Rapport Partie Son Projet PILE

Maël Derio

22 février 2007

Chapitre 1

Etat de l'art

Introduction

L'étude de la production vocale du bébé intéresse de nombreux acteurs dans des domaines de recherche à priori assez éloignés comme la psychologie, la phonétique, le traitement du signal, les interactions homme machine. Nous cherchons pour notre part à comprendre l'évolution de la production vocale du bébé à partir d'un traitement et d'une classification automatique des vocalisations du bébé. Pour cela nous avons d'abord cherché dans la bibliographie des exemples de classification et de traitement automatique du signal. Cela nous a conduit à nous poser la question du matériau qu'il est pertinent d'étudier : analysons nous l'étendue complète des vocalisations du bébé ou une partie d'entre elles et comment pouvons nous qualifier, et quantifier, l'étendue des vocalisations du bébé ou ses sous parties.

1.1 Classifications automatiques

Plusieurs équipes de recherche se sont penchées sur la question de la classification automatique des vocalisations du bébé. Il est intéressant en premier lieu de connaitre les objectifs qu'ils souhaitaient atteindre dans leur recherche et les méthodes de classifications qu'ils ont employées, et ensuite de discuter de leur choix de vocalisations.

1.1.1 Présentation des études

Les différentes équipes ont essentiellement pour objectif le diagnostic de pathologies diverses. La première équipe fonctionne grâce à une collaboration entre les départements Computer and Information sciences et Speech-Language Pathology and Audiology de la Northeastern University de Boston. L'équipe (Fell et al. 1996, 1998, 2002, 2003, 2005) a pour but de détecter automatiquement les bébés qui auront des troubles de la communication. Leur système, nommé EVA (Early Vocalisation Analyser) repose sur l'analyse de l'age du bébé. Si l'âge trouvé pour un bébé à partir de l'analyse de ses babbils et de

la connaissance de l'évolution normale du bébé au cours des âges, est différent (dans une certaine mesure) de son âge réel, il est diagnostiqué comme pouvant présenter des troubles du langage. Les tests se sont fait avec un enfant diagnostiqué avec apraxia, un avec le syndrome de Down, un avec hydrocéphalie, et trois (dont un prématuré) avec des retards du développement moteur. Les auteurs ont également développés un système de soins reposant sur l'EVA, le VisiBabble. Ce système est utilisé avec des enfants victimes de troubles du langage. Il détecte les vocalisations proches du langage (comme définies par les auteurs) et donne un retour visuel censé encourager l'enfant dans cette voie. Le système permet aussi l'analyse des vocalisations du bébé qui sont enregistrées au cours de la séance. L'équipe travaille également sur un système censé stimuler les vocalisations des enfants sourds, le Deaf-Infant Babble Stimulator. La seconde équipe (Lederman 2002, Lederman et al. 2002) vient du département d'ingénierie électrique et informatique de la faculté d'ingénierie de l'université Ben Gourion (Beer Sheva, Israël)en collaboration avec plusieurs équipes dans le domaine médical. Elle réalise une classification automatique des cris des bébés suivant la pathologie. Plusieurs bases de données sont utilisées, la première comprend les cris provenant d'une centaine d'enfants malades et sains, arrivés à terme ou avant terme. La seconde base de donnée contient des cris produites par 7 enfant ayant une fente palatine. Chacun des enfants a été enregistré pendant plusieurs semaines avec et sans plaque palatine. La troisème base de donnée contient les cris de 992 enfants exposés à différents types de drogues in utero. Enfin la quatrième base de donnée inclut les enregistrements d'enfants affectés de diverses pathologie comme le Cri-du-Chat ou le syndrome de Down. Il est à noter que la plupart des enfants sont des nouveaux nés de moins de 3 mois. Quelques uns ont plus de trois mois. L'idée des auteurs repose sur des études qui montrent que le cri change suivant l'état psychologique du bébé, faim, douleur, peur par exemple, et suivant la pathologie du bébé. De ce point de vue, le cri peut être considéré comme un outil de diagnostic précoce non invasif. La troisième équipe vient du département de traitement de la parole de l'université Oriente de Cuba et de l'université de Twente aux Pays Bas . Ils ont (Ortiz et al. 2003, Ekkel 2002) pour but de classer automatiquement les cris produits par les enfants souffrants d'atteintes au systeme nerveux central dues a l'hypoxia d'un côté et par les enfants sains d'autre part. Les enfants sont des nouveaux nés. L'idée de départ de l'étude est que le cri de douleur produit par le bébé est un mouvement spontané, un réflexe produit par le système nerveux central et qu'un enfant souffrant d'hypoxia aura par conséquent une réponse différente d'un enfant sain. La quatrième équipe travaille à l'institut national d'astrophysique optique et électronique (INAOE) de Mexico. Leur objectif (Orozco Garcia 2003a, 2003b, Reyes Galaviz 2004) est de classer automatiquement les cris d'enfants sourds et sains, et également d'enfants victime d"'asphyxia". Les enfants sont des nouveaux nés. Leur idée, reposant sur des études neuropsychologiques, est que le cri permet de connaitre l'état physique du bébé et donc de détecter différentes pathologies, principalement cérébrales. D'autres équipes travaillent sur le sujet, à priori plutôt sur l'étude des cris produits par les bébés et toujours dans une optique de diagnostic.

1.1.2 Description et classification des vocalisations

Descripteurs

Différents descripteurs ont été utilisés sur ces différentes études. Voici les descripteurs utilisés suivant les études :

Ortiz et al. (2003) et Ekkel (2002) définissent une unité temporelle de cri comme une exhalation ininterrompue dans laquelle un son est produit entre deux inhalations se succédant. A noter que le bébé est soumis à un stimulus. Le descripteur first latency correspond à la durée entre ce stimulus et le début du cri.

f0
melody contour
formants (f1,f2)
voisement
énergie
longueur d'une unité de cri
First latency
Présence d'aberrations sonores

Pour les équipes de l'INAOE et de Ben Gourion, les descipteurs utilisés sont (respectivement) les suivants :

INAOE
MFCC
LPC

Ben Gourion
LPC
LPCC
PARCOR
LAR
filter banks
energy
log-energy
Del-LPCC
Del-log-energy
Voiced cry time
Silent time

Les descripteurs utilisés dans le travail de Fell et al. sont assez différents car le type de vocalisations étudiées n'est pas le même. Les auteurs se reposent sur la théorie de détection des landmarks de Stevens (1992) et l'adaptent pour les bébés. Ce système permet de reconnaitre les différents types de babbils produit par le bébé et de les classer. A cela viennent se rajouter un taux de vocalisations par unité de temps, la durée d'une utterance (ici définie par une vocalisation suivie d'un silence assez long) et la fréquence fondamentale. Ces différentes descriptions se combinent pour arriver à 93 descripteurs.

Méthodes de classification

Les méthodes utilisés par les différentes équipes sont les suivantes :

Ortiz, Ekkel	Probabilistic Neural Network + bootstrap aggregation
Orozco Garcia	PCA puis feed forward network avec utilisation de la méthode Scale Conjugate Gradient
Reyes Galaviz	Input Delay Neural Network entrainé avec l'algorithme adaptive learning rate back propagation
Lederman	Continuous Density Hidden Markov Model
Fell	Principal Component Analysis

1.1.3 Choix des vocalisations étudiées

Les différentes études montrées plus haut donnent à priori de bons résultats de classification. Ces bons résultats sont dûes principalement au choix des vocalisations étudiées. En effet aucun de ces travaux n'étudie l'ensemble de la production vocale du bébé. Fell et al. étudient les syllabes produites par le bébé et les autres travaux concernent les cris du bébé. Pour les cris, le choix est relativement simple à faire quand un stimulus est exercé sur le bébé, un cri est défini précisemment comme la réponse à ce stimulus. Par exemple, dans la méthode utilisée par Ekkel, la vocalisation produite par le bébé après un test sanguin est un cri de douleur. Si aucun stimulus n'est appliqué, le choix qu'une vocalisation est un cri ou non est fait par un expert, un pédiatre dans le cas de l'équipe de l'INAOE (Orozco Garcia, Reyes Galaviz). Pour les syllabes, dans le travail de Fell et al., une segmentation manuelle du signal est d'abord effectuée avant l'analyse, même si les auteurs souhaitent à terme que cette segmentation soit automatique. Tous les sons qu'ils considèrent n'être pas de la parole ainsi que ce qu'ils appellent les sons végétatifs et réflexifs sont éliminés. Les sons qui ne sont pas de la parole incluent les cris, le rire, les sons d'effort. Les sons végétatifs et réflexifs comprennent la toux, les claquements de lèvres, les hoquets, les éternuements. Tous les bruits, les sons produits par d'autres personnes, les sons produits par le bébé mais masqués par un jouet ou autres sont éliminés. Ce qui permet donc dans tous ces cas une bonne classification est un bon ciblage du matériau sonore à analyser, un choix d'une sous catégorie très précise de sons produits par le bébé. Les choix fait dans ces études, comme nous l'avons vu, de sous catégories (cris, babils, sons végétatifs, ...) ont été réalisés par des experts, ou sur le fondement de travaux antérieurs définissant ces sous catégories. Ces sous catégories ne sont pas définies et quantifiés précisemment d'un point de vue acoustique directement dans ces travaux. Ces sous catégories ne semblent par ailleurs pas forcément cohérentes entre les différents travaux, ni les termes utilisés pour décrire la production vocale du bébé constants entre les auteurs. Il nous semble donc utile à ce stade d'étudier la catégorisation de la production vocale du bébé, ce que nous allons faire dans le chapitre suivant.

Conclusion

Nous avons présenté les différents travaux portant sur la classification automatique des sons produits par le bébé. Différentes méthodes de classification ont été utilisées ainsi que des descriptions du son différentes. Les résultats, dans une optique de diagnostic sont bons, mais il nous semble que la qualité de ces résultats dépend de la préparation des données, de la segmentation faite au préalable, qu'elle soit automatique ou non. Dans les travaux que nous avons décrit, des critères ou méthode de segmentation assez précis ont été décidés (par exemple, les vocalisations étudiées sont celles qui viennent après un stimulus douloureux), qui permettent d'avoir un ensemble de vocalisation assez homogène à étudier d'un point de vue acoustique. Pour définir une méthode de segmentation, ces différentes études se sont appuyées sur des travaux antérieurs cherchant la meilleure manière de catégoriser la production vocale du bébé. Ces derniers peuvent nous apporter un éclairage sur la meilleure manière d'aborder notre problème de classification et de segmentation, nous allons donc étudier cet aspect dans le chapitre suivant.

1.2 Catégorisation de la production vocale du bébé

Introduction

L'évolution de la production vocale du bébé a interessé de nombreux auteurs. L'évolution de la hauteur des vocalisations au cours des âges, des contours mélodiques, l'âge auquel apparait une caractéristique spécifique comme la production de babils, etc ... Ces études, comme la notre ont besoin d'un fondement commun qui permet de comparer les mesures. Nous avons besoin d'une catégorisation commune qui nous permette de dire que telle ou telle caractéristique de la production vocale du bébé évolue dans un sens ou dans un autre. Sans ce fondement commun, les études arrivent à des résultats différents. Il a par exemple été longtemps admis que la production de babils apparait vers 6 mois en général. Des études récentes avancent plutôt l'âge de 4 mois. Ce qui différencie ces travaux, ce n'est pas le nombre de bébés étudiés ou la méthode utilisée pour enregistrer les sons, mais la façon de catégoriser la production vocale du bébé et la prise en compte, ou non, de certaines vocalisations. Ces catégorisations se développent à partir d'à priori sur le développement de la production vocale du bébé, sur la fonction des vocalisations, sur la psychologie du bébé ou sur un fondement physiologique ou acoustique. Voyons quelles sont les différentes interprétations qui existent et les problèmes que cela engendre.

1.2.1 Catégorisations

Au travers de nombreuses études qui s'intéressent à la production vocale du bébé et à son évolution vers le langage (ou non) (Rothgänger, 2003, Fell et al., 1996, Petitto et al., 2004, de Boysson-Bardies 1989, Hsu et al., 2000, Esling, 2004, ...), on constate la nécéssité d'une catégorisation de la production vocale du bébé. Si l'on veut étudier les babils, il faut définir ce que c'est et le différencier des autres productions vocales du bébé, comme le fait

par exemple de Boysson Bardies (1989). Elle s'appuie pour cela sur des catégorisations construites par plusieurs auteurs cités également dans toutes ces études : Oller, Stark, Koopmans van Beinum, pour ne citer que les principaux. Nous allons en faire un rapide historique reconstitué à partir des informations données par les différents auteurs que nous avons cités (entre autres) ainsi que Nathani et al. (2001) ou Bettany, (2004).

Historique

Les recherches se sont d'abord focalisees sur les sons "canoniques", c'est à dire les syllabes bien formées, qui arrivent en general apres 6 mois, car facilement comparables aux syllabes produites à l'âge adulte.Des tentatives ont d'abord été faites pour catégoriser les sons produits par le bébé a l'aide l'aphabet phonétique international (IPA) ou de manière acoustique. D'après Nathani et al., ce fut un échec car cette catégorisation ne prend pas en compte la "cible", c'est à dire le langage mature. C'est ainsi que Oller et al. ont construit une catégorisation partant d'une analyse acoustique, mais augmentée d'une couche interprétative qui relie le son produit par le bébé aux sons matures, une couche que ces auteurs nomment "infraphonologique". Cette couche interprétative permet de voir à quel point un son produit par le bébé s'approche du langage mature, ce que ne permettent pas les méthodes strictement acoustiques ou reposant sur l'IPA. Cette méthode permet de prendre en compte les sons "précanoniques", c'est à dire intervenant avant 5-6 mois pour les enfants normaux et qui ne sont pas des syllabes bien formées. D'autres perspectives ont été développées d'après Nathani et al, comme celle de Stark et al. Ils ont plutôt utilisé une approche motrice en regardant la complexité articulatoire des sons produits par le bébé. Enfin Koopmans Van Beinum et al. ont utilisé une approche sensorimotrice en s'appuyant sur les caractéristiques phonatoires et articulatoires de ces sons. Ces méthodes de catégorisation n'ont en général pas pris en compte une partie des sons émis par le bébé, mais seulement les "speech-like sounds", c'est à dire les sons ressemblant de près ou de loin au langage suivant leurs critères. Ils mettent de côté pour la plupart les sons qu'ils appellent végétatifs ou réflexifs et reliés d'après ces auteurs aux fonctions biologiques, donc peu intéressant pour la recherche sur le langage. Ce point de vue est discuté dans les recherches d'Esling et al. (Esling, 2004, 2005, Bettany, 2004). Ils partent d'un modèle articulatoire qui leur permet de décrire l'ensemble de la production vocale de l'adulte, à partir des configurations des deux articulateurs du larynx que sont la glotte et le sphincter aryepiglottique. Ces auteurs appliquent ce modèle au bébé, ce qu'il leur permet de prendre en compte (d'après eux) l'ensemble de la production vocale du bébé. Ainsi les sons produits par le bébé sont catégorisés suivant que le sphincter aryepiglottique est plus ou moins constricté et suivant la position de la glotte. Leur approche théorique est différente des autres auteurs évoqués dans le sens où ils décrivent "les vocalisations en tant que pratique phonétiques où l'enfant développe de nouvelles capacités articulatoires" (Esling, 2004). Ainsi ce que certains auteurs ont pu nommer sons végétatifs ou réflexifs correspond en général aux sons constrictés (sphincter aryepiglottique fermé) et peuvent être vu comme une production vocale servant à s'approprier telle ou telle configuration articulatoire ou à s'entrainer à produire une hauteur donnée de son par exemple. Ajoutons à cela que Es-

ling et al. font une catégorisation qui normalise le vocabulaire employé pour parler de la production vocale du bébé, les termes employés pouvant être multiples pour désigner des vocalisations similaires. Finissons en montrant un tableau tiré de Bettany (2004) décrivant la façon dont Esling et al. catégorisent la production vocale du bébé par rapport aux autres travaux sur le sujet, dont ceux des auteurs que nous avons déjà cité :

Esling	**Autres auteurs**
Constricted	
	Strained
	Tense
	Noisy excitation
	Turbulent noise
Whisper	
	Cough
	Pulse
	Pharyngeal friction
	Whispering
	Grunts
Harsh Voice	
Low Pitch	
	Crying
	Fusses
	Hyperphonation
	Disphonation
	Gurgle
	Grunts
	Growling
	Cough
	Gutteral
	Gazouilles
	Moan
	Groan
	Chaos
	Pharyngeal friction
Mid Pitch	
	Crying
	Fusses
	Hyperphonation
	Disphonation
	Gurgle
	Grunts
	Growling
	Cough
	Gutteral
	Gazouilles
	Moan
	Groan
	Chaos
	Pharyngeal friction
High Pitch	
	Fusses
	Squealing
	Screaming
	Shrieking
	Pharyngeal friction
Creaky Voice	
	Pulse
	Glottal pulse
	Vocal Fry
	Vocal tremor
Unconstricted	
	Normal phonation
	Speech like
Modal Voice	
	Gooing
	Cooing
	Yelling
Falsetto	
	Squealing

Discussion

L'historique que nous avons faite permet de voir que le choix d'une catégorisation implique aussi un choix théorique sur la signification que nous accordons aux vocalisations du bébé. Certains articles font par exemple également des distinctions entre "non-distress sounds" (sons non provoqués par la douleur) et autres sons ou entre sons porteurs d'une intention ou non. La catégorisation d'Oller et al. estime que nous sommes capable d'évaluer une proximité entre sons produits par le bébé et les sons matures, produits par les adultes. Les catégorisations se reposant sur un modèle de motricité, comme celui de Stark ou d'Esling repose sur l'hypothèse que toute vocalisation a son importance. Une catégorisation qui se reposerait sur l'acoustique uniquement et qui permettrait d'évaluer une évolution du bébé oriente aussi les résultats dans une direction donnée. Le choix d'une bonne catégorisation a donc toute son importance, nous allons voir maintenant quelles sont les difficultés qui y sont liées.

1.2.2 Segmentation

Pour catégoriser les sons produits par le bébé, il faut auparavant l'enregistrer et segmenter les séquences sonores obtenues en catégories, les coder. Deux approches sont possibles, l'une est l'approche de Esling et al., qui repose sur des observations laryngoscopiques. Des profils laryngeaux des vocalisations adultes ont été développés et adaptés aux bébés. L'autre est celle décrite par Nathani (2001), qui repose sur un ensemble de critères prédéfinis et sur un jugement humain. Les points qu'elle souligne sont mis en relation avec la catégorisation "infraphonologique" de Oller et al., il s'agit donc d'un exemple de critères servant à la segmentation. L'idée principale est que des critères doivent être définis qui permettent à des personnes différentes (et à une même personne à des moments différents) de coder les vocalisations de la même façon. Le problème dans ce genre de recherche est que la fiabilité inter-juges qui peut être faible, même dans des cas simple comme compter le nombre d'utterances (unités temporelles) dans une séquence sonore.
Voici donc les principaux critères sur lesquels il est nécéssaire de s'accorder :

- Choix d'un seuil d'audibilité. A partir de quelle amplitude inclut on un son ?
- Problème de la distinction entre babils et sons végétatifs (éructations, toux, éternuement, ...). La distinction entre ces types de sons n'est pas toujours claire, en particulier lors des sons de grognements ou geignards ou les babils aux milieu d'un rire, d'un hurlement.
- Problème de la segmentation en uttérances. Plusieurs techniques ont été utilisées : celle reposant sur les groupes de respiration a donné de meilleurs résultats que celle reposant sur des pauses entre les uttérances par exemple. Il reste quand meme des différences suivant les laboratoires.
- Reconnaissance des syllabes et de simili-syllabes. Si la syllabe est l'unité de base dans la parole, pour pouvoir comparer ce que produit le bébé a la parole, il faut utiliser un mode de description proche, c'est a dire trouver un équivalent pour la syllabe et pour la transition entre les syllabes, les consonnes. On peut alors compter le nombre

de syllabes par exemple.
- Qualité phonatoire : il n'est pas toujours clair si un son a une qualité phonatoire normale ou s'il a une qualité plus proche d'un couinement ou d'un grognement auquel cas il est classé en tant que couinement ou grognement.
- Classification en types de protophones. Il est utile de classer les segments ou syllabes en differents types. L'approche infraphonologique de Oller (2000) donne une base logique a cette classification. Les protophones sont les sons qui apparaissent avoir une statut fonctionnel pour le bébé. Ils sont opposés aux sons végétatifs qui possèdent des liens avec les fonctions biologiques. Les protophones sont classés différemment suivant leur ressemblance avec les caractéristiques du syllabes canoniques, matures. Il existe 4 types majeurs de protophones : quasi-resonant nuclei, fully resonant nuclei, marginal syllables, et canonical syllables.

Les auteurs se rejoignent sur l'idée que les différents laboratoires doivent utilisent des bases de données d'annotations communes, pour pouvoir utiliser le même vocabulaire pour décrire la production vocale du bébé, ce qui n'est pas le cas encore aujourd'hui.

Conclusion

Nous avons vu qu'il existe différentes manières d'aborder la catégorisation des sons produits par le bébé. Chacune est relié à une façon de voir le bébé et influence les résultats que nous pourrons avoir dans la perspective de l'étude de l'évolution des vocalisations du bébé au cours des mois. Il apparait qu'une seule de ces catégorisations, à notre connaissance, englobe l'ensemble de la production vocale du bébé et est réalisée à partir de critères objectifs, c'est à dire de mesures, plutôt qu'à partir d'un jugement humain, celle de Esling. Pour pouvoir utiliser sa catégorisation ou une autre, l'idéal est d'avoir une base de donnée d'exemples de sons annotés, ou d'avoir l'aide d'un expert entrainé depuis quelques années (d'après les auteurs) à catégoriser les sons d'enfant à partir d'une méthode précise.

Chapitre 2

Travaux effectués

Introduction

L'idée principale de PILE est de comprendre l'émergence du langage et de voir comment s'articulent les différentes modalités geste, voix, regard dans ce but. Nous nous intéressons pour notre part à la partie son. Notre but est de comprendre comment évolue le son produit par le bébé de 3 à 9 mois et de voir quelles différences peuvent présenter les évolutions d'enfants atteints de pathologies qui font croitre la probabilité de ne pas développer le langage. Cela revient pour nous globalement à un problème de classification, nous souhaitons retrouver les classes correspondant à une évolution normale ou à une évolution pathologique. Nous avons à notre disposition pour cela une multitude d'enregistrements vidéos et sonores effectués selon le protocole PILE. Les séquences sonores sont composées essentiellement de 4 types d'évenements sonores : les moments où le bébé s'exprime seul, les moments où la maman s'exprime seule, les moments où le bébé et sa maman s'expriment de concert et enfin les moments ou seul du bruit est présent (bruit de fond, coups, ...). Nous voulons faire un traitement automatique de ces séquences sonores et pour cela nous avons besoin d'extraire, automatiquement, chacun de ces événements sonores. Il s'agit de la première étape, essentielle, de notre travail qui nous a occupé jusque là. Elle consiste donc en la mise en place d'un outil de segmentation automatique.

2.1 Description générale

Notre but est de reconnaitre chacun des événements sonores présents dans les séquences sonores. Nous avons donc un problème de classification à 4 classes. Nous n'utilisons pas la vidéo, nous cherchons donc à reconnaitre chaque classe suivant ses caractéristiques acoustiques. Nous ne connaissons pas à priori ces dernières, nous avons donc choisi de faire une classification supervisée. Cela nécessite d'avoir au préalable un certain nombre d'exemples de chaque classe. Nous avons pour cela annoté manuellement une cinquantaine de séquences sonores. Chacune de ces séquences est découpée en segments correspondant chacun à un événement sonore. Chaque événement sonore est annoté suivant l'une des 4 classes que nous

avons défini plus haut. Nous pouvons à partir de ces exemples, calculer les caractéristiques acoustiques de chaque classe. Cela revient à calculer un certain nombre de descripteurs audio sur les exemples de chaque classe. A partir de cette description audio de chaque classe, nous pouvons tenter de retrouver a quelle classe appartient chaque événement sonore dans une séquence inconnue.

2.1.1 Descripteurs audio

Trames et supertrames

Les données que nous avons à disposition étant sous forme de séquences temporelles d'une durée chacune d'une à plusieurs minutes, nous allons nous intéresser à des descripteurs instantannés, dont la valeur va varier au cours du temps. Chaque valeur est évaluée sur une fenêtre temporelle de courte durée dans laquelle les propriétés locales du signal sont stationnaires : une trame. Les trames se recouvrent partiellement afin de couvrir le signal entier. A chaque trame est associé un label qui correspond à la classe. Les moyennes et variances de ces descripteurs sur des durées plus longues peuvent ensuite être calculées. Ces durées plus longues sont les supertrames et correspondent à plusieurs trames d'affilée. Le label qui leur est associé est le label majoritaire des trames la constituant.

MFCC

Ce descripteur est un descripteur multidimensionnel utilisé de manière classique en reconnaissance de la parole. Il se repose sur un modèle de la parole source-filtre. La source est la glotte et le filtre est le conduit vocal. Chaque être humain possède un conduit vocal différent des autres, ce qui permet de différencier les différentes voix. Le conduit vocal d'un enfant est également différent de celui d'un adulte. Les différences de forme du conduit vocal entrainent des différences dans la forme de l'enveloppe du spectre des signaux produits. C'est cette enveloppe que les MFCC modélisent, en la pondérant également par une modélisation de la façon dont l'oreille humaine perçoit le son.

f0

Les différences de hauteurs entre les voix peuvent être évaluées à partir de la fréquence fondamentale du spectre, c'est à dire la fréquence la plus basse d'un spectre périodique. Ce descripteur modélise la fréquence d'ouverture/fermeture de la glotte dans le modèle source-filtre. Nous avons utilisé ici un algorithme développé à l'Ircam, Yin (Cheveigné et al, 2002), réputé être performant pour le signal de parole. Il utilise la fonction d'autocorrélation pour calculer la fréquence fondamentale.

apériodicité

La glotte vibre plus ou moins (ou pas du tout) suivant les sons que l'on produit. Le spectre n'est ainsi pas forcément périodique et n'a de cette manière pas toujours de

fréquence fondamentale. Le caractère plus ou moins périodique du spectre, le pourcentage d'énergie apériodique du spectre est calculé grâce à ce descripteur. Ce descripteur est calculé également à partir de la fonction d'autocorrélation dans l'algorithme Yin.

spectral centroid

Il s'agit du barycentre du spectre. Il est calculé à l'aide de la toolbox ircamdescriptor développée par Geoffroy Peeters à l'Ircam. Ce descripteur possède 6 dimensions. Chacune de ces dimensions correspond à un calcul du barycentre du spectre suivant différentes échelles d'amplitude ou de fréquence.

- Dimension 1 : amplitude linéaire/fréquence linéaire
- Dimension 2 : calcul à partir du spectre de puissance/fréquence linéaire
- Dimension 3 : amplitude logarithmique/fréquence linéaire
- Dimension 4 : amplitude linéaire/fréquence logarithmique
- Dimension 5 : calcul à partir du spectre de puissance/fréquence logarithmique
- Dimension 6 : amplitude logarithmiqueåmplitude fréquentielle

noisiness

C'est le rapport entre l'énergie de la partie non harmonique du signal et l'énergie totale. Plus un signal est bruité, plus sa "noisiness" est importante. Le calcul de ce descripteur repose sur le calcul des pics du spectre qui sont eux mêmes calculés à partir des valeurs de f0 trouvées dans chaque fenêtre d'analyse. Pour avoir la valeur de ce descripteur avec ircamdescriptor, il faut donc avoir d'abord avoir calculé la f0 avec par exemple l'algorithme Yin ou avec f0 (algorithme développé dans l'équipe Analyse/Synthèse à l'Ircam). Nous avons utilisé pour l'instant f0, mais il pourrait être interessant d'utiliser Yin, ce dernier pouvant être plus performant pour des signaux de parole. Les trajectoires des partiels sont ensuite estimés avec le package additive développé également dans l'équipe Analyse/Synthèse.

Loudness

Il s'agit de l'énergie du signal, de son "volume". Il est calculé dans ircamdescriptor.

Spectral Flatness

Ce descripteur analyse le caractère bruité du signal. Une grande valeur indique un spectre qui possède la même quantité d'énergie dans toutes les bandes spectrales, alors qu'une faible valeur indique un spectre dont l'énergie est concentrée dans un petit nombre de bandes. Il se calcule en faisant le rapport de la moyenne géométrique à la moyenne arithmétique du spectre de puissance. Ce calcul se fait dans 4 bandes de fréquence différentes : 250 à 500 hz, 500 à 1000 hz, 1000 à 2000 hz et 2000 à 4000 hz. Ce descripteur possède donc 4 dimensions. Il fait partie des descripteurs calculés dans ircamdescriptor.

Spectral crest factor

Ce descripteur est proche du précédent. Il se calcule par le rapport entre la valeur maximale dans chaque bande de fréquence et la moyenne arithmétique du spectre de puissance. Ce descripteur possède donc aussi 4 dimensions. Il est calculé également avec ircamdescriptor.

Total Energy

Energie du signal calculée avec ircamdescriptor.

Harmonic Energy

Energie de la partie harmonique du signal. Ce descripteur est calculé à partir de l'estimation de l'amplitude des partiels.

Noise Energy

Energie de la partie bruitée du signal. Ce descripteur est calculé en soustrayant la partie harmonique du signal.

Dérivées temporelles

Nous avons également calculé les dérivées temporelles premières et deuxièmes de ces descripteurs. Cela revient à passer le signal dans des filtres passe haut successifs.

2.1.2 Sélection de descripteurs

Les descripteurs peuvent être redondants ou interférer. Il est utile de faire un choix dans les descripteurs utiles, dans les dimensions utiles à la séparation des données en classes. Nous utilisons pour cela un algorithme de sélection de descripteurs nommé IRMFSP (Inertia Ratio Maximization Feature Space Projection) développé par Geoffroy Peeters (peeters,2003). Nous avons implémenté nous même une version de cet algorithme.

L'algorithme fonctionne de manière itérative : On sélectionne d'abord le descripteur ayant la plus forte valeur du critère r suivant :

$$r = \frac{B}{T} = \frac{\sum_{k=1}^{K} \frac{N_k}{N}(m_{i,k} - m_i)(m_{i,k} - m_i)'}{\frac{1}{N}\sum_{n=1}^{N}(f_{i,n} - m_i)(f_{i,n} - m_i)'}$$

B est l'inertie interclasse, T est l'inertie totale. N est le nombre total de données, N_k est le nombre de données appartenant à la classe k, m_i est le centre de gravité du descripteur fi pour les données appartenant à la classe k.

Ainsi si le rapport est grand, l'inertie inter-classes est grande devant l'inertie intraclasse, ce qui veut dire des classes compactes et bien séparées (l'inertie totale étant la somme de l'inertie intraclasse et de l'inertie interclasse).

Pour éviter la redondance d'un descripteur, une procédure d'orthogonalisation est ensuite appliqué au descripteur choisi. Le ratio r est ensuite calculé à nouveau et si la valeur du critère est assez grande par rapport aux descripteurs sélectionnés précédemment, ce descripteur est sélectionné.

2.1.3 Modélisation

GMM

La distribution de chaque descripteur pour chaque classe est modélisée par une mixture de gaussiennes (GMM). Nous utilisons la toolbox Netlab pour les GMM et l'algorithme EM.

$$p(x|\theta) = \sum_{i=1}^{M} \alpha_i p(x|\theta_i)$$

$p(x|\theta)$ est la distribution des données (valeurs des distributeurs pour chaque trame) pour une classe donnée. $\theta = (\alpha_1, ..., \alpha_M, \theta_1, ..., \theta_M)$ correspond aux paramètres de la distribution des données de d'une classe. α_i est un coefficient de mixture. Il correspond au "poids" de la composante i dans la mixture de gaussiennes. M est le nombre de composantes de la mixture. $p(x|\theta_i)$ est la distribution des données de la composante i. θ_i correspond aux paramètres de la distribution de la composante i. La distribution étant gaussienne par hypothèse, ces paramètres sont la moyenne et la variance (matrice de variance/covariance).

Nous voulons trouver les paramètres de la distibution pour chaque classe, ce qui peut se faire en maximisant la log-vraisemblance $p(x|\theta)$ de chaque classe. Cette opération étant en général trop compliquée à faire directement, nous nous servons de l'algorithme couramment utilisé nommé EM (Expectation Maximization).

EM

Une astuce est utilisée dans cette algorithme qui permet de simplifier le problème. Elle consiste à présumer que nous connaissons la composante de la mixture de gaussiennes qui a généré chaque donnée. Autrement dit nous associons à chaque donnée x_i un label y_i pour former un jeu de donnée "complètées" $z_i = (x_i, y_i)$. Cela veut dire que pour une donnée x_i générée par la composante k, nous avons le label $y_i = k$. A partir de là nous n'allons plus chercher les paramètres qui maximisent la (log)vraisemblance des données incomplètes mais celle des données complétées. Plus précisement, nous allons maximiser l'espérance des données complétées connaissant les données et les paramètres de leurs distributions (fonction Q). Ces paramètres sont inconnus au départ, nous les initialisons donc avec un algorithme des k-moyennes (k-means) (en fixant le nombre de composantes des mixtures). Cette maximisation va nous donner de nouveaux paramètres en augmentant la vraisemblance des données. On fonctionne alors de manière itérative, pour arriver à un maximum local.

La fonction Q se définit comme suit :

$$Q(\theta, \theta^{(i-1)}) = E[log(p(x, y|\theta)|x, \theta^{(i-1)})]$$

Ce qui peut se mettre sous la forme :

$$Q(\theta, \theta^{(i-1)}) = E[log(p(x, y|\theta)p(y|x, |\theta^{i-1}))]$$

Le premier terme $log(p(x, y|\theta))$ se calcule de la façon suivante : Les données étant indépendantes et identiquement distribuées par hypothèse,

$$log(p(x, y|\theta) = log(\prod_{i=1}^{N} p(x_i, y_i|\theta)) = \sum_{i=1}^{N} log(p(x_i, y_i|\theta))$$

La formule des probabilités conditionnelles donne :

$$\sum_{i=1}^{N} log(p(x_i, y_i|\theta)) = \sum_{i=1}^{N} log(p(x_i|y_i, \theta)p(y|\theta))$$

Par définition la donnée x_i est générée par la composante de label y_i, dont la distribution de probabilité a pour paramètres θ_{y_i}. Donc $p(x_i|y_i, \theta) = p(x_i|y_i, \theta_i) = p_{y_i}(x_i|y_i)$ ou p_{y_i} est la distribution de la composante de label y_i. De la même façon, $p(y_i|\theta) = p(y_i|\theta_i) = \alpha_i$.

Au final :

$$log(p(z|\theta) = log(p(x, y|\theta) = \sum_{i=1}^{N} log(\alpha_i p_{y_i}(x_i|y_i))$$

Le deuxième terme $log(p(y|x, |\theta^{i-1}))$ se calcule en utilisant la formule de bayes et en initialisant les paramètres avec les k-means.

$$p(y_i|x_i, \theta^{(i_1)}) = \frac{\alpha_{y_i}^{(i-1)} p_{y_i}(x_i|\theta_{y_i}^{(i-1)})}{p(x_i|\theta^{(i-1)})} = \frac{\alpha_{y_i}^{(i-1)} p_{y_i}(x_i|\theta_{y_i}^{(i-1)})}{\sum_{k=1}^{M} \alpha_k^{(i-1)} p_k(x_i|\theta_k^{(i-1)})}$$

La partie E de l'algorithme EM consiste a calculer la fonction Q et la partie M à trouver les paramètres qui maximisent cette fonction :

$$\theta^{(i)} = \arg\max_{\theta} Q(\theta, \theta^{(i-1)})$$

2.1.4 Classification

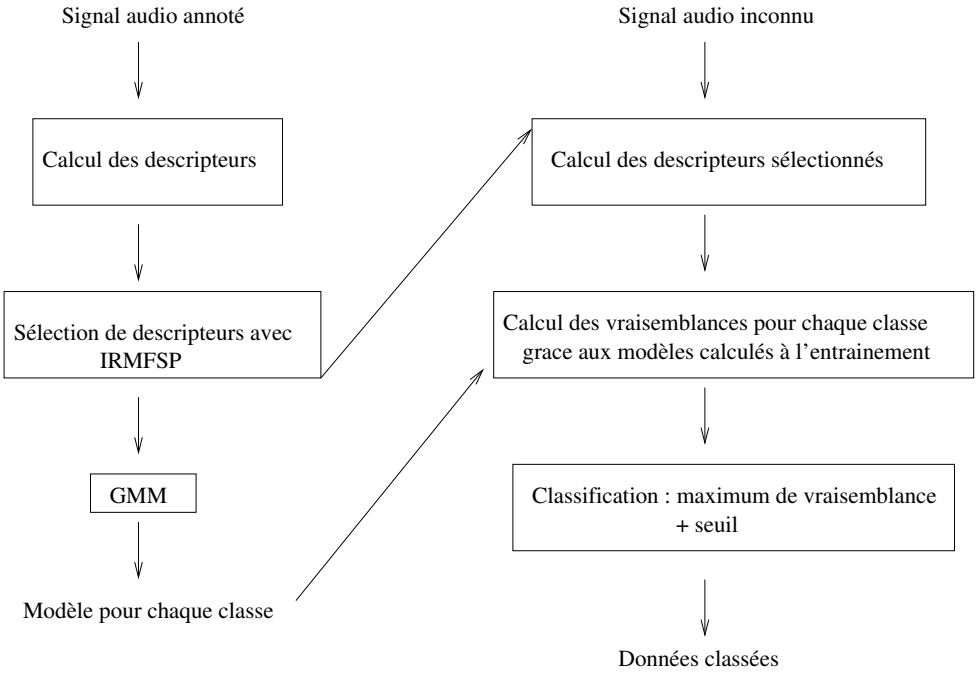
Une fois que les paramètres sont évalués pour chaque classe, la probabilité d'appartenance de nouvelles données peuvent être évaluées.

Il s'agit simplement de calculer la vraisemblance $p(x|\theta)$ des nouvelles données connaissant les paramètres θ des classes. Les données sont classées dans une classe si la vraisemblance de cette donnée est supérieure relativement à cette classe que par rapport aux autres classes. Il est possible également de mettre un seuil de décision. Pour qu'une donnée soit classée, il faut que la vraisemblance la plus élevée soit supérieure d'un seuil fois la deuxième vaisemblance plus élevée. Si ce critère n'est pas rempli, la donnée n'est pas classée. Cela implique qu'une partie partie des données ne sera pas classée, mais que le reste le sera avec plus de précision (voir plus loin pour la définition de la précision).

2.1.5 Lissage

Les résultats de la classification sont ensuite lissés. Un segment contenant 16 trames est défini, à laquel est affecté un label égal au label majoritaire des trames le constituant : si par exemple 12 trames sur 16 sont classées avec le label 1 (classe 1, par exemple la maman), le segment reçoit le label 1.

2.1.6 Schéma général



2.2 Méthode d'évaluation

2.2.1 Validation croisée

Pour évaluer les résultats de classification, il faut pouvoir vérifier si le classifieur fait des erreurs ou non. Nous devons donc utiliser des données annotées. Cela oblige à utiliser uniquement une partie des données pour l'apprentissage, l'entrainement. Les modèles pour chaque classe sont donc construits sur cette partie de la base de donnée. Une autre partie de la base de donnée sert au test. Cela veut dire que nous allons chercher à classer ces données grâce aux modèles que nous avons construit sur les données d'entrainement. Pour assurer une meilleure généralisation, nous utilisons une méthode de validation croisée. La base de donnée est découpée en K blocs, dont K-1 servent pour l'entrainement, le dernier servant pour le test. Les blocs sont ensuite permutés circulairement. Les résultats sont ensuite moyennés sur les K itérations.

2.2.2 Mesures

Deux mesures sont utilisées de façon classique pour évaluer les résultats de la classification : le recall et la précision. Nous utilisons pour notre part la f-measure qui combine ces deux mesures.

Recall

Le recall est le le nombre de données d'une classe bien classés par rapport au nombre de données appartenant à cette classe. Un recall important n'indique pas forcément un bon résultat de classification. En effet on peut atteindre les 100% de recall très facilement en classant bien toute les données de cette classe, plus ceux des autres classes dans celle ci. Cette mesure ne donne pas la proportion de données qui est classée par erreur dans cette classe, il faut une mesure supplémentaire, par exemple la précision.

Précision

La précision est le nombre de données d'une classe bien classées par rapport au nombre de données qui ont été classées dans cette classe (données bien classées et données mal classées).

f-measure

La f-measure se calcule de la façon suivante :

$$F = \frac{2(precision * recall)}{precision + recall}$$

Nous voulons pour notre tâche de classification un maximum de données classées avec un maximum de précision, ce qui correspond à une f-measure la plus importante possible.

2.2.3 Trames et supertrames

Nous avons choisi de montrer les résultats sur les supertrames, les résultats étant meilleurs et le temps de calcul étant moins long. Le label associé à une supertrame est le label majoritaire des trames la constituant. De ce fait, les supertrames comportent pour certaines une partie de leurs données qui n'appartient pas à la classe dont elles ont le label. Par exemple, 19.60% des supertrames de 0.16 secondes contiennent aucune d'autres données que celle appartenant à la classe de leur label. Et 4.72% des trames contenues dans ces supertrames recoivent le mauvais label. Si la taille de la supertrame n'est pas trop élevée comme nous le verrons ensuite, cette erreur sur le label peut être négligée, les résultats sur des supertrames s'avérant meilleurs que ceux sur les trames.

2.3 Résultats

2.3.1 Constitution de la base de donnée

Nous avons constitué une base de données à partir des enregistrements sonores provenant du protocole PILE et disponibles en format DV. Nous avons transféré ces données et nous les avons extrait en wav. Nous avons découpé pour 43 séquences les minutes choisies dans le protocole PILE. 24 de ces minutes ont été annotées en entier. Pour les autres, seules les classes bébé et maman/bébé ont été annotées, ces classes étant en général sous représentées. Enfin 5 séquences entières, d'environ 10 à 15 minutes chacune ont été annotées en bébé et maman/bébé uniquement. Nous avons donc en tout des données provenant de 48 séquences. Nous avons 25 couples mère bébé différents dans ces 48 séquences, plusieurs séquences étant des enregistrements du même couple à des périodes différentes ou à des minutes différentes dans la même séquence.

Nous avons annoté manuellement ces séquences pour en retirer les données nécéssaires à l'entrainement. 4 classes ont été définies : maman s'exprimant seule, maman et bébé s'exprimant de concert, bébé seul et bruit. Les sons ne paraissant pas appartenir à l'interaction entre la mère et le bébé n'ont pas été annotés. Par exemple les respirations ou la toux ne sont pas annotées. De manière générale les sons qui paraissent involontaires n'ont pas été annotés. Les sons ambigus, qui ne peuvent pas être classés de façon certaine dans une des classes ne sont pas annotés. Les sons ou se superposent du bruit aux autres classes ne sont pas annotés si le rapport signal sur bruit est trop faible.

Ces règles ont été définies au fur et à mesure de l'annotation et comportent une part de subjectivité. Il existe donc une incertitude sur la qualité de l'annotation d'une partie (minoritaire) des données.

En l'état la base de donnée est constituée de la manière suivante en terme de durées.

maman	bébé	maman-bébé	bruit
9 mn 22	2 mn 40	5 mn	11 mn 45

La moitié des segments a une durée comprise entre 0.12 secondes et 0.54 secondes. La médiane se situant à 0.25 secondes. Les segments les plus courts ont des durées de 0.01 seconde et les plus longs une durée de 1.2 secondes. Quelques rares segments vont jusqu'à une durée maximale de 10 secondes ou descendent en dessous de 0.02 secondes. Les durées des segments annotés sont donc très courtes si on compare aux durées que l'on peut avoir dans le domaine de la segmentation parole musique (segments d'au moins deux secondes).

2.3.2 Base d'entrainement, base de test

Pour assurer l'indépendance des données pour la validation croisée, nous avons regroupé les séquences avec le même couple mère bébé. La base de test et la base d'entrainement ne peuvent jamais posséder de cette manière un couple en commun même s'ils proviennent de séquences différentes. Par contre ces bases d'entrainement et de test n'ont pas forcément la même taille suivant l'itération, chaque couple mère bébé n'ayant pas le même nombre de données annotées.

2.3.3 Choix de la taille des supertrames

Pour choisir la taille des supertrames, on compare les résultats de classification avec les MFCC en ramenant les résultats à des résultats sur trames. Pour ce faire, nous donnons à chaque trame appartenant à une supertrame le label qui a été donné à la supertrame lors de la classification. Nous avons évalué la f-measure moyenne pour une validation croisée à 8 blocs. Une gaussienne par mixture est utilisée avec une matrice de variance/covariance pleine. Il n'est pas forcément utile et plus long pour cette tâche de comparaison d'utiliser plus de gaussiennes par mixture.

Calcul de la f-measure sur les trames :

maman	bébé	maman-bébé	bruit
0.6848	0.4001	0.4965	0.8572

Calcul de la f-measure sur les supertrames :

Longueur de la supertrame	maman	bébé	maman-bébé	bruit
0.05s	0.6978	0.4080	0.5186	0.8630
0.07s	0.6992	0.4190	0.5255	0.8612
0.09s	0.7023	0.4180	0.5257	0.8623
0.11s	0.7040	0.4128	0.5261	0.8585
0.13s	0.7022	0.4176	0.5336	0.8575
0.15s	0.7039	0.4159	0.5313	0.8576
0.17s	0.7034	0.4200	0.5359	0.8553
0.19s	0.7119	0.4310	0.5389	0.8533
0.21s	0.6964	0.4291	0.5376	0.8447
0.23s	0.7036	0.4370	0.5519	0.8445
0.25s	0.6942	0.4182	0.5493	0.8412
0.27s	0.6965	0.4329	0.5452	0.8378
0.29s	0.6933	0.4347	0.5375	0.8324
0.31s	0.6850	0.4323	0.5353	0.8263
0.33s	0.6858	0.4452	0.5216	0.8260
0.35s	0.6857	0.4194	0.5313	0.8230
0.37s	0.6826	0.4204	0.5442	0.8170
0.39s	0.6717	0.4142	0.5396	0.8078
0.41s	0.6808	0.4157	0.5312	0.8146
0.43s	0.6732	0.4249	0.5262	0.8070
0.45s	0.6684	0.4234	0.5362	0.8058
0.47s	0.6770	0.4170	0.5398	0.8056
0.49s	0.6747	0.4081	0.5140	0.7989
0.75s	0.6295	0.4004	0.5247	0.7554
1.01s	0.6017	0.3700	0.5120	0.7352
2.01s	0.5279	0.2724	0.4643	0.6084

Les résultats avec des supertrames, sont donc bien supérieurs à ceux sur trames. On voit que pour des supertrames de durée comprises entre 0.09s et 0.29s la f-measure pour chaque classe reste assez constante. Au dessus et en dessous, la f-measure décroit pour chacune des classes. Nous voulons prendre des supertrames les plus grandes possibles, pour diminuer la taille de la base de donnée et donc le temps de calcul, qui assurent une bonne classification. Nous prendrons dans la suite des supertrames de 0.19 secondes, celles ci nous paraissant représenter le meilleurs compromis (f-measure assez élevée pour toutes les classes).

2.3.4 Sélection de descripteurs

Nous avons calculé les descripteurs décrits plus haut sur les données, puis leur dérivées premières et secondes. Ce qui donne en tout 135 descripteurs (chaque dimension de chaque "descripteur"comme les MFCC est comptée comme un descripteur). Nous avons ensuite calculé la moyenne et la variance de ces 135 descripteurs sur des supertrames de 0.19 secondes. Cela porte donc en tout le nombre de descripteurs à 270. Les descripteurs sélectionnés avec l'algorithme IRMFSP que nous avons implémenté sont les suivants :

moy(noisiness)
moy(spectral centroid 3)
moy(MFCC 4)
ectyp(MFCC 1)
moy(MFCC 1)
moy(loudness)
moy(MFCC 8)
moy(MFCC 3)
moy(MFCC 5)
moy(spectral flatness 4)
moy(MFCC 9)
ectyp(Delta Delta MFCC 5)
moy(spectral flatness 1)
moy(Spectral Centroid 5)
ectyp(noisiness)
moy(MFCC 24)
moy(MFCC 23)
moy(MFCC 20)
ectyp(Delta spectral crest 2)
moy(MFCC 10)
moy(spectral centroid 1)
moy(MFCC 15)
ectyp(MFCC 3)
ectyp(Delta Delta MFCC 3)
ectyp(Delta harmonic energy)
ectyp(Delta spectral centroid 5)
moy(MFCC 19)
moy(MFCC 17)
moy(MFCC 11)
moy(MFCC 7)
moy(spectral flatness 3)
ectyp(spectral flatness 1)
moy(MFCC 13)
ectyp(Delta MFCC 3)
ectyp(Delta Delta spectral flatness 2)
moy(MFCC 12)

Quand le nom d'un descripteur est suivi d'un nombre N, il s'agit de la dimension N du descripteur. Par exemple ectyp(Delta spectral centroid 5) correspond à l'écart type de la dérivée temporelle première de la cinquième dimension du spectral centroid. Cette dimension, comme nous l'avons décrit plus haut est calculée à partir du spectre de puissance, et la fréquence est en échelle logarithmique.

2.3.5 f-measure

Les f-measure moyennes pour les 4 classes sont les suivantes après classification en validation croisée avec 8 itérations avec les descripteurs sélectionnés au dessus. Nous avons utilisé une gaussienne par classe avec des matrices de variance/covariance pleines. Il est possible d'utiliser plus de gaussiennes par mixture, mais nous n'avons pas encore testé quelle combinaison donne les meilleurs résultats.

maman	maman/bébé	bébé	bruit
73.55%	47.67%	57.47%	87.08%

Pour comparaison, les résultats avec la moyenne des MFCC seuls (et non ramenés à la trame comme plus haut) sur des supertrames de 0.19 secondes sont :

maman	maman/bébé	bébé	bruit
73.42%	44.18%	54.52%	87%

2.3.6 Matrice de confusion

La matrice de confusion est la suivante :

	maman	maman/bébé	bébé	bruit
maman	2198	261	167	362
maman/bébé	185	480	161	9
bébé	237	420	796	122
bruit	369	18	71	3205

La même matrice mais en indiquant le pourcentage de chaque classe indiquée par la ligne classée dans la classe indiquée par la colonne :

	maman	maman/bébé	bébé	bruit
maman	73.6%	8.7%	5.6%	12.1%
maman/bébé	22.2%	57.5%	19.3%	1.1%
bébé	15.0%	26.7	50.5%	7.7%
bruit	10.1%	0.5%	1.9%	87.5%

2.3.7 Seuils

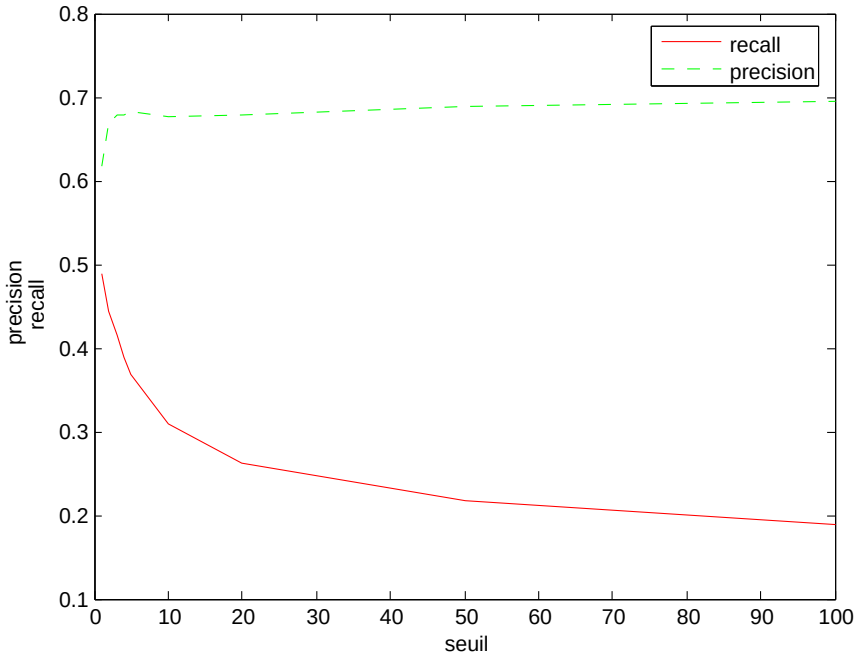
Voici comment se comportent la précision moyenne et le recall moyen au fur et à mesure que l'on augmente le seuil à partir duquel une supertrame est classée (tel que nous l'avons expliqué plus haut dans la partie classification de la section description générale).

Si $pmax$ est la vraisemblance maximale et $pmax_{-1}$ la vraisemblance la plus forte ensuite, nous pouvons définir :

$$RapportVraisemblance = \frac{pmax}{pmax_{-1}}$$

A partir de là nous définissons un seuil : si $RapportVraisemblance < seuil$, alors les données ne sont pas classées.

Nous montrons la précision et le recall moyen pour la classe bébé suivant la valeur du seuil que nous avons fixé.



2.3.8 Discussion

Les f-measure pour la maman et le bruit sont bien plus élevées que pour les classes bébé et maman/bébé. Pour la maman, la matrice de confusion montre que la plus grande part des confusions se fait avec le bruit. Cela reste cependant limité. Pour la classe maman/bébé, les confusions se font, comme on pouvait s'y attendre entre la mère et le bébé essentiellement. Peu de confusions s'opèrent avec le bruit. Pour la classe bébé, une bonne partie des données sont classées comme maman/bébé (420) et une part non négligeable comme maman. La part des bébés qui est classée comme bruit est moins importante. Pour la classe bruit, les confusions se font très majoritairement avec la classe maman, dans une proportion qui reste cependant faible par rapport au nombre de données bien classées.

Il faut noter que le nombre de données entre les classes est inégal. La classe maman en contient 2988. La classe maman/bébé 835. La classe bébé en contient 1575. La classe bruit

contient 3663. Cela correspond respectivement à 9 minutes 22 pour la maman, 2 minutes 40 pour la classe maman/bébé, 5 minutes pour le bébé et 11 minutes 45 pour le bruit. Ce nombre de donnée général est assez faible et la disproportion entre les classes explique en partie la disproportion des taux de reconnaissance suivant les classes.

Nous pensons cependant que des meilleurs taux de reconnaissance peuvent être atteint en utilisant ou construisant des descripteurs mieux adaptés à notre problème. Pour l'instant, les descripteurs qui discriminent le mieux les données sont les MFCC. Les autres descripteurs que nous avons utilisés ne rajoutent que 3% supplémentaires. La visualisation des distributions de ceux ci suivant les classes montrent que dans la plupart des cas elles se recouvrent. Ceci ne préjuge pas cependant de ce que peuvent donner les distributions jointes.

Enfin, les confusions entre classes étant encore trop nombreuses, le seuillage que nous avons mis en place avait pour but d'augmenter la précision de la classification en limitant le nombre de données reconnues. Il apparait cependant que la courbe du recall diminue beaucoup plus vite, et vers des valeurs très faibles .La courbe de la précision atteint elle une limite en dessous de 70% trop faible pour pouvoir classer le (très petit) nombre de données reconnues.

Conclusion et travaux en cours

Nous avons construit un système abouti, mais qui a pour l'instant des résultats trop faibles pour réaliser une classification automatique de toutes les données. Nous pouvons envisager de l'utiliser pour faire de la classification semi automatique, en classant automatiquement les données puis en corrigeant à la main. Nous avons programmé, avec un stagiaire, d'élargir la base de donnée d'entrainement drastiquement, ce qui amènera très probablement à une grosse amélioration des résultats.

Nous étudions en ce moment les signaux mal classés pour voir si nous pouvons intégrer de nouveaux descripteurs ou en construire de nouveaux. Nous avons par exemple constaté qu'une partie des signaux des bébés possèdent une modulation en fréquence. Nous étudions actuellement si un descripteur adapté pourrait améliorer la classification.

Nous avons vu également que plusieurs des descripteurs que nous avons utilisés ont pour base un calcul de f0. Nous pouvons améliorer les résultats en utilisant l'algorithme Yin plutôt que f0.

Par ailleurs nous tentons d'améliorer l'architecture de notre système. Pour l'utilisation des supertrames par exemple, nous pensons qu'un système adaptatif à la durée de chaque segment pourrait avoir de meilleurs résultats. Nous le mettons en place actuellement grâce à un algorithme de détection de transitoires développée dans l'équipe.

Enfin, sur le classifieur lui même, il est possible à terme d'en utiliser un autre plus performant comme les SVM par exemple. Nous avons fait quelques essais en ce sens non concluants pour l'instant sans doute en raison du faible nombre de données. La priorité est donc pour l'instant à l'élargissement de la base de donnée et à l'utilisation ou la construction de nouveaux descripteurs.

Bibliographie

[1] Kevin Bailly, Jocelyne Kiss, Valérie Desjardins, and Bernard Golse. Les productions vocales du bébé : hyperfréquences et processus d'attachement.

[2] Lisa Danielle Bettany. Range exploration of phonation and pitch in the first six months of life. Master's thesis, University of Victoria, 2004.

[3] René Boite, Hervé Bourlard, Thierry Dutoit, Joël Hancq, and Henri Leich. *Traitement de la parole*. Presses polytechniques et universitaires romandes, 2000.

[4] S.F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27 :113–120, 1979.

[5] J.-F. Bonastre, P. Delacourt, C. Fredouille, S. Meignier, T. Merlin, and C. Wellekens. Différentes stratégies pour le suivi du locuteur. In *Reconnaissances des Formes et Intelligence Artificielle, RFIA'2000*, 2000.

[6] Joseph P. Campbell. Speaker recognition : A tutorial. In *Proceedings of the IEEE, Vol. 85, No. 9*, 1997.

[7] M.J. Carey, Parris E.S., and H. Lloyd-Thomas. A comparison of features for speech, music discrimination. In *ICASSP'99*, 1999.

[8] A. de Cheveigné and H. Kawahara. Yin, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.*, 111 :1917–1930, 2002.

[9] Valérie Desjardins. Comodalite, mantelement, enveloppe de mouvements rythmee.

[10] Taco Ekkel. Neural network-based classification of cries from infants suffering from hypoxia-related cns damage. Master's thesis, University of Twente, The Netherlands, 2002.

[11] J.H. Esling, Allison Benner, Lisa Bettany, and Chakir Zeroual. Le controle articulatoire phonétique dans le prébabillage. In *Actes des XXVes Journés d' Etude sur la Parole*, 2004.

[12] John H. Esling, Lisa D. Bettany, Allison Benner, and Rose Spencer. Phonetic structure and acquisition of laryngeal and pharyngeal articulations : Text-analysis considerations. In *CHWP*. CHWP, 2005.

[13] Harriet J. Fell, Linda J. Ferrier, Zehra Mooraj, Etienne Benson, and Dale Schneider. Eva, an early vocalization analyzer. an empirical validity study of computer categorization. In *Proceedings of ASSETS*, 1996.

[14] Harriet J. Fell and J. MacAuslan. Vocalization analysis tools. In *Proceedings of MAVEBA*, 2005.

[15] Harriet J. Fell, J. MacAuslan, C.J. Cress, and L.J. Ferrier. Using early vocalization analysis for visual feedback. In *Proceedings of MAVEBA*, 2003.

[16] Harriet J. Fell, Joel MacAuslan, Karen Chenausky, and Linda J. Ferrier. Automatic babble recognition for early detection of speech related disorders. In *Proceedings of ASSETS*, 1998.

[17] Harriet J. Fell, Joel MacAuslan, Linda J. Ferrier, Susan G. Worst, and Chenausky Karen. Vocalization age as a clinical tool. In *Proceedings of ICSLP*, 2002.

[18] L. A. Gerken. Child phonology. past research, present questions, future directions. In *Handbook of psycholinguistics*. Academic press, 1994.

[19] O. Gillet and G. Richard. Comparing audio and video segmentations for music videos indexing. In *Proceedings of ICASSP 2006*, 2006.

[20] Maya Gratier. Expressive timing and interactional synchrony between mothers and infants : cultural similarities, cultural differences, and the immigration experience. *Cognitive Development*, 18 :533–554, 2003.

[21] Hani Hamdan. *Développement de méthodes de classification pour le contrôle par émission acoustique d'appareils à pression*. PhD thesis, Université de Technologie de Compiègne, 2005.

[22] J. Fell Harriet, Linda J. Ferrier, Carol Espy-Wilson, Susan G. Worst, Eric A. Craft, Karen Chenausky, Joel MacAuslan, and Glenna Hennessey. Analysis of infant babbles. In *Proceedings of ASHA*, 2000.

[23] Nathalie Henrich. *Etude de la source glottique en voix parlée et chantée*. PhD thesis, Uiversité Paris 6, 2001.

[24] Perfecto Herrera-Boyer, Geoffroy Peeters, and Shlomo Dubnov. Automatic classification of musical instrument sounds. *Journal of New Music Research*, 32(1) :3–21, 2003.

[25] Marie Holzer. Projet pile : Localisation en trois dimensions des mains d'un bébé à partir de deux vidéos et réalisation d'une base de donnée pour le projet pile, 2004.

[26] I.S. Howard and M.A. Huckvale. Learning to control an articulatory synthesizer by imitating real speech. *ZAS Papers in Linguistics*, 40 :63–78, 2005.

[27] Hui-Chin Hsu, Alan Fogel, and Rebecca B. Cooper. Infant vocal development during the first 6 months : Speech quality and melodic complexity. *Infant and Child Development*, 9 :1–16, 2000.

[28] F. Koopmans-van Beinum and J. van der Stelt. Early stages in the development of speech movements. In *Precursors of early speech*. NY : stockton press, 1986.

[29] Biing-Hwang Juang Lawrence Rabiner. *Fundamentals of speech recognition*. Prentice Hall PTR, 1993.

[30] Olivier Le Blouch. Méthode de segmentation parole/non-parole, 2005.

[31] D. Lederman, A. Cohen, E. Zmora., K. Wermke, S. Hauschildt, and A. Stellzig-Eisenhauer. Automatic classification of the cry of infants with cleft palate. In *Proceedings of the 2nd European Medical & Biomedical Engineering Conference in Vienna*, 2002.

[32] Dror Lederman. Automatic classification of infant's cry. Master's thesis, Ben-Gurion University of the Negev, Faculty of Engineering, Sciences Department of Electrical and Computer Engineering, 2002.

[33] Stephen N. Malloch. Mothers and infants and communicative musicality. *Musicae Scientae*, Special Issue 1999-2000 :29–57, 1999.

[34] Dan Mateescu. *English phonetics and phonological theory. 20th century approach.* Universitatea din Bucuresti, 2003.

[35] Sirko Molau, Michael Pitz, Ralf Schlüter, and Hermann Ney. Computing mel-frequency cepstral coefficients on the power spectrum. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001.

[36] Suneeti Nathani and Kimbrough Oller. Beyond ba-ba and gu-gu : Challenges and strategies in coding infant vocalizations. *Behavior Research Methods, Instruments, & Computers*, 33(3) :321 – 330, 2001.

[37] D. K. Oller. The emergence of the sounds of speech in infancy. In John E. Bernthal, editor, *Child Phonology*, volume I, chapter 6, pages 93–112. Academic Press, 1980.

[38] D. K. Oller. Metaphonology and infant vocalizations. In *Precursors of early speech.* NY : stockton press, 1986.

[39] Alan V. Oppenheim. Speech analysis-synthesis system based on homomorphic filtering. *The Journal of the Acoustical Society of America*, 45 :458–465, 1968.

[40] Alan V. Oppenheim, Ronald W. Schafer, and John R. Buck. *Discrete-time signal processing (second edition).* Prenctice Hall Signal Processing Series, 1999.

[41] Jose Orozco Garcia and Carlos A. Reyes Garcia. Detecting pathologies from infant cry applying scaled conjugate gradient neural networks. In *ESANN'2003 proceedings*, 2003.

[42] Jose Orozco Garcia and Carlos A. Reyes Garcia. Mel-frequency cepstrum coefficients extraction from infant cry for classification of normal and pathological cry with feed-forward neural networks. In *Proceedings of IJCNN*, 2003.

[43] Sergio D.Cano Ortiz, Daniel I.Escobedo Beceiro, and Taco Ekkel. A radial basis function network oriented for infant cry classification. In *Proceedings of the fifth Congreso de la Sociedad Cubana de Bioingenieria*, 2003.

[44] P. Narcy P. Contentin. Sténoses laryngées de l'enfant. In *Encyclopédie Médico-Chirurgicale*, volume 20 of *A*, oto-rhino-laryngologie 10. Paris, elsevier edition, 1998.

[45] G. Peeters. Instrument sound description in the context of mpeg-7. In *Proceedings of ICMC2000*, 2000.

[46] G. peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. 2004.

[47] Laura Ann Petitto, Siobhan Holowka, Lauren E. Sergio, Bronna Levy, and David J. Ostry. Baby hands that move to the rythm of language : hearing babies acquiring sign languages babble silently on the hands. *Cognition*, 93 :43–73, 2004.

[48] Laura Ann Petitto, Siobhan Holowka, Lauren E. Sergio, and David J. Ostry. Language rhythms in baby hand movements. *Nature*, 413 :35–36, 2001.

[49] Louis Pols. Current developments in phonetics. In *From Sound to Sense*, 2004.

[50] M. Ramona. Approches automatiques pour la segmentation parole/musique. Master's thesis, Master Recherche ATIAM, 2006.

[51] Joseph Razik, Dominique Fohr, Odile Mella, and Nathalie Parlangeau-VallÃ¨s. Segmentation parole/musique pour la transcription automatique. In *Actes des XXVes JournÃ©es d'Etude sur la Parole - JEP'2004, FÃ¨s, Maroc*, 2004.

[52] Orion F. Reyes-Galaviz and Carlos Alberto Reyes-Garcia. A system for the processing of infant cry to recognize pathologies in recently born babies with neural networks. In *Proceedings of SPECOM*, 2004.

[53] Axel Roebel. Transient detection and preservation in the phase vocoder. In *International Computer Music Conference (ICMC). Singapore : Octobre 2003*, 2003.

[54] Harmut Rothgänger. Analysis of the sounds of the child in the first year of age and a comparison to the language. *Early Human Development*, 75 :55–69, 2003.

[55] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *ICASSP '97*, 1997.

[56] R. Stark. Prespeech segmental feature development. In *Precursors of early speech*. NY : stockton press, 1986.

[57] Rachel E. Stark. Stages of speech development in the first year of life. In John E. Bernthal, editor, *Child Phonology*, volume I, chapter 5, pages 73–92. Academic Press, 1980.

[58] Colwyn Trevarthen and Maya Gratier. Voix et musicalité : Nature, emotion, relations et culture.

[59] M. V. Vihman. Individual differences in babbling and early speech : predicting to age three. In *Precursors of early speech*. NY : stockton press, 1986.