lip roll	Bi-labial_Latéralisé_Inspiré_Explosif _Non-pulmonaire_Roulé	BiLa_Lat_Ins_Exp_NonPul_Rou
Lips clic	Bi-labial_Clic	BiLa_Cli
Blow	Bi-labial_Fricatif	BiLa_Fri
Wistle	Bi-labial_Fricatif_Sifflé	BiLa_Fri_Sif
Spit snare	Bi-labial_Constrictif	BiLa_Con

NOM BEATBOX	FORMULE PHONETIQUE	FORMULE PHONETIQUE ABREGEE
Zipper	Bi-labial_Inspiré_Constrictif_Sifflé	BiLa_Ins_Con_Sif
None zipper	Bi-labial_Inspiré_Non-pulmonaire_Constrictif_Sifflé	BiLa_Ins_NonPul_Con_Sif
Lip vibration	Bi-labial_Roulé	BiLa_Rou
lip roll	Bi-labial_Latéralisé_Inspiré_Roulé	BiLa_Lat_Ins_Rou
Vocal lip roll	Bi-labial_Latéralisé_Inspiré_Roulé_Voisé	BiLa_Lat_Ins_Rou_Voi
Lips oscillation	Bi-labial_Oscillé	BiLa_Osc
Vocal lips oscillation	Bi-labial_Oscillé_Voisé	BiLa_Osc_Voi
Lip perc	Labio-dental_Explosif	LaDe_Exp
Flute	Labio-dental_Inspiré_Fricatif_Sifflé	LaDe_Ins_Fri_Sif
Fric	Dento-labial_Fricatif	DeLa_Fri
Teeth tap	Bi-dental_Projeté	BiDe_Pro
Tong Kick	Apico-labial_Explosif	ApLa_Exp
Tutu	Apico-postlabial_Constrictif	ApPosLa_Con
Reversed snake sanre	Apico-dental_Inspiré_Fricatif_Bi-labial_Projeté	ApDe_Ins_Fri_BiLa_Pro
Hi-Hat close	Apico-postdental_Explosif	ApPosDe_Exp
Hi-Hat open	Apico-postdental_Explosif_Prédorso-prépalatal_Fricatif	ApPosDe_Exp_PreDoPrePa_Fri
Brush symbal	Apico-postdental_Explosif_Labio-dental_Fricatif	ApPosDe_Exp_LaDe_Fri
Closing Hi-Hat	Apico-postdental_Explosif_Prédorso-prépalatal_Fricatif_Apico-postdental_Projeté	ApPosDe_Exp_PreDoPrePa_Fri_ApPosDe_Pro

NOM BEATBOX	FORMULE PHONETIQUE	FORMULE PHONETIQUE ABREGEE
Speed charley chain	Apico-postdental_Roulé_Labio-dental_Fricatif	ApPosDe_Rou_LaDe_Fri
Dry Kick	Apico-alvéolaire_Explosif	ApAl_Exp
Sic	Prédorso-prépalatal_Fricatif	PreDoPrePa_Fri
Tong oscillation	Apico-alvéolaire_Oscillé	ApAl_Osc
Vocal tong oscillation	Apico-alvéolaire_Oscillé_Voisé	ApAl_Osc_Voi
Vocal cowbell	Apico-palatal_Explosif	ApPa_Exp
808 snare	Apico-palatal_Explosif_Apico-alvéolaire_Fricatif	ApPa_Exp_ApAl_Fri
808 snare roll	Apico-palatal_Explosif_Apico-alvéolaire_Roulé_Fricatif	ApPa_Exp_ApAl_Rou_Fri
Middle 808 snare	Apico-palatal_Explosif_Dorso-palatal_Fricatif	ApPa_Exp_DoPa_Fri
Back 808 snare	Apico-palatal_Explosif_Postdorso-vélaire_Fricatif	ApPa_Exp_PosDoVe_Fri
Clic clap	Apico-palatal_Clic_Plancho-Sublingual_Projeté	ApPa_Cli_PlSu_Pro
Helicopter clic roll	Apico-palatal_Inspiré_Roulé_Bi-labial_Constrictif	ApPa_Ins_Rou_BiLa_Con
Clic roll	Apico-palatal_Inspiré_Roulé	ApPa_Ins_Rou
None clic roll	Apico-palatal_Inspiré_Non-pulmonaire_Roulé	ApPa_Ins_NonPul_Rou
Hollow clap	Plancho-Sublingual_Projeté	PlSu_Pro
Hollow clap wistle	Plancho-Sublingual_Projeté_Oro-pharyngal_Fricatif_Bi-labial_Fricatif_Sifflé	PlSu_Pro_OrPh_Fri_BiLa_Fri_Sif

NOM BEATBOX	FORMULE PHONETIQUE	FORMULE PHONETIQUE ABREGEE
Hollow clap wistle non pulmonaire	Plancho-Sublingual_Projeté_Bilabial_Non pulmonaire_Fricatif_Sifflé	PlSu_Pro_BiLa_NonPul_Fri_Sif
Clic	Prédorso-prépalatal_Clic	PreDoPrePa_Cli
Chic	Apico-alvéolaire_Fricatif	ApAl_Fri
The cat	Postdorso-vélaire_Fricatif	PosDoVe_Fri
K perc	Dorso-palatal_Explosif	DoPa_Exp
K perc snare	Dorso-palatal_Explosif_Fricatif	DoPa_Exp_Fri
Middle Clic	Dorso-palatal_Clic	DoPa_Cli
Dog Call	Dorso-palatal_Latéralisé_Clic	DoPa_Lat_Cli
K rimshot	Postdorso-vélaire_Latéralisé_Explosif	PosDoVe_Lat_Exp
K Snare	Postdorso-vélaire_Latéralisé_Explosif _Fricatif	PosDoVe_Lat_Exp_Fri
Inward K Snare	Postdorso-vélaire_Latéralisé_Inspiré _Explosif_Fricatif	PosDoVe_Lat_Ins_Exp_Fri
Inward K rimshoh	Postdorso-vélaire_Latéralisé_Inspiré _Clic_Fricatif_Non-pulmonaire	PosDoVe_Lat_Ins_Cli_Fri_NonPul
Sucker punch	Postdorso-vélaire_Inspiré_Fricatif_Apico-postdental_Projeté	PosDoVe_Ins_Fri_ApPosDe_Pro
Snore Bass	Postdorso-uvulaire_Roulé	PoDoUv_Rou
Vocal Snore	Postdorso-uvulaire_Roulé_Voisé	PoDoUv_Rou_Voi
Inward Snore bass	Postdorso-uvulaire_Inspiré_Roulé	PoDoUv_Ins_Rou
Inward Vocal Snore bass	Postdorso-uvulaire_Inspiré_Roulé_Voisé	PoDoUv_Ins_Rou_Voi

NOM BEATBOX	FORMULE PHONETIQUE	FORMULE PHONETIQUE ABREGEE
Trumpet	Oro-pharyngal_Fricatif_Voisé_Bi-labial_Fricatif_Vrombi	OrPh_Fri_Voi_BiLa_Fri_Vro
House kick	Apico-labial_Inspiré_Explosif_Bi-labial_Projeté_Gloto-pharyngal_Explosif	ApLa_Ins_Exp_BiLa_Pro_GlPh_Exp _GlPh_Exp
Throat bass	Laryngo-pharyngal_Roulé	LaPh_Rou
Throat bass nasalisé	Laryngo-pharyngal_Roulé_Nasalisé	LaPh_Rou_Nas
Techno Dry Kick	Gloto-pharyngal_Explosif	GlPh_Exp
Techno Kick	Gloto-pharyngal_Explosif_Voisé_Nasalisé	GlPh_Exp_Voi_Nas
Esh snare	Gloto-pharyngal_Explosif_Apico-alvéolaire_Fricatif_Voisé	GlPh_Exp_ApAl_Fri_Voi
Cought snare	Gloto-pharyngal_Explosif_Oro-pharyngal_Constrictif	GlPh_Exp_OrPh_Con
Inward cought snare	Gloto-pharyngal_Inspiré_Explosif_Voisé	GlPh_Ins_Exp_Voi
Inward Non Techno Kick	Gloto-pharyngal_Inspiré_Explosif_Non pulmonaire	GlPh_Ins_Exp_NonPul
Long kick	Bi-labial_Explosif_Nasalisé_Oro-pharyngal_Fricatif	BiLa_Exp_Nas_OrPh_Fri
Blowed kick	Bi-labial_Explosif_Fricatif	BiLa_Exp_Fri

PHONETIQUE	ABREVIATION
Bi-labial	BiLa
Labio-dental	LaDe
Dento-labial	DeLa
Apico-labial	ApLa
Apico-postlabial	ApPosLa
Bi-dental	BiDe
Apico-dental	ApDe
Apico-postdental	ApPosDe
Apico-alvéolaire	ApAl
Apico-palatal	ApPa
Plancho-Sublingual	PlSu
Prédorso-prépalatal	PreDoPrePa
Dorso-palatal	DoPa
Postdorso-vélaire	PosDoVe
Postdorso-uvulaire	PoDoUv
Naso-pharyngal	NaPh
Oro-pharyngal	OrPh
Gloto-pharyngal	GlPh
Laryngo-pharyngal	LaPh
Explosif	Exp
Projeté	Pro
Clic	Cli
Fricatif	Fri
Constrictif	Con
Roulé	Rou
Oscillé	Osc
Inspiré	Ins
Non pulmonaire	NonPul
Latéralisé	Lat
Nasalisé	Nas
Voisé	Voi
Sifflé	Sif
Vrombi	Vro

Annexe C

Résultats

Lecture des résultats Sous la ligne des titres, les deux première lignes concernent le *beatboxeur* amateur. L'identifiant comprend les trois premières lettres du nom de famille et la première lettre du prénom. Sur la ligne "raw", nous pouvons observer les résultats bruts. La colonne "#SENT" représente le nombre de phrases et "#WORD" le nombre de mots dans la référence. La cinquième colonne indique le nombre de mots correctement reconnus, les quatre colonnes suivantes indiquent les substitutions, les insertions et les délétions ainsi que le nombre d'erreurs. La dernière colonne indique le nombre de phrases mal reconnues. La deuxième ligne est sensiblement la même, la seule différence étant que les résultats des colonnes 5 à 10 sont en pourcentages. L'avant-dernière colonne représente donc le WER. Les lignes "SUM" représentent les résultats généraux du système selon le même principe.

C.1 Résultats des décodages sur le corpus petit vocabulaire

paramètres

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_CON_A	raw	17	188	164	6	7	18	31	12
ID_CON_A	sys	17	188	87.23	3.19	3.72	9.57	16.49	70.59
ID_PIN_A	raw	22	271	215	18	13	38	69	16
ID_PIN_A	sys	22	271	79.34	6.64	4.80	14.02	25.46	72.73
SUM	raw	39	459	379	24	20	56	100	28
SUM	sys	39	459	82.57	5.23	4.36	12.20	21.79	71.79

FIGURE C.1 – Résultats sur les monophones. (mono)

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_CON_A	raw	17	188	158	8	5	22	35	11
ID_CON_A	sys	17	188	84.04	4.26	2.66	11.70	18.62	64.71
ID_PIN_A	raw	22	271	220	13	14	38	65	15
ID_PIN_A	sys	22	271	81.18	4.80	5.17	14.02	23.99	68.18
SUM	raw	39	459	378	21	19	60	100	26
SUM	sys	39	459	82.35	4.58	4.14	13.07	21.79	66.67

FIGURE C.2 – Résultats sur les triphones. (tri2a)

C.2 Impact du type de microphone

C.2.1 microphone ambiance

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	54	252	174	58	57	20	135	53
ID_Con_A_m_amateur	sys	54	252	69.05	23.02	22.62	7.94	53.57	98.15
ID_Pin_A_m_Pro	raw	80	701	371	237	49	93	379	77
ID_Pin_A_m_Pro	sys	80	701	52.92	33.81	6.99	13.27	54.07	96.25
SUM	raw	134	953	545	295	106	113	514	130
SUM	sys	134	953	57.19	30.95	11.12	11.86	53.93	97.01

FIGURE C.3 – Résultats des décodages avec MA monos pour le microphone d'ambiance

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	54	252	156	60	45	36	141	53
ID_Con_A_m_amateur	sys	54	252	61.90	23.81	17.86	14.29	55.95	98.15
ID_Pin_A_m_Pro	raw	80	701	359	213	31	129	373	77
ID_Pin_A_m_Pro	sys	80	701	51.21	30.39	4.42	18.40	53.21	96.25
SUM	raw	134	953	515	273	76	165	514	130
SUM	sys	134	953	54.04	28.65	7.97	17.31	53.93	97.01

FIGURE C.4 – Résultats des décodages avec MA tri2a pour le microphone d'ambiance

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	54	252	177	60	44	15	119	52
ID_Con_A_m_amateur	sys	54	252	70.24	23.81	17.46	5.95	47.22	96.30
ID_Pin_A_m_Pro	raw	80	701	374	216	42	111	369	77
ID_Pin_A_m_Pro	sys	80	701	53.35	30.81	5.99	15.83	52.64	96.25
SUM	raw	134	953	551	276	86	126	488	129
SUM	sys	134	953	57.82	28.96	9.02	13.22	51.21	96.27

FIGURE C.5 – Résultats des décodages avec MA tri2b pour le microphone d'ambiance

C.2.2 microphone beta

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	54	246	106	120	46	20	186	53
ID_Con_A_m_amateur	sys	54	246	43.09	48.78	18.70	8.13	75.61	98.15
ID_Pin_A_m_Pro	raw	80	723	319	325	96	79	500	74
ID_Pin_A_m_Pro	sys	80	723	44.12	44.95	13.28	10.93	69.16	92.50
SUM	raw	134	969	425	445	142	99	686	127
SUM	sys	134	969	43.86	45.92	14.65	10.22	70.79	94.78

FIGURE C.6 – Résultats des décodages avec MA monos pour le microphone encapsulé

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	54	252	156	60	45	36	141	53
ID_Con_A_m_amateur	sys	54	252	61.90	23.81	17.86	14.29	55.95	98.15
ID_Pin_A_m_Pro	raw	80	701	359	213	31	129	373	77
ID_Pin_A_m_Pro	sys	80	701	51.21	30.39	4.42	18.40	53.21	96.25
SUM	raw	134	953	515	273	76	165	514	130
SUM	sys	134	953	54.04	28.65	7.97	17.31	53.93	97.01

FIGURE C.7 – Résultats des décodages avec MA tri2a pour le microphone encapsulé

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	54	252	177	60	44	15	119	52
ID_Con_A_m_amateur	sys	54	252	70.24	23.81	17.46	5.95	47.22	96.30
ID_Pin_A_m_Pro	raw	80	701	374	216	42	111	369	77
ID_Pin_A_m_Pro	sys	80	701	53.35	30.81	5.99	15.83	52.64	96.25
SUM	raw	134	953	551	276	86	126	488	129
SUM	sys	134	953	57.82	28.96	9.02	13.22	51.21	96.27

FIGURE C.8 – Résultats des décodages avec MA tri2b pour le microphone encapsulé

C.2.3 microphone brauner

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	54	252	119	80	25	53	158	51
ID_Con_A_m_amateur	sys	54	252	47.22	31.75	9.92	21.03	62.70	94.44
ID_Pin_A_m_Pro	raw	80	701	291	281	75	129	485	78
ID_Pin_A_m_Pro	sys	80	701	41.51	40.09	10.70	18.40	69.19	97.50
SUM	raw	134	953	410	361	100	182	643	129
SUM	sys	134	953	43.02	37.88	10.49	19.10	67.47	96.27

FIGURE C.9 – Résultats des décodages avec MA monos pour le microphone brauner

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	54	252	136	75	30	41	146	51
ID_Con_A_m_amateur	sys	54	252	53.97	29.76	11.90	16.27	57.94	94.44
ID_Pin_A_m_Pro	raw	80	701	314	258	60	129	447	78
ID_Pin_A_m_Pro	sys	80	701	44.79	36.80	8.56	18.40	63.77	97.50
SUM	raw	134	953	450	333	90	170	593	129
SUM	sys	134	953	47.22	34.94	9.44	17.84	62.22	96.27

FIGURE C.10 – Résultats des décodages avec MA tri2a pour le microphone brauner

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	54	252	144	80	31	28	139	51
ID_Con_A_m_amateur	sys	54	252	57.14	31.75	12.30	11.11	55.16	94.44
ID_Pin_A_m_Pro	raw	80	701	377	217	62	107	386	75
ID_Pin_A_m_Pro	sys	80	701	53.78	30.96	8.84	15.26	55.06	93.75
SUM	raw	134	953	521	297	93	135	525	126
SUM	sys	134	953	54.67	31.16	9.76	14.17	55.09	94.03

FIGURE C.11 – Résultats des décodages avec MA tri2v pour le microphone brauner

C.2.4 microphone cravate

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	53	248	127	60	24	61	145	45
ID_Con_A_m_amateur	sys	53	248	51.21	24.19	9.68	24.60	58.47	84.91
ID_Pin_A_m_Pro	raw	80	701	401	217	45	83	345	78
ID_Pin_A_m_Pro	sys	80	701	57.20	30.96	6.42	11.84	49.22	97.50
SUM	raw	133	949	528	277	69	144	490	123
SUM	sys	133	949	55.64	29.19	7.27	15.17	51.63	92.48

FIGURE C.12 – Résultats des décodages avec MA monos pour le microphone cravate

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	53	248	123	45	15	80	140	44
ID_Con_A_m_amateur	sys	53	248	49.60	18.15	6.05	32.26	56.45	83.02
ID_Pin_A_m_Pro	raw	80	701	415	182	47	104	333	79
ID_Pin_A_m_Pro	sys	80	701	59.20	25.96	6.70	14.84	47.50	98.75
SUM	raw	133	949	538	227	62	184	473	123
SUM	sys	133	949	56.69	23.92	6.53	19.39	49.84	92.48

FIGURE C.13 – Résultats des décodages avec MA tri2a pour le microphone brauner

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	53	248	139	48	31	61	140	50
ID_Con_A_m_amateur	sys	53	248	56.05	19.35	12.50	24.60	56.45	94.34
ID_Pin_A_m_Pro	raw	80	701	353	252	59	96	407	80
ID_Pin_A_m_Pro	sys	80	701	50.36	35.95	8.42	13.69	58.06	100.00
SUM	raw	133	949	492	300	90	157	547	130
SUM	sys	133	949	51.84	31.61	9.48	16.54	57.64	97.74

FIGURE C.14 – Résultats des décodages avec MA tri2b pour le microphone brauner

C.2.5 microphone SM58 éloigné

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	54	252	171	65	52	16	133	51
ID_Con_A_m_amateur	sys	54	252	67.86	25.79	20.63	6.35	52.78	94.44
ID_Pin_A_m_Pro	raw	80	701	426	230	75	45	350	78
ID_Pin_A_m_Pro	sys	80	701	60.77	32.81	10.70	6.42	49.93	97.50
SUM	raw	134	953	597	295	127	61	483	129
SUM	sys	134	953	62.64	30.95	13.33	6.40	50.68	96.27

FIGURE C.15 – Résultats des décodages avec MA monos pour le microphone SM58 éloigné

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	54	252	174	57	45	21	123	50
ID_Con_A_m_amateur	sys	54	252	69.05	22.62	17.86	8.33	48.81	92.59
ID_Pin_A_m_Pro	raw	80	701	439	206	59	56	321	78
ID_Pin_A_m_Pro	sys	80	701	62.62	29.39	8.42	7.99	45.79	97.50
SUM	raw	134	953	613	263	104	77	444	128
SUM	sys	134	953	64.32	27.60	10.91	8.08	46.59	95.52

FIGURE C.16 – Résultats des décodages avec MA tri2a pour le microphone SM58 éloigné

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	54	252	132	93	51	27	171	52
ID_Con_A_m_amateur	sys	54	252	52.38	36.90	20.24	10.71	67.86	96.30
ID_Pin_A_m_Pro	raw	80	701	157	436	64	108	608	80
ID_Pin_A_m_Pro	sys	80	701	22.40	62.20	9.13	15.41	86.73	100.00
SUM	raw	134	953	289	529	115	135	779	132
SUM	sys	134	953	30.33	55.51	12.07	14.17	81.74	98.51

FIGURE C.17 – Résultats des décodages avec MA tri2b pour le microphone SM58 éloigné

C.2.6 microphone SM58 proche

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	55	252	143	79	37	30	146	52
ID_Con_A_m_amateur	sys	55	252	56.75	31.35	14.68	11.90	57.94	94.55
ID_Pin_A_m_Pro	raw	80	701	407	243	79	51	373	79
ID_Pin_A_m_Pro	sys	80	701	58.06	34.66	11.27	7.28	53.21	98.75
SUM	raw	135	953	550	322	116	81	519	131
SUM	sys	135	953	57.71	33.79	12.17	8.50	54.46	97.04

FIGURE C.18 – Résultats des décodages avec MA monos pour le microphone SM58 proche

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	55	252	147	82	29	23	134	52
ID_Con_A_m_amateur	sys	55	252	58.33	32.54	11.51	9.13	53.17	94.55
ID_Pin_A_m_Pro	raw	80	701	388	254	61	59	374	79
ID_Pin_A_m_Pro	sys	80	701	55.35	36.23	8.70	8.42	53.35	98.75
SUM	raw	135	953	535	336	90	82	508	131
SUM	sys	135	953	56.14	35.26	9.44	8.60	53.31	97.04

FIGURE C.19 – Résultats des décodages avec MA tri2a pour le microphone SM58 proche

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	55	252	138	85	51	29	165	52
ID_Con_A_m_amateur	sys	55	252	54.76	33.73	20.24	11.51	65.48	94.55
ID_Pin_A_m_Pro	raw	80	701	202	488	349	11	848	80
ID_Pin_A_m_Pro	sys	80	701	28.82	69.61	49.79	1.57	120.97	100.00
SUM	raw	135	953	340	573	400	40	1013	132
SUM	sys	135	953	35.68	60.13	41.97	4.20	106.30	97.78

Figure C.20 – Résultats des décodages avec MA tri2b pour le microphone SM58 proche

C.3 Influence de la variabilité

C.3.1 Sur SM58 éloigné

Normal

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	55	376	177	46	155	153	354	52
ID_Con_A_m_amateur	sys	55	376	47.07	12.23	41.22	40.69	94.15	94.55
ID_Pin_A_m_pro	raw	79	164	97	9	2	58	69	41
ID_Pin_A_m_pro	sys	79	164	59.15	5.49	1.22	35.37	42.07	51.90
SUM	raw	134	540	274	55	157	211	423	93
SUM	sys	134	540	50.74	10.19	29.07	39.07	78.33	69.40

Figure C.21 – Résultats du décodage sur le mode normal du microphone SM58 éloigné avec des MA monos

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	55	376	195	64	232	117	413	52
ID_Con_A_m_amateur	sys	55	376	51.86	17.02	61.70	31.12	109.84	94.55
ID_Pin_A_m_pro	raw	79	164	108	10	2	46	58	38
ID_Pin_A_m_pro	sys	79	164	65.85	6.10	1.22	28.05	35.37	48.10
SUM	raw	134	540	303	74	234	163	471	90
SUM	sys	134	540	56.11	13.70	43.33	30.19	87.22	67.16

Figure C.22 – Résultats du décodage sur le mode normal du microphone SM58 éloigné avec des MA monos+LDA et MLLT

Débutant

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Pin_A_m_pro	raw	79	161	66	62	8	33	103	57
ID_Pin_A_m_pro	sys	79	161	40.99	38.51	4.97	20.50	63.98	72.15
SUM	raw	79	161	66	62	8	33	103	57
SUM	sys	79	161	40.99	38.51	4.97	20.50	63.98	72.15

Figure C.23 – Résultats du décodage sur le mode débutant du microphone SM58 éloigné avec des MA monos

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Pin_A_m_pro	raw	79	161	85	61	14	15	90	55
ID_Pin_A_m_pro	sys	79	161	52.80	37.89	8.70	9.32	55.90	69.62
SUM	raw	79	161	85	61	14	15	90	55
SUM	sys	79	161	52.80	37.89	8.70	9.32	55.90	69.62

FIGURE C.24 – Résultats du décodage sur le mode débutant du microphone SM58 éloigné avec des MA monos+LDA et MLLT

Avec variations

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	20	106	30	44	2	32	78	17
ID_Con_A_m_amateur	sys	20	106	28.30	41.51	1.89	30.19	73.58	85.00
ID_Pin_A_m_pro	raw	79	446	132	211	24	103	338	73
ID_Pin_A_m_pro	sys	79	446	29.60	47.31	5.38	23.09	75.78	92.41
SUM	raw	99	552	162	255	26	135	416	90
SUM	sys	99	552	29.35	46.20	4.71	24.46	75.36	90.91

FIGURE C.25 – Résultats du décodage sur le mode variations du microphone SM58 éloigné avec des MA monos

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	20	106	29	51	5	26	82	19
ID_Con_A_m_amateur	sys	20	106	27.36	48.11	4.72	24.53	77.36	95.00
ID_Pin_A_m_pro	raw	79	446	142	204	36	100	340	77
ID_Pin_A_m_pro	sys	79	446	31.84	45.74	8.07	22.42	76.23	97.47
SUM	raw	99	552	171	255	41	126	422	96
SUM	sys	99	552	30.98	46.20	7.43	22.83	76.45	96.97

FIGURE C.26 – Résultats du décodage sur le mode variations du microphone SM58 éloigné avec des MA monos+LDA et MLLT

C.3.2 Sur SM58 proche

Normal

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	55	376	177	46	155	153	354	52
ID_Con_A_m_amateur	sys	55	376	47.07	12.23	41.22	40.69	94.15	94.55
ID_Pin_A_m_pro	raw	79	164	97	9	2	58	69	41
ID_Pin_A_m_pro	sys	79	164	59.15	5.49	1.22	35.37	42.07	51.90
SUM	raw	134	540	274	55	157	211	423	93
SUM	sys	134	540	50.74	10.19	29.07	39.07	78.33	69.40

FIGURE C.27 – Résultats du décodage sur le mode normal du microphone SM58 proche avec des MA monos

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	55	376	195	64	232	117	413	52
ID_Con_A_m_amateur	sys	55	376	51.86	17.02	61.70	31.12	109.84	94.55
ID_Pin_A_m_pro	raw	79	164	108	10	2	46	58	38
ID_Pin_A_m_pro	sys	79	164	65.85	6.10	1.22	28.05	35.37	48.10
SUM	raw	134	540	303	74	234	163	471	90
SUM	sys	134	540	56.11	13.70	43.33	30.19	87.22	67.16

FIGURE C.28 – Résultats du décodage sur le mode normal du microphone SM58 proche avec des MA monos+LDA et MLLT

Débutant

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Pin_A_m_pro	raw	79	161	66	62	8	33	103	57
ID_Pin_A_m_pro	sys	79	161	40.99	38.51	4.97	20.50	63.98	72.15
SUM	raw	79	161	66	62	8	33	103	57
SUM	sys	79	161	40.99	38.51	4.97	20.50	63.98	72.15

FIGURE C.29 – Résultats du décodage sur le mode débutant du microphone SM58 proche avec des MA monos

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Pin_A_m_pro	raw	79	161	85	61	14	15	90	55
ID_Pin_A_m_pro	sys	79	161	52.80	37.89	8.70	9.32	55.90	69.62
SUM	raw	79	161	85	61	14	15	90	55
SUM	sys	79	161	52.80	37.89	8.70	9.32	55.90	69.62

FIGURE C.30 – Résultats du décodage sur le mode normal du microphone SM58 proche avec des MA monos+LDA et MLLT

Avec variations

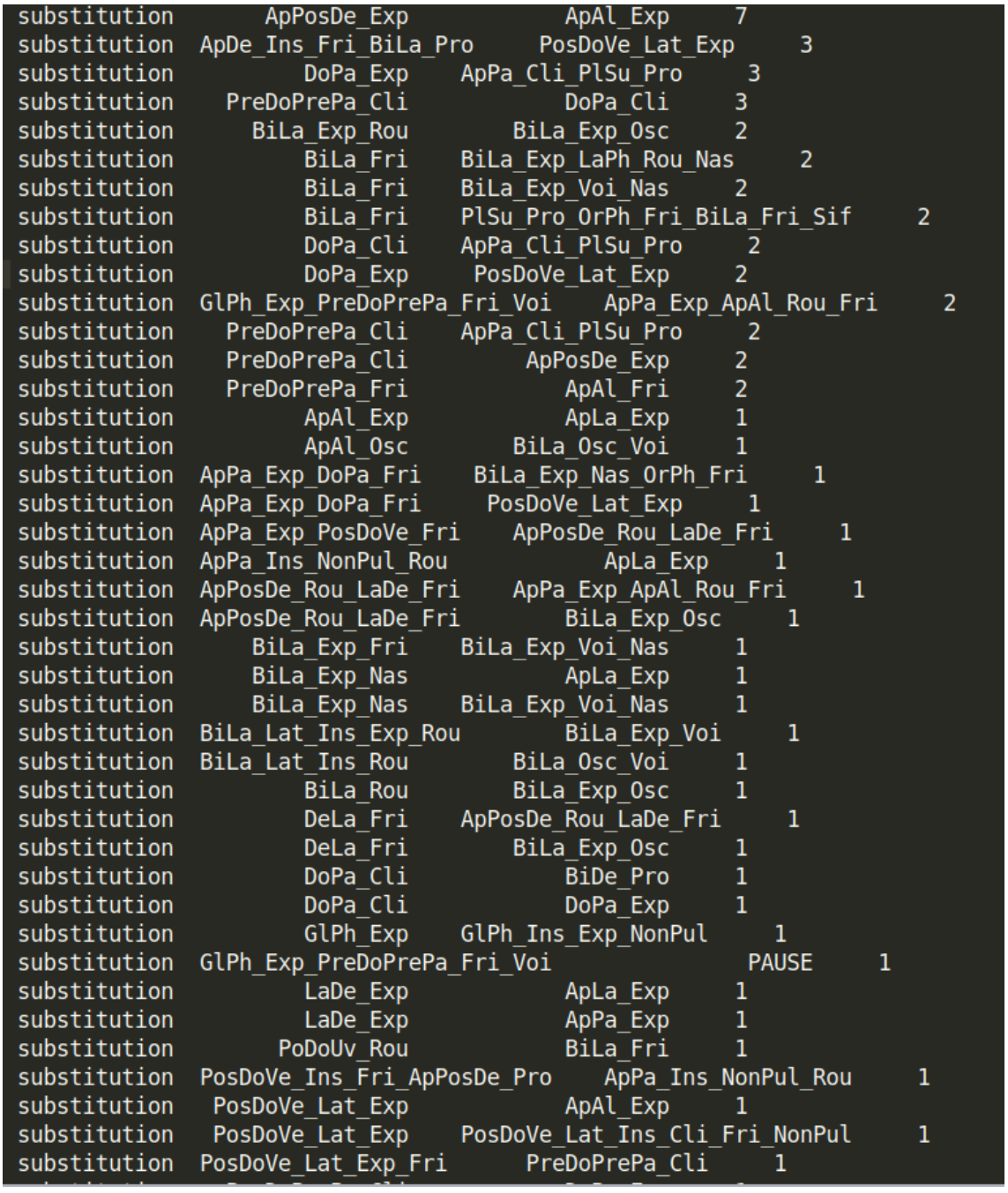
SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	20	106	30	44	2	32	78	17
ID_Con_A_m_amateur	sys	20	106	28.30	41.51	1.89	30.19	73.58	85.00
ID_Pin_A_m_pro	raw	79	446	132	211	24	103	338	73
ID_Pin_A_m_pro	sys	79	446	29.60	47.31	5.38	23.09	75.78	92.41
SUM	raw	99	552	162	255	26	135	416	90
SUM	sys	99	552	29.35	46.20	4.71	24.46	75.36	90.91

FIGURE C.31 – Résultats du décodage sur le mode variations du microphone SM58 proche avec des MA monos

SPEAKER	id	#SENT	#WORD	Corr	Sub	Ins	Del	Err	S.Err
ID_Con_A_m_amateur	raw	20	106	29	51	5	26	82	19
ID_Con_A_m_amateur	sys	20	106	27.36	48.11	4.72	24.53	77.36	95.00
ID_Pin_A_m_pro	raw	79	446	142	204	36	100	340	77
ID_Pin_A_m_pro	sys	79	446	31.84	45.74	8.07	22.42	76.23	97.47
SUM	raw	99	552	171	255	41	126	422	96
SUM	sys	99	552	30.98	46.20	7.43	22.83	76.45	96.97

FIGURE C.32 – Résultats du décodage sur le mode variations du microphone SM58 proche avec des MA monos+LDA et MLLT

C.3.3 Détail des substitutions pour les meilleurs systèmes sur monophones et monophones avec adaptations LDA et MLLT [1]



```
substitution       ApPosDe_Exp            ApAl_Exp     7
substitution  ApDe_Ins_Fri_BiLa_Pro     PosDoVe_Lat_Exp     3
substitution         DoPa_Exp    ApPa_Cli_PlSu_Pro     3
substitution    PreDoPrePa_Cli            DoPa_Cli     3
substitution      BiLa_Exp_Rou        BiLa_Exp_Osc     2
substitution         BiLa_Fri    BiLa_Exp_LaPh_Rou_Nas     2
substitution         BiLa_Fri    BiLa_Exp_Voi_Nas     2
substitution         BiLa_Fri    PlSu_Pro_OrPh_Fri_BiLa_Fri_Sif     2
substitution         DoPa_Cli    ApPa_Cli_PlSu_Pro     2
substitution         DoPa_Exp     PosDoVe_Lat_Exp     2
substitution  GlPh_Exp_PreDoPrePa_Fri_Voi    ApPa_Exp_ApAl_Rou_Fri     2
substitution    PreDoPrePa_Cli    ApPa_Cli_PlSu_Pro     2
substitution    PreDoPrePa_Cli         ApPosDe_Exp     2
substitution    PreDoPrePa_Fri            ApAl_Fri     2
substitution         ApAl_Exp            ApLa_Exp     1
substitution         ApAl_Osc        BiLa_Osc_Voi     1
substitution  ApPa_Exp_DoPa_Fri    BiLa_Exp_Nas_OrPh_Fri     1
substitution  ApPa_Exp_DoPa_Fri     PosDoVe_Lat_Exp     1
substitution  ApPa_Exp_PosDoVe_Fri    ApPosDe_Rou_LaDe_Fri     1
substitution  ApPa_Ins_NonPul_Rou            ApLa_Exp     1
substitution  ApPosDe_Rou_LaDe_Fri    ApPa_Exp_ApAl_Rou_Fri     1
substitution  ApPosDe_Rou_LaDe_Fri        BiLa_Exp_Osc     1
substitution      BiLa_Exp_Fri    BiLa_Exp_Voi_Nas     1
substitution      BiLa_Exp_Nas            ApLa_Exp     1
substitution      BiLa_Exp_Nas    BiLa_Exp_Voi_Nas     1
substitution  BiLa_Lat_Ins_Exp_Rou        BiLa_Exp_Voi     1
substitution  BiLa_Lat_Ins_Rou        BiLa_Osc_Voi     1
substitution         BiLa_Rou        BiLa_Exp_Osc     1
substitution         DeLa_Fri    ApPosDe_Rou_LaDe_Fri     1
substitution         DeLa_Fri        BiLa_Exp_Osc     1
substitution         DoPa_Cli            BiDe_Pro     1
substitution         DoPa_Cli            DoPa_Exp     1
substitution         GlPh_Exp    GlPh_Ins_Exp_NonPul     1
substitution  GlPh_Exp_PreDoPrePa_Fri_Voi               PAUSE     1
substitution         LaDe_Exp            ApLa_Exp     1
substitution         LaDe_Exp            ApPa_Exp     1
substitution       PoDoUv_Rou            BiLa_Fri     1
substitution  PosDoVe_Ins_Fri_ApPosDe_Pro    ApPa_Ins_NonPul_Rou     1
substitution   PosDoVe_Lat_Exp            ApAl_Exp     1
substitution   PosDoVe_Lat_Exp    PosDoVe_Lat_Ins_Cli_Fri_NonPul     1
substitution  PosDoVe_Lat_Exp_Fri      PreDoPrePa_Cli     1
```

FIGURE C.33 – Substitutions lors du décodage avec une probabilité d'apparition du silence à 0,7, un contexte pause gauche et droit dans le lexique et des HMM à 5 états. Modèle acoustique monophone.

1. Le son PlSu_Pro_OrPh_Fri_BiLa_Fri_Sif n'est pas à prendre en compte car a souffert d'un problème de référencement dans le dictionnaire.

```
substitution  ApPosDe_Rou_LaDe_Fri    PlSu_Pro_OrPh_Fri_BiLa_Fri_Sif     4
substitution         DeLa_Fri    PlSu_Pro_OrPh_Fri_BiLa_Fri_Sif     4
substitution  GlPh_Exp_PreDoPrePa_Fri_Voi    PlSu_Pro_OrPh_Fri_BiLa_Fri_Sif     4
substitution  ApDe_Ins_Fri_BiLa_Pro            ApPa_Exp     3
substitution         BiLa_Osc        BiLa_Osc_Voi     3
substitution      ApPosDe_Exp            ApAl_Exp     2
substitution  ApPosDe_Exp_LaDe_Fri    PlSu_Pro_OrPh_Fri_BiLa_Fri_Sif     2
substitution         DoPa_Exp    ApPa_Cli_PlSu_Pro     2
substitution         LaDe_Exp            ApLa_Exp     2
substitution       PoDoUv_Rou    PoDoUv_Ins_Rou_Voi     2
substitution         ApAl_Osc    PlSu_Pro_OrPh_Fri_BiLa_Fri_Sif     1
substitution  ApDe_Ins_Fri_BiLa_Pro            LaDe_Exp     1
substitution  ApLa_Ins_Exp_BiLa_Pro_GlPh_Exp    BiLa_Exp_Voi_Nas     1
substitution  ApPa_Exp_DoPa_Fri    PlSu_Pro_OrPh_Fri_BiLa_Fri_Sif     1
substitution  ApPa_Exp_PosDoVe_Fri         PosDoVe_Fri     1
substitution  ApPosDe_Rou_LaDe_Fri    BiLa_Exp_LaDe_Fri     1
substitution         BiDe_Pro    PlSu_Pro_OrPh_Fri_BiLa_Fri_Sif     1
substitution         BiLa_Fri            BiLa_Osc     1
substitution         BiLa_Fri    PlSu_Pro_OrPh_Fri_BiLa_Fri_Sif     1
substitution     BiLa_Fri_Sif    ApPa_Cli_PlSu_Pro     1
substitution  BiLa_Lat_Ins_Rou        BiLa_Osc_Voi     1
substitution  BiLa_Lat_Ins_Rou    PlSu_Pro_OrPh_Fri_BiLa_Fri_Sif     1
substitution         BiLa_Rou        BiLa_Exp_Osc     1
substitution         DoPa_Cli            DoPa_Exp     1
substitution         DoPa_Exp            ApAl_Exp     1
substitution  GlPh_Exp_PreDoPrePa_Fri_Voi    ApPa_Exp_ApAl_Fri     1
substitution  GlPh_Ins_Exp_NonPul            GlPh_Exp     1
substitution         LaDe_Exp        BiLa_Exp_Voi     1
substitution     LaPh_Rou_Nas    PlSu_Pro_OrPh_Fri_BiLa_Fri_Sif     1
substitution  PoDoUv_Ins_Rou_Voi    PlSu_Pro_OrPh_Fri_BiLa_Fri_Sif     1
substitution       PoDoUv_Rou    PlSu_Pro_OrPh_Fri_BiLa_Fri_Sif     1
substitution  PosDoVe_Ins_Fri_ApPosDe_Pro            ApAl_Exp     1
substitution  PosDoVe_Ins_Fri_ApPosDe_Pro    PlSu_Pro_OrPh_Fri_BiLa_Fri_Sif     1
substitution   PosDoVe_Lat_Exp    ApPa_Exp_DoPa_Fri     1
substitution  PosDoVe_Lat_Exp_Fri    PlSu_Pro_OrPh_Fri_BiLa_Fri_Sif     1
substitution  PosDoVe_Lat_Exp_Fri         PoDoUv_Rou     1
substitution    PreDoPrePa_Cli            DoPa_Cli     1
substitution    PreDoPrePa_Fri    ApPosDe_Rou_LaDe_Fri     1
substitution    PreDoPrePa_Fri    PlSu_Pro_OrPh_Fri_BiLa_Fri_Sif     1
```

FIGURE C.34 – Substitutions lors du décodage avec une probabilité d'apparition du silence à 0,7 et un contexte pause gauche et droit dans le lexique. Modèle acoustique monophone avec adaptations LDA et MLLT.