Université du Québec
Institut national de la recherche scientifique
Centre Énergie Matériaux Télécommunications

ACOUSTIC AND PROSODIC ANALYSIS OF PRE-VERBAL VOCALIZATIONS OF 18-MONTH OLD TODDLERS WITH AUTISM SPECTRUM DISORDER

By

Stefany Bedoya Jaramillo

Mémoire présenté for the degree of
Master of Science, MSc
in Telecommunications

Evaluation Committee

Internal evaluator and committee president:	Prof. Leszek Szczecinski
External evaluator:	Prof. Patrick Cardinal
Research supervisor:	Prof. Tiago H. Falk
Research co-supervisor:	Prof. Douglas O'Shaughnessy

Acknowledgements

First, I would like to express my sincere gratitude to my advisor Dr. Tiago H. Falk for the continuous support and encouragement during my master program, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I will always be thankful for his help and understanding in this process.

Besides my advisor, I would like to extend my thanks to my director Dr. Douglas O'Shaughnessy and to the members of the evaluation committee, as well INRS-EMT staff and faculty. This research was made possible by funding from the Natural Sciences and Engineering Research Council of Canada (NSERC).

My sincere thanks also goes to the children and their families for participating in the Canadian Infant Sibling Study and the ARU sibs study team.

I thank my friends and lab partners for the encouragement, comments, ideas, discussions and loyal friendship offered during these years.

I would not have been here today if it were not for the love and care of my family. Special thanks to my mother Rubiela and my brother Jeferson. They always give me the more real and pure love. Your love keep me strong!.

"The real voyage of discovery consists not in seeking new landscapes, but in having new eyes." Marcel Proust

Abstract

Autism Spectrum Disorder (ASD) covers a wide spectrum of symptoms with the main ones relating to problems with social communication and interaction. Definite ASD diagnosis is based on the presence of certain symptoms and their severity levels and, according to current standards, occurs typically at 36 months of age. Recent statistics show that about 1 in 68 children are diagnosed with autism and there is a recurrence rate of 18.7% for the biological siblings of autistic individuals. As such, early detection is critical, as it may allow for intense therapy to be initiated, thus tapping into a young brain's plasticity properties and increasing odds of success. Today, researchers and clinicians have joined efforts to understand and identify new markers of the disorders, thus allowing for early diagnosis, ideally around 18 months of age. To this end, acoustic analysis of toddler vocalizations has emerged as a promising area, even for pre-verbal children. Prosodic and acoustic disorders have been reported for babble and speech-like vocalizations. As such, pitch, energy and voice quality related features have been explored for early ASD diagnosis. In this work, we build upon these findings and propose the use of wavelet-based and speech modulation spectral features for ASD diagnosis based not only on speech-like verbalizations, but also on cries, laughs, and other sounds made by the toddlers. We show that the proposed features are complementary to existing ones and, on a cohort of forty-three 18-month old toddlers, a support vector machine classifier was capable of correctly discriminating the ASD group from the typically-developing toddlers with accuracies above 80%, thus outperforming existing methods. More importantly, we show that with these new features, vocalizations such as cries, squeals, whines and shouts showed to be more discriminative than babble and speech-like vocalizations. It is hoped that these findings will lead to more accurate early diagnosis of ASD symptoms.

Index terms— Autism spectrum disorder, diagnosis, prosody, wavelets, speech modulation spectrum

Contents

List of Tables

List of Figures

List of acronyms

AD	Autistic disorder
ADOS	Autism diagnostic observation schedule
AS	Asperger syndrome
ASD	Autism spectrum disorders
AUC	Area under curve
CPP	Cepstral peak prominence
CR	Classification-rate
CV	Coefficient variation
CV	Cross-validation
DD	Developmental delayed
DFT	Discrete fourier transform
DWT	Discrete wavelet transform
ERB	Equivalent rectangular bandwidth
F0	Fundamental frequency
FN	False negatives
FP	False positives
HFA	high-functioning autism
HNR	Harmonic-to-noise ratio
MI	Mutual information
PDD-NOS	Pervasive developmental disorder not otherwise specified

PR	Plurality vote
RBF	Radial basis function
RMSE	root mean squared error
RT	Reaction time
ST	Spectro-temporal
SVM	Support vector machine
SWV	Simple weighted vote
TD	Typically developing
TN	True negatives
TP	True positives
WPD	Wavelet packet decomposition

Synopsis

0.1 Introduction

L'Association Américaine de Psychologie définit l'autisme comme un trouble envahissant du développement qui est lié à une triade de déficiences: (1) développement atypique dans l'interaction sociale réciproque; (2) communication atypique; et (3) des comportements restreints, stéréotypés et répétitifs [2]. En fait, la définition a récemment été mise à jour pour inclure un large éventail de symptômes et de niveaux d'altération, de sorte que la terminologie des troubles du spectre autistique (ASD) a été incorporée [2]. Des statistiques récentes suggèrent qu'environ 1 enfant sur 68 ont un diagnostic d'autisme et il y a un taux de récurrence de 18,7% pour les frères et sœurs biologiques des personnes autistes [3, 4]. Les troubles courants dans le spectre peuvent inclure: le trouble autistique (AD), le syndrome d'Asperger (AS) et le trouble de développement omniprésent non spécifié autrement (PDD-NOS). Le diagnostic définitif (ou stable) de la TSA repose sur la présence de certains symptômes et de leurs niveaux de gravité et se produit habituellement environ à l'âge de 36 mois.

Des recherches récentes, cependant, ont suggéré que le diagnostic peut être accompli vers l'âge de 18 mois [5]. Le diagnostic précoce peut permettre aux parents d'aller de l'avant avec un soutien pédagogique approprié [6] et des cliniciens pour lancer des interventions intenses qui profitent des propriétés de plasticité du jeune cerveau [7].

Cependant, de nouveaux outils sont nécessaires pour diagnostiquer correctement les ASD à un âge précoce. Le programme d'observation diagnostique de l'autisme (ADOS) est l'un des outils d'évaluation standard de l'or utilisés dans le diagnostic des ASD. Récemment, un élément a été ajouté à l'évaluation pour marquer le domaine de communication d'un enfant, ce qui concerne l'existence d'une intonation, d'un volume, d'un rythme et d'un taux de vocalisation particuliers. Il a été inclus dans l'évaluation suite à l'observation clinique largement notée selon laquelle de nombreux enfants sur le spectre autistique présentent une prosodie atypique de vocalisations et de verbalisations [8, 9, 10, 11].

Ces études ont montré que les vocalisations des personnes autistes sont plus difficiles à reconnaître en termes de pairs affectifs et retardés au développement (DD) [12, 13]. De plus, les observations montrent que les individus sur le spectre avaient une qualité de voix plus rousse, plus sévère, plus hyper-nasale, avec une récidive plus élevée des grinçons, des croûtes et des hurlements [14]. Plusieurs termes ont été utilisés pour décrire la prosodie dans les ASD et comprennent: monotone, exagérée, robotique, pédante et raide, pour n'en nommer que quelques-uns. Quel que soit le résultat, il est souvent remarqué par les contacts sociaux et peut représenter un obstacle important à l'acceptation par les pairs. Les atypicités dans la prosodie des sujets atteints de ASD ont été notées dans la première description de l'autisme et les études montrent que les déficiences du comportement vocal sont évidentes, même chez les enfants autistes préverbaux et peuvent représenter une caractéristique précoce de l'ASD [15, 16, 17, 12, 18].

Les premières études ont examiné la perception et la production de la prosodie chez les personnes atteintes d'autisme, mais un problème souvent abordé dans la littérature de l'ASD est la prévalence inconnue de la prosodie atypique et l'hétérogénéité de la population ASD. De plus, l'absence des méthodes d'enquête standardisées, qui sont utilisées pour quantifier la prosodie vocale et les lacunes du jugement perceptif, ont donné des résultats incohérents et parfois contradictoires [14, 19, 20]. Des recherches récentes se sont intéressées à l'intégration de l'analyse acoustique informatisée dans le

diagnostic de la ASD, ce qui a permis de surmonter certaines de ces limitations impliquées dans les études qualitatives [9, 10, 11]. Les caractéristiques acoustiques de la production de la parole ont été mesurées et quantifiées dans plusieurs études en tant que marqueurs possibles d'ASD. La plupart des études existantes, cependant, ont été menées sur des enfants plus âgés ou avec des enfants plus jeunes, mais avec une vaste tranche d'âge entre 18 et 36 mois. Compte tenu des changements naturels des caractéristiques vocales (par exemple, la croissance et la maturation des voies vocales) qui se produisent pendant l'enfance, il est important que l'analyse acoustique se concentre sur des individus d'environ le même âge, éliminant ainsi le biais potentiel des changements acoustiques liés à l'âge naturel et et de la variabilité et permettre au système de se concentrer uniquement sur les différences liées au désordre. Le travail décrit ici comble cet écart.

0.1.1 Prosodie dans l'ASD

La prosodie désordonnée a longtemps été considérée comme une caractéristique de l'ASD, et l'intonation atypique a été reconnue comme un graveur important du diagnostic d'ASD. La caractérisation prosodique des enfants atteints d'ASD est une zone sous-traitée, en particulier pour les enfants très jeunes et préverbaux, bien que les études suggèrent que l'atypicité vocale peut représenter un symptôme apparaissant tôt chez les ASD. Les études ont traditionnellement mesuré et quantifié des paramètres d'intonation, de volume et de qualité vocale afin d'identifier les différences entre l'ASD et les groupes de comparaison. La découverte la plus persistante dans la littérature est que les individus atteints de ASD présentent une plus grande variabilité et une gamme plus large de mesures de fréquence fondamentale (F0) [21, 22, 23, 14, 9, 24, 25, 26, 18]. Cette constatation est dans une mesure contraire à l'intuition car les enfants sur le spectre autistique sont souvent décrits comme ayant un discours monotone ou robotique. Peu d'études ont également suggéré que les altérations dans l'utilisation des caractéris-

tiques liées au ton dans les individus autistes sont liées à un traitement anormal des stimuli auditifs au niveau du tronc cérébral [27, 28, 29]. Plus récemment, un coefficient plus élevé de variation de hauteur (CV) a été trouvé pour le groupe TD par rapport au groupe ASD à l'âge scolaire, alors qu'aucune différence significative entre les deux classes n'a pu être observée à l'âge préscolaire [11].

En plus du ton désordonné, on a déclaré que les personnes atteintes d'ASD avaient un discours qualifié de «trop rapide» ou «trop lent» [14]. La plupart des études qui ont mesuré la durée des phrases et des mots ont signalé une durée plus longue chez les personnes atteintes d'ASD [14, 30, 10, 31, 32]. Dans [10] par exemple, une analyse perceptuelle et acoustique chez les enfants atteints du syndrome d'Asperger a été employée. L'étude a montré que même lorsque les enfants atteints d'ASD et de TD se comportaient de manière similaire à la perception des modèles d'intonation et à l'utilisation de caractéristiques prosodiques pour exprimer une signification grammaticale, les enfants atteints d'ASD présentaient une altération de la durée, de la hauteur moyenne et maximale. Ces résultats sont conformes aux résultats présentés dans [32], où une différence significative de durée pour le discours émotionnel (énoncés tristes ayant une durée plus longue que les énoncés heureux ou en colère) a été trouvée pour les groupes TD et AS, tandis que les sujets ayant un autisme élevé fonctionnellement (HFA) n'a montré aucune différence. Cependant, deux études récentes n'ont pas révélé de différences significatives de la durée entre les enfants atteints de ASD et TD [21, 33]. Encore une fois, de tels résultats contradictoires peuvent être attribuables à une vaste gamme d'âge des participants à l'étude, ce qui peut entraîner une baisse des tendances en raison des changements acoustiques et de la variabilité liés à l'âge naturel.

De plus, quelques études ont utilisé une analyse multi variée à l'aide de techniques d'apprentissage par machine. Dans [34], par exemple, quatre groupes de caractéristiques: qualité de la voix, caractéristiques liées à l'énergie, spectrale, et cépstrale ont été comparés pour une base de données recueillie pour évaluer les capacités des enfants à l'imitation des différents types de contours de prosodie. Les résultats ont montré

que la qualité de la voix améliorait les performances de la classification sur les autres groupes de fonctionnalités. D'autres études ont fait des mesures liées à la qualité de la voix telles que la gigue, le rapport harmonique-bruit (HNR) et la prévalence du pic cépstral (CPP), qui ont tendance à augmenter chez les enfants atteints d'autisme avec une augmentation de la gravité d'ASD [35, 9, 25, 24].

0.1.2 Modèles vocaux primaires en ASD

Les mères d'enfants atteints d'ASD peuvent avoir de la difficulté à reconnaître la signification affective de la vocalisation de leurs nourrissons, et des preuves récentes apparaissent pour indiquer que l'atypicité vocale peut apparaître chez des jeunes très jeunes atteints d'ASD [12, 16, 36, 37, 13]. Cependant, la plupart des études ont eu tendance à se concentrer sur les adolescents et les adultes verbaux, bien que certains aient également étudié des enfants d'âge scolaire et d'âge préscolaire. Les vocalisations infantiles sont la première forme de communication vocale. Ils jouent un rôle important dans le développement de la relation parent-enfant et de l'acquisition. Les nourrissons commencent à produire des sons végétatifs ou réfléchis, comme tousser ou pleurer après la naissance, et à travers plusieurs étapes, les enfants développent leurs sons et leurs vocalisations pour devenir plus parlants [38]. La croissance et la restructuration anatomique du tractus vocal pendant la première moitié de la vie et l'apprentissage vocal sont les principaux facteurs qui induisent un changement dans les vocalisations infantiles [38]. Récemment, peu d'études ont analysé le développement de paramètres acoustiques dans les vocalisations de l'enfant comme outil non invahissant pour mesurer la maturation musculaire-vocale [39, 12, 16, 40]. Les atypicités dans la production de modèles vocaux pourraient impliquer un traitement anormal des commentaires auditifs ou des problèmes dans les mécanismes de production de la parole. Dans les prochaines sous-sections, nous explorons certaines études employées chez les enfants autistes préverbaux. En parti-

culier, les vocalisations de babillages canoniques et de pleurs ont été étudiées dans la littérature ASD.

Les pleurs

Pleurer est la première forme de communication vocale du nourrisson. Il a été exploré en raison de leur relation avec le système nerveux central [41, 12, 16, 42, 12, 17, 43, 44, 45, 37]. Les anomalies acoustiques dans les cris du nourrisson ont été associées à certains troubles tels que l'asphyxie, le faible poids à la naissance, les troubles métaboliques, les symptômes neurologiques et l'exposition au plomb, parmi d'autres. La plupart d'entre eux prouvent un pas élevé et variable [41]. Dans la littérature d'ASD, les cris des enfants plus tard diagnostiqués avec de l'autisme ont affiché une valeur F0 plus élevée et des pauses plus courtes que les cris des enfants retardés ou généralement développés au développement [12, 16, 36, 37]. De plus, la fréquence fondamentale a diminué pour les enfants en bonne santé pendant la première et la deuxième année de vie, contrairement au cas où les enfants ont plus tard été diagnostiqués avec ASD [36, 16]. Sheinkopf et les collèges ont utilisé une analyse acoustique grise des nourrissons de 6 mois [46] et ont montré qu'à 6 mois, les enfants à risque élevé ont commencé à montrer une valeur de hauteur plus élevée et variable que ceux à faible risque. Plus tard, lorsque les mêmes sujets à risque élevé ont été analysés à l'âge de 36 mois, ils ont montré un F0 encore plus élevé. De plus, d'autres études ont analysé la variabilité de la fréquence fondamentale (plage de hauteur), mais aucune différence n'a été trouvée entre les deux groupes [37, 46].

Une étude plus récente dans l'ASD a utilisé une tâche catégorielle du temps de réaction (RT) pour analyser les réponses des adultes des cris d'enfants âgés de 36 à 40 mois atteints d'ASD [16]. Ils ont constaté que les différences de comportement vocal chez les enfants plus tard diagnostiqués avec ASD ont amené les adultes à percevoir les cris comme affligés et plus difficiles à traiter et à interpréter que les cris des enfants

typiquement en développement, ainsi que des cris d'animaux de mammifères et des sons de contrôle du bruit environnemental. Esposito et ses collègues, à leur tour, ont examiné les caractéristiques acoustiques des vocalisations de criblage infantile d'un groupe d'enfants et d'enfants atteints d'ASD généralement développés, âgés de 13 mois [12], et ont constaté que la durée de pause était plus importante pour la perception de la détresse dans Les enfants atteints d'ASD que la fréquence fondamentale ou la fréquence des sons de cri par unité de temps à travers un épisode de pleurs.

Babillage canonique

Le babillage commence peu après la naissance et est bien établi à 10 mois. Tout retard dans l'apparition du babillage canonique est lié au retard de la langue ou à d'autres troubles du développement [17]. Quelques études ont porté sur l'analyse des balbutiements canoniques chez les enfants atteints d'ASD [17, 43, 44]. Patten et ses collègues ont étudié le statut canonique et la fréquence de vocalisation d'un groupe de 37 nourrissons atteints d'ASD par rapport à un groupe de développement typique à 9-12 et 15-18 mois [17]. Les personnes ayant diagnostiqué avec l'autisme plus tard ont produit des taux significativement plus faibles et ont eu un début de balbutiement canonique plus tard que le groupe témoin. Ces résultats sont cohérents avec les résultats obtenus dans l'analyse de la production de syllabes canoniques chez les nourrissons âgés de 16 à 48 mois [44]. Cependant, certains résultats signalés ont été contradictoires. Par exemple, Sheinkopf et ses collègues ont étudié la nature des comportements vocaux précoces chez les jeunes enfants atteints d'autisme axés sur le balbutiement canonique et la qualité vocale atypique (définie comme le taux de production de la phonation atypique) [43] et le groupe avec l'ASD n'a pas montré des différences significatives par rapport au groupe témoin retardé par le développement en termes de taux de balbutiement canonique, malgré leurs vocalisations montrant une qualité vocale atypique [17, 44].

0.1.3 Objectifs et contributions de la thèse

Il est de plus en plus évident que les enfants atteints d'ASD ont une trajectoire de développement retardée de la parole et que les marqueurs précoces peuvent être présents au cours des premières années de vie. Puisque l'identification des différences acoustiques dans les vocalisations peut aider au développement d'interventions antérieures ciblées, plus efficaces, l'objectif de cette thèse est de développer un système automatisé pour distinguer entre les groupes ASD et non-ASD. À cette fin, une analyse acoustique des vocalisations préverbales de 43 enfants (23 avec ASD, 20 témoins) de 18 mois est réalisée. Nous étendons les travaux précédents dans la détection automatisée de l'autisme de trois façons.

Tout d'abord, nous analysons les contributions de différents types de vocalisations (par exemple, le cri, la parole, le babillage, le rire et le grognement) pour la tâche à accomplir. Compte tenu des différentes tendances signalées pour différents types de vocalisation, on s'attend à ce qu'elles contribuent différemment pour la performance globale du diagnostic des ASD. Cette connaissance n'a pas encore été signalée dans la littérature. Deuxièmement, nous proposons l'utilisation de deux nouveaux ensembles de fonctionnalités, à savoir des fonctions à base de wavelet et de modulation de la parole pour le diagnostic ASD. Les caractéristiques des ondelettes ont été explorées dans le passé pour l'analyse des pleurs des nourrissons (p. Ex., [47, 48])). On s'attend à ce qu'elles caractérisent avec précision les vocalisations tels que les cris, les hurlements, etc. Les caractéristiques spectrales de la modulation de la parole, à leur tour, ont été largement utilisées pour la caractérisation de l'émotion de la parole [49] et pour l'analyse de la voix pathologique (p. Ex., [50, 51]). Nous explorons leurs avantages individuellement, ainsi que combinés avec les fonctionnalités prosodiques existantes. Au mieux de notre connaissance, c'est la première tentative de combiner les caractéristiques spectrales des ondelettes et de la modulation dans le but de détecter l'autisme. Troisièmement, en utilisant les données d'une évaluation de routine de 18 mois, les résultats présentés ici

sont basés sur les données obtenues auprès de participants ayant la tranche d'âge la plus basse et la plus sévère. Étant donné que les vocalités des enfants sont connues pour changer continuellement pendant l'enfance, compte tenu de la maturation de leurs voies vocales, une telle gamme d'âge réduit les biais potentiels et permet aux classificateurs de se concentrer sur les caractéristiques d'ASD discriminatives existantes.

0.2 Méthodes et matériaux

0.2.1 Collection de données

Les données ont été tirées de l'étude Canadienne prospective future«Infant Sibling Study» en cours à l'unité de recherche sur l'autisme de l'hôpital SickKids de Toronto [52]. Les participants étaient 23 (15 masculins / 8 féminins) frères et sœurs plus jeunes de probands atteints d'ASD, recrutés en raison d'un risque plus élevé connu pour exposer le trouble [3]. Les participants ont été suivis tous les 3 à 6 mois dans les premiers 24 mois, en recherchant des signes précoces d'ASD et ils ont été diagnostiqués de façon indépendante avec l'ASD à l'âge de 36 mois en utilisant le programme d'observation de diagnostic d'autisme (ADOS) et Autism Diagnostic Interview- Révisé (ADI-R). Un groupe de comparaison correspondant à un âge de 20 (13 masculin / 7 féminin) sera à partir du groupe témoin de l'étude plus vaste. Ils ont reçu les mêmes évaluations de suivi et ont été déterminés à 36 mois d'âge pour ne pas avoir d'ASD. Toutes les vocalisations des tout-petits ont été segmentées et extraites des évaluations ADOS enregistrées par vidéo menées à l'âge de 18 mois. La Table 2.2 résume la quantité d'énoncés de vocalisation et le type de vocalisation pour chaque groupe. La classe intitulée «émotions négatives» combine des énoncés de vocalisation tels que le cri et le gémissement. Comme on l'a vu, le nombre de vocalisations par groupe et par type de vocalisation n'est pas équilibré avec le babillage et le discours étant le plus important et le moins de rire .

0.2.2 Extraction des caractéristiques

Décomposition des paquets d'ondelettes (WPD)

La décomposition des paquets d'ondelettes (WPD) est une méthode de généralisation de la Transformée d'ondelettes discrètes (DWT) qui permet une analyse multi-résolution temps-fréquence d'un signal d'entrée. Contrairement au DWT, les bandes à faible et haute fréquence sont utilisées pour une nouvelle décomposition dans WPD. Le processus de décomposition et l'analyse multi-résolution peuvent être considérés comme l'application d'une banque de filtres. Plus précisément, le signal d'entrée passe à travers un filtre passe-bas et passe haut, qui correspond à la fonction d'échelle et d'ondelettes [53, 54]. La bande de fréquence inférieure donne les coefficients d'approximation et la bande de fréquence supérieure des coefficients de détail. La Figure 2.1 montre un exemple de décomposition de paquets d'ondelettes à deux niveaux pour un signal de babillage généré à partir d'un contrôleur et d'un participant ASD. Pour les niveaux de décomposition de n, le WPD produit 2^n sous-bandes ou nœuds de largeur de fréquence égale . Les fréquences indiquées dans Hertz pour le niveau n de décomposition sont décrites par:

$$\left[\frac{kfs}{2^{n+1}} \frac{(k+1)fs}{2^{n+1}}\right] \quad k = 0, 1, ..2^{n-1}, \tag{1}$$

Où f_s est la fréquence d'échantillonnage du signal original [47].

Fonctionnalités des ondelettes

Étant donné le WPD décrit ci-dessus, nous proposons l'utilisation des caractéristiques basées sur l'énergie et l'entropie pour notre analyse, calculées à partir de chaque coefficient de paquet d'ondelettes $C^P_{n,k}$ comme suit [53, 47, 55], respectivement:

$$E_{(n,k,l)} = \sum_{l=-\infty}^{+\infty} |C^P_{n,k}(m)|^2 w(l-m), \tag{2}$$

$$S_{(n,k,l)} = \sum_{l=-\infty}^{+\infty} -|C_{n,k}^{P}(m)|^2 \log |C_{n,k}^{P}(m)|^2 w(l-m), \tag{3}$$

Où l est le nombre des largeurs des fenêtres, P est l'indice de l'échelle, n représente le niveau de décomposition et $k = 0, 1, ..., 2^{n-1}$ est le numéro du noeud . Dans nos expériences, chaque coefficient de paquet d'ondelettes sous-bandes est divisé en trames de 40 ms et les images successives ont été chevauchées de 50%. Enfin, des mesures statistiques tels que la moyenne ($\bar{X}_{(n,k)}$), écart-type ($std_{(n,k)}$), asymétrie ($g_{(n,k)}$), et le kurtosis ($G_{(n,k)}$) sont calculés sur les mesures globales par image sur l'énoncé complet de la vocalisation. Le vecteur de la caractéristique finale est un vecteur de caractéristique 8-dimensionnel calculé par noeud k et par niveau de décomposition i, composé de la moyenne, de l'écart type, de l'asymétrie et du kurtosis de l'énergie et des valeurs d'entropie. La Figure 2.2 représente un schéma fonctionnel du schéma d'extraction de la fonctionnalité proposé.

Représentation spectrale de la modulation de la parole

La Figure 2.3 montre les étapes utilisées dans notre approche pour calculer la représentation spectro-temporelle (ST) d'un signal d'entrée [49]. Dans la première étape, le niveau du signal actif est normalisé à -26dBov (surcharge dB) [56]. Ensuite, une banque de 23 filtres gammatone à bande critique est utilisée pour modéliser la réponse en fréquence de la membrane basilaire. La banque de filtres est conçue comme une série de filtres passe-bande indépendants inspirés du modèle cochléaire de Patterson [57]. Le premier filtre est centré à 125 Hz et le dernier à la moitié de la fréquence d'échantillonnage du signal analysé; Les bandes passantes des filtres sont caractérisées par la bande passante rectangulaire équivalente (ERB).

L'enveloppe temporelle $e_i(n)$ est calculée pour chacune des sorties de la banque de filtres$\hat{s}_i(n)$ à l'aide de la transformation Hilbert $\mathcal{H}\{\cdot\}$. L'enveloppe du filtre passe-bande

i est donnée par:

$$e_i(n) = \sqrt{\hat{s}_i(n)^2 + \mathcal{H}\{\hat{s}_i(n)\}^2} \qquad i = 1 \cdots, 23 \tag{4}$$

Le spectre de modulation du signal est calculé à l'aide de la transformée de Fourier discrète (DFT). Plus précisément, le signal d'enveloppe $e_i(n)$ est divisé en trames de 256 ms tous les 40 ms à l'aide d'une fenêtre Hamming. La notation $e_i(m)$ est utilisée pour indiquer le cadre m de l'enveloppe fenêtre. L'étape suivante consiste à prendre le DTF $\mathcal{F}\{\cdot\}$ de chaque image. Ensuite, le spectre de modulation est défini par

$$E_i(m, f) = |\mathcal{F}(e_i(m))|, \tag{5}$$

Où f est la fréquence de modulation. Enfin, une banque de filtres de modulation inspirée de l'audition nous permet de construire une représentation en fonction de la fréquence acoustique et des éléments de fréquence de modulation temporelle. L'énergie de modulation du signal de la bande critique i est regroupée en 8 bandes, chaque bande est désignée par $\varepsilon_{i,k}(m), k = 1, \ldots, 8$, où k est le filtre de modulation k.

Caractéristiques spectrales de la modulation

Compte tenu de la représentation spectrale de la modulation ci-dessus, l'ensemble des caractéristiques proposées initialement dans [1] est extrait. Le premier ensemble de fonctionnalités, $\Phi_{1,m}(k)$, représente la distribution d'énergie le long de la fréquence de modulation. Il est défini comme la moyenne des échantillons d'énergie par rapport à la chaîne de modulation k:

$$\Phi_{1,m}(k) = \frac{\sum_{i=1}^{N} \varepsilon_m(i,k)}{N}. \tag{6}$$

Le deuxième ensemble, $\Phi_{2,m}(k)$, est défini comme le rapport de la moyenne géométrique d'une mesure d'énergie spectrale et de sa valeur moyenne arithmétique, représentant ainsi la planéité spectrale du spectre. Une valeur de planéité spectrale proche de 1 est liée à un spectre plat, tandis qu'une valeur proche de 0 suggère un spectre avec des variations élevées de son amplitude spectrale. Cette mesure est calculée comme suit:

$$\Phi_{2,m}(k) = \frac{\sqrt[N]{\Pi_{i=1}^{N} \varepsilon_m(i,k)}}{\Phi_{1,m}(k)}. \tag{7}$$

La troisième mesure$\Phi_{3,m}(k)$ correspond au centre de masse de chaque canal de modulation, où $f(i)$ est l'indice de la bande critique. Le centroïde spectral pour le canal de modulation i est donné par:

$$\Phi_{3,m}(k) = \frac{\sum_{i=1}^{N} f(i)\varepsilon_m(i,k)}{\varepsilon_m(i,k)}. \tag{8}$$

Afin de mesurer la relation de différents canaux de modulation, les 23 canaux acoustiques sont regroupés en cinq niveaux: $D_1 = [1-4]$, $D_2 = [5-8]$, $D_3 = [9-12]$, $D_4 = [13-18]$ et $D_5 = [19-23]$. Les canaux de modulation dans chaque catégorie sont additionnés et utilisés pour calculer le centroïde spectral $\Phi_{4,m}(k)$ dans le domaine de fréquence de modulation pour D_l comme suit:

$$E_m(l,k) = \sum_{i \in D_l} \varepsilon_m(i,k), \tag{9}$$

$$\Phi_{4,m}(k) = \frac{\sum_{k=1}^{8} k\mathsf{E}_m(l,k)}{\sum_{k=1}^{8} \mathsf{E}_m(l,k)}. \tag{10}$$

Les deux dernières mesures captent le taux de variation de chaque région de fréquence acoustique, fournissant ainsi une indication de la dynamique temporelle des énoncés. Le coefficient de régression linéaire $\Phi_{5,m}(k)$ (pente) et l'erreur de régression correspondante $\Phi_{6,m}(k)$ (erreur moyenne quadratique, RMSE) sont calculés. Ces mesures sont associées

au modèle polynomial de premier degré utilisé pour s'adapter à $\mathsf{E}_m(l,k)$. Le vecteur de fonctionnalité finale comprend 184 fonctionnalités d'énergie de spectre de modulation $\varepsilon_{i,k}(m), k = 1, \ldots, 8$ plus les 39 fonctionnalités décrites ci-dessus, ce qui représente 223 fonctionnalités.

Mesures acousto-prosodiques

Les caractéristiques acoustiques prosodiques et leurs variations entre les groupes ont été les caractéristiques les plus utilisées dans l'analyse des troubles du spectre autistique, comme mentionné dans le Chapitre 0.1. Ici, les énoncés de vocalisation ont été analysés acoustiquement à l'aide de la boîte à outils VoiceSauce MATLAB du laboratoire UCLA SPAP. Dans nos expériences, les deux mesures sont optimisées pour la vocalisation des enfants. Les paramètres liés au ton et aux formants ont été calculés en utilisant une gamme de fréquence fondamentale (F0) comprise entre 60-1600 Hz et une fréquence nominale F1 de 1250 Hz, ce qui correspond à la fréquence nominale d'un tractus vocal de 7 cm. Les caractéristiques ont été extraites à partir de cadres de 25 ms toutes les 10 ms.

Au total, 26 paramètres acoustiques sont extraits, comme indiqué dans le Tableau 2.4. Afin d'explorer les variations des caractéristiques prosodiques entre ASD et les groupes de contrôle, trois combinaisons de fonctionnalités prosodiques différentes sont proposées. Le premier groupe (PF1) comprend la valeur moyenne des caractéristiques signalées dans le Tableau 2.4 pour chaque énoncé de vocalisation. Le deuxième groupe (PF2) a combiné la moyenne de distribution et l'écart-type, et finalement, la moyenne, l'écart-type et la portée sont inclus dans le troisième groupe (PF3). Une telle partition a été utile dans [35].

0.2.3 Design du classificateur

Dans nos expériences, une machine de vecteur de support (SVM) est utilisée pour classer entre ASD et les classes de contrôle. Trois SVMs sont formés séparément sur les énoncés de vocalisation des trois différents groupes de caractéristiques, à savoir les ondelettes, la modulation spectrale et l'acoustique-prosodique. La mise en œuvre de SVM dans [58] a été adoptée et un noyau RBF a été choisi car il a entraîné une performance améliorée au cours de nos expériences pilotes. Les paramètres de conception des SMVs sont sélectionnés à l'aide d'une méthodologie de recherche par quadrillage 4 fois. Afin d'évaluer la performance du système de classification des SVMs, une validation croisée à 10 fois est adoptée en utilisant les hyper paramètres trouvés avec le processus décrit ci-dessus et la performance est calculée par participant. Avec cette approche, un nourrisson est étiqueté comme témoin ou ASD en utilisant un schéma basé sur le score des décisions prises par le SVM. Cette méthode a été choisie de manière empirique en raison de sa performance supérieure par rapport à la méthode commune de vote partagé. Plus précisément, les résultats SVMs des énoncés de vocalisation pour chaque sujet sont comparés et la vocalisation avec le score de vraisemblance le plus élevé décide de la prédiction de la classe finale. Ainsi, si $c(x_i)$ correspond au score de prévision pour l'échantillon x_i, le score final de prédiction peut être calculé comme suit:

$$C = \arg\max[c(x_1) \cdots c(x_i)], \tag{11}$$

Où $c(x_i)$ correspond à la distance de l'échantillon x_i à l'hyperplan séparateur.

0.2.4 Plans de fusion

Fusion au niveau de la décision

Dans nos expériences, tous les échantillons x_i appartenant au sujet s_l des différents classificateurs sont utilisés pour le problème de combinaison afin de faire un diagnostic par sujet. Trois fonctions différentes combinées sont proposées:

1. Le vote sur la pluralité (PV): c'est la méthode de fusion la plus simple et la plus commune. Le sujet s_l est attribué à la classe c_j qui a obtenu le plus grand nombre de votes. Dans ce cas, tous les poids du classificateur sont égaux, c'est-à-dire $w_k = 1/K \ \forall K$.

2. Votes de probabilité maximum (MPV): Dans ce schéma de fusion, les probabilités pour les échantillons x_i appartenant à la matière s_l sont comparées et l'échantillon avec la probabilité la plus élevée décide de la prédiction finale:

$$C = \arg\max[p(x_1) \cdots p(x_i)]. \tag{12}$$

3. Votes de probabilité moyenne (APV): Les probabilités conditionnelles par échantillon pour chaque classe sont en moyenne et la classe c_j qui a obtenu la probabilité moyenne la plus élevée décide de la prédiction finale. Ainsi, si $p(c_j/x_i)$ est la probabilité conditionnelle que x_i soit au dessous de la classe c_j, la prédiction finale est faite comme suit:

$$C = \arg\max\left[\frac{\sum_i p(c_1/x_i)}{i}, \frac{\sum_i p(c_2/x_i)}{i}\right]. \tag{13}$$

Dans nos expériences, les classificateurs dans les ensembles sont comparables dans le sens où ils ont été formés sur les mêmes ensembles de données et en utilisant le même partitionnement.

Fusion au niveau des fonctionnalités

Dans la fusion au niveau des fonctionnalités, les ensembles des caractéristiques provenant de différentes sources sont concaténés en un seul vecteur de fonctionnalité avant le processus de classification. Le principal avantage de cette méthode est que les fonctionnalités corrélées dans et entre différents ensembles de caractéristiques peuvent être supprimées par des outils de réduction de la dimensionnalité, améliorant ainsi la générateur du système. Dans nos expériences, un algorithme basé sur l'information mutuelle (MI) a été utilisé afin de mesurer le degré de parenté entre les valeurs de la caractéristique [59, 60]. L'IM des caractéristiques définies est calculé pour l'algorithme en utilisant la méthode de voisinage le plus proche. Les détails de la méthode sont présentés dans [59].

0.3 Expériences

0.3.1 Expérience 1: Sélection de la mère Wavelet

Notre premier objectif est d'étudier et de comparer l'efficacité des différents types d'ondelettes de la mère et la répartition de l'information dans plusieurs niveaux de décomposition pour la discrimination des troubles du spectre autistique. Pour ce faire, la méthodologie de fonctionnalité d'extraction proposée dans la Section 2.4.1 est employée à l'aide de différentes familles d'ondelettes, tels que daubechies (db), coiflet (coif), symétrie (sym), biorthogonal (bior) , et biorthogonal reversé(rb). Les caractéristiques d'énergie et d'entropie sont extraites du coefficient de paquet d'ondelettes à plusieurs niveaux de décomposition et comparées entre elles pour chaque type d'ondelette mère. Un résumé des meilleurs résultats de performance est présenté dans le Tableau 3.6. La précision de la reconnaissance maximale de 81,5% a été obtenue à l'aide d'une décomposition de premier niveau et d'une ondelette mère de 18ème ordre. De plus, il a été

constaté que l'augmentation du niveau de décomposition des ondelettes améliorait la performance générale de la plupart des ondelettes de la mère.

0.3.2 Expérience 2: Comparaisons des ensembles de fonctionnalités

Les résultats expérimentaux pour les différents ensembles de caractéristiques proposés dans le Chapitre 2 sont rapportés dans le Tableau 3.7. Comme on peut le voir, les caractéristiques d'ondelettes proposées ont atteint la meilleure précision, atteignant un taux de reconnaissance moyen de 81,5%. Ils sont suivis par les caractéristiques spectrales de modulation avec une performance de 79,0%. Les trois ensembles de caractéristiques prosodiques de référence ont réalisé des performances similaires avec PF1 obtenant la meilleure performance globale parmi le PF1-PF3 (voir la Section 2.4.3 pour plus de détails).

Les taux de la reconnaissance moyenne dans le Tableau 3.7 sont signalés par participant, comme cela est décrit dans la Section 2.5.2. Au total, 43 enfants (23 avec ASD, 20 témoins) ont été diagnostiqués pour chaque modèle de classification séparément. La vocalisation avec le score le plus élevé par sujet a fait le diagnostic final. Le Tableau 3.8 et la Table 3.9 présentent les détails du diagnostic effectué entre les classes de contrôle et l'ASD pour chaque groupe de fonctionnalités. Plus précisément, le Tableau 3.8 indique le nombre d'enfants correctement classés (vrais positifs et vrais négatifs) et la vocalisation qui ont le plus contribué au diagnostic final par enfant. La Table 3.9, à son tour, suit la même méthodologie et montre les enfants qui ont été incorrectement classés pour chaque modèle (faux positif et faux négatif).

Comme le montre le Tableau 3.8, les vocalisations avec le score le plus élevé parmi tous les groupes de caractéristiques sont les émotions négatives. Ce groupe comprend des vocalisations liées à la douleur ou à la colère tels que le cri et le gémissement. Les

caractéristiques prosodiques ont abouti au nombre le plus élevé de sujets ASD classés correctement (n = 15) par des vocalisations des "émotions négatives", suivies des caractéristiques spectrales de modulation (n = 10). À son tour, les caractéristiques de modulation spectrale et les ondelettes ont principalement contribué à attribuer correctement l'étiquette de contrôle dans la classe de vocation "autres". D'autre part, alors que les vocalisations des «émotions négatives» ont été utiles pour le diagnostic des ASD, elles ont contribué négativement à l'étiquetage des cas de contrôle, comme le montre le Tableau 3.9.

0.3.3 Expérience 3: Fusion au niveau des décisions et des fonctions

Tout d'abord, la fusion au niveau de la décision a été réalisée en combinant les décisions des classificateurs qui ont été formés et testés par des fonctions prosodiques, d'ondelettes et de modulation indépendamment. Trois schémas différents de fusion au niveau décisionnel, tels que décrits dans la Section 2.6, ont été utilisés. La Table 3.10 présente les résultats de classification pour les schémas de fusion de pluralité (PV), de probabilité maximale (MPV) et de probabilité moyenne (APV). En outre, chaque schéma de fusion est testé sous différents ensembles où WF, MF et FC1 correspondent aux ondelettes, la modulation et (moyenne) les caractéristiques prosodiques, respectivement. Comme le montre le Tableau 3.10, la combinaison des caractéristiques d'ondelettes et de modulation a atteint les performances les plus élevées sur toutes les méthodes testées. La fusion au niveau de la décision n'a pas amélioré la performance obtenue avec les classificateurs individuels, comme indiqué dans le Tableau 3.7.

Ensuite, on a utilisé une fusion au niveau des caractéristiques, associée à un schéma de réduction de la dimension MI. À la fin, un classificateur SVM a été formé sur les fonctionnalités du top 17 et les résultats sont rapportés au bas de la Table 3.10. Comme on peut le voir, la fusion au niveau des fonctions a permis d'améliorer la précision et la

spécificité du meilleur classement ASD versus non ASD, tout en maintenant le niveau de sensibilité à environ 90 %. La Table 3.11 répertorie les 17 principales fonctionnalités utilisées par ce classificateur. Comme on peut le voir, la plupart des caractéristiques proviennent de la classe spectrale de modulation.

0.4 Discussion

Au cours de la dernière décennie, la caractérisation acoustique-prosodique des enfants sur le spectre autistique a été explorée comme un marqueur possible pour une détection très précoce. Ici, nous avons exploré deux nouveaux ensembles de fonctionnalités, à savoir les caractéristiques dérivées d'une décomposition de paquets d'ondelettes et des caractéristiques dérivées d'une représentation de caractéristiques spectro-temporelles inspirées de l'audition. Nous avons montré que, dans une cohorte d'enfants en bas âge de 18 mois, nous avons pu discerner avec précision les deux groupes avec des précisions supérieures à celles obtenues avec des caractéristiques prosodiques proposées précédemment [35]. De tels résultats sont importants car une détection précoce peut permettre de commencer les premières interventions, ce qui améliore notamment le pronostic [6]. Dans les sections à suivre, nous discutons en détail les principales conclusions de notre étude à la lumière de la littérature existante.

Ondelettes mères

Avec la décomposition des paquets d'ondelettes, le signal est décomposé en versions mises à l'échelle et traduites d'une ondelette mère. Comme chaque famille d'ondelettes mère présente différentes caractéristiques tels que la symétrie, l'orthogonalité, la longueur du filtre et l'ordre de disparition, différentes propriétés du signal peuvent être capturées par différentes ondelettes de la mère. Dans cette enquête, Daubechies 8 ('db8') a été considérée comme la meilleure ondelette de la mère pour discriminer les deux

groupes témoins et ASD parmi d'autres ondelettes mère testées, y compris: coiffe, symétrie, biorthogonal et biorthogonal inverse. En outre, nos résultats ont montré que l'augmentation du niveau de décomposition a conduit à des caractéristiques plus détaillées et, par conséquent, à une meilleure performance de classification. Cela était vrai pour toutes les ondelettes mère testées, à l'exception de db8, dans laquelle une décomposition à 1 niveau était optimale (voir Tableau 3.6).

Dans la littérature de traitement du langage, l'ondelette db8 mère a été largement utilisée dans de nombreuses applications, y compris l'amélioration, la compression et la reconnaissance, pour n'en nommer que quelques [61]. La décomposition des vagues a également été utilisée dans le passé pour le crique pathologique et l'analyse du discours pathologique [47, 62, 53, 63]. C'est la première fois, cependant, que les caractéristiques d'ondelettes ont été explorées pour le diagnostic du spectre autistique. Dans [47], par exemple, les signaux de pleurs ont été décomposés en cinq niveaux à l'aide de quatre ondelettes mammaires différentes de la famille Daubechies: db1, db4, db10 et db20. La précision de classification la plus élevée a été atteinte au cinquième niveau de décomposition en utilisant l'ondelette mère "db20".

Alors que les niveaux de décomposition supérieurs ont peut-être aidé à la détection pathologique des pleurs, une décomposition simple à un niveau a été démontrée ici pour être optimale pour la tâche à accomplir. Une telle décomposition a probablement suffi à mesurer les différences d'énergie généralement signalées dans la littérature ASD (par exemple, [44]), l'entropie à haute fréquence représentative des sons respirants et sévères, ainsi que des énergies à haute fréquence représentatives de la qualité cri/hurlement [44]. Comme le montre le Tableau 3.8, les caractéristiques des ondelettes étaient utiles pour les classes de vocalisation vocale et négative, capturant ainsi ces qualités, respectivement.

Caractéristiques prosodiques

Selon le Tableau 3.8, les caractéristiques prosodiques ont été particulièrement utiles pour discriminer les ASD et les contrôles dans la classe de vocalisation des émotions négatives, ce qui contribue à la classification correcte d'environ la moitié des participants. De tels résultats corroborent ceux précédemment publiés dans la littérature qui ont montré des cris pour avoir différentes fréquences F0 et formantes entre les ASD et les contrôles [41, 12, 16, 42, 12, 17, 43, 44, 45, 37, 64]. Les catégories «autres» et «rires» ont été les premières à contribuer principalement à la classification correcte. De tels résultats corroborent également ceux dans la littérature qui ont montré le rire pour affecter l'inclinaison spectrale, F0 et les premières amplitudes formantes [65]. Les caractéristiques prosodiques ont généralement été explorées dans la littérature et sont utilisées ici comme référence pour le système proposé, ainsi que pour fournir des informations complémentaires aux caractéristiques spectrales des ondelettes et de la modulation proposées.

Caractéristiques des ondelettes

Selon la Table 3.8, l'émotion négative, les autres et la parole ont été les trois premières classes de vocalisation, respectivement, contribuant à une classification correcte lorsqu'on utilise uniquement des caractéristiques d'ondelettes. Les caractéristiques d'ondelettes calculées à partir de la décomposition à un niveau explorent essentiellement les niveaux d'énergie et la variabilité dans les plages de fréquences élevées et faibles, ainsi que l'entropie spectrale. Dans le passé, de tels détails, bien qu'ils ne soient pas calculés via WPD, se sont révélés discriminatoires entre les deux groupes. L'entropie spectrale, par exemple, était liée à des signaux rythmiques, à la respiration et à la dureté et pouvait discriminer les enfants autistes et généralement en développement: [40]. La variabilité énergétique, à son tour, s'est révélée être un aspet corrélatif de l'atypicité de la prosodie perçue dans ASD [66]. Les caractéristiques des ondelettes ont

également été utiles pour la détection des pleurs pathologiques et sur la reconnaissance de l'émotion par la parole [67, 68]. Dans [67], par exemple, les fonctionnalités basées sur les ondelettes ont été utiles pour la détection de la colère, alors que dans [68], elles ont été utiles pour discriminer les émotions en colère et dégoûtées. En calculant séparément les caractéristiques d'ondelettes pour différentes classes de vocalisation, on peut mesurer différents attributs, contribuant ainsi positivement à la détection ASD.

Caractéristique de la modulation

Les caractéristiques de modulation inspirées de l'audition ont été utilisées dans le passé pour la caractérisation pathologique [69, 50, 70] et la reconnaissance d'émotion de la parole [49]. Des modulations à haute fréquence ont également été liées au bruit de turbulence présent dans les plaintes de non-douleur [40], alors que certaines fréquences de modulation du crieur ont été liées aux troubles du système nerveux central [40]. En plus, des recherches récentes ont suggéré une diminution de l'extraction du rythme de la parole à partir de modèles de modulation temporelle dans la disclexion de la dyslexie du développement, une représentation neurale altérée de la structure sonore des mots typiquement observés avec des individus sur le spectre [71]. Le travail décrit ici illustre la première tentative d'utilisation des fonctions de modulation pour la détection ASD. De tels résultats corroborent le fait observé que les caractéristiques spectrales de modulation qui ont le plus contribué à la tâche en cours ont été calculées à partir d'émotions négatives, d'autres et de classes de parole (voir Tableau 3.8).

0.5 Conclusions et perspectives

La caractérisation acousto-prosodique des énoncés de vocalisation des tout-petits a été utile pour le diagnostic du trouble du spectre autistique. Les études existantes ont généralement exploré des irrégularités pendant les vibrations du pli vocal et l'utilisation

inappropriée du volume chez les individus atteints d'autisme [21, 22, 23, 14, 9, 24, 25, 26, 18]. Compte tenu du large éventail d'âge des participants, et de l'évolution rapide des caractéristiques du trait vocal pendant l'enfance, bon nombre des résultats rapportés ont été contradictoires [14]. De plus, les résultats ont généralement été rapportés en utilisant uniquement des énoncés semblables à la parole (par exemple, babillage) [17, 43, 44] ou le crie [41, 12, 16, 42, 12, 17, 43, 44, 45, 37], il n'est donc toujours pas clair quels types de vocalisation contribuent le plus à la classification. Enfin, une vaste littérature a exploré l'utilisation de caractéristiques d'ondelettes pour l'analyse de la parole et des pleurs infantile (par exemple, [53, 47, 48, 55]), ainsi que le spectre de la modulation de la parole (par exemple, [49, 50, 51]). Pour le meilleur de la connaissance de l'auteur, cependant, de telles fonctionnalités n'ont pas encore été explorées pour le diagnostic des ASD. Cette thèse vise à combler ces trois lacunes.

Plus précisément, nous explorons l'utilisation des caractéristiques des ondelettes, de la modulation spectrale et du prosodie pour classifier quarante-trois enfants en bas âge de 18 mois (dont 23 ont été diagnostiqués comme autistes à l'évaluation de 36 mois et un groupe de 20 témoins appariés selon l'âge) en groupes ASD et non-ASD. En se concentrant uniquement sur les données de 18 mois, la variabilité de la maturation des voies vocales est minimisée, ce qui met en lumière les caractéristiques vraiment discriminatives des ASD. Enfin, nous explorons les contributions de différents types de vocalisation, à savoir le babillage, le discours, le rire, les émotions négatives (groupage de vocalisations tels que les pleurs, les gémissements, les criques et les cris) et d'autres et explorer les effets que présentent différentes caractéristiques sur certains types de vocalisation pour le diagnostic global des ASD.

Dans l'ensemble, on a constaté qu'une précision de 81,5%, une sensibilité de 91,6% et une spécificité de 70% pourrait être obtenue avec un classificateur SVM individuel formé sur les fonctionnalités d'énergie et d'entropie à base d'ondelettes. Lorsqu'il a été formé avec des caractéristiques spectrales de modulation de la parole, une précision de 79%, une sensibilité de 80% et une spécificité de 75% pourraient être obtenues. Ces

précisions se sont comparées favorablement aux caractéristiques prosodiques précédemment proposées dans la littérature [35]. En plus, bien que la fusion au niveau de la décision n'améliore pas les performances globales, la fusion au niveau des fonctionnalités associée à la sélection des fonctionnalités a atteint une précision de 86,5%, une sensibilité de 90% et une spécificité de 80%, ce qui représente une amélioration relative par rapport à l'individu classificateur de 5% et 10% en termes de précision et de spécificité, respectivement. Une inspection approfondie des 17 principales caractéristiques sélectionnées a montré que les caractéristiques les plus importantes correspondaient aux caractéristiques spectrales de modulation. Fait intéressant, il a été observé que les vocalisations tels que les grincements et les cris semblaient être plus discriminatifs entre les classes que les vocations de parole, de babillage ou de rire. De tels résultats pourraient aider les cliniciens dans les évaluations futures, qui mettent actuellement l'accent sur les nuances prosodiques lors d'énoncés semblables à la parole.

En ce qui concerne les futures orientations de recherche, le travail présenté ici peut être développé de deux façons: (1) les enregistrements audio des évaluations ADOS antérieures (par exemple, à 12 mois) peuvent être explorés pour voir si des caractéristiques discriminatives peuvent être trouvées même à des âges antérieurs et (2) les données provenant de plus de jeunes enfants de 18 mois peuvent être incorporées pour valider les résultats obtenus, ainsi que pour explorer d'autres solutions. Plus de données, par exemple, pourraient permettre des classes de type de vocalisation plus équilibrées. Cela permettrait d'explorer de nouveaux schémas de fusion, tel qu'une combinaison optimale de fusion de classe de fonctionnalité par vocation. Une distribution de type de vocalisation plus équilibrée permettrait également des expériences plus précises explorant les effets de chaque sous-type sur la classification globale.

De telles explorations sont laissées pour le travail futur en raison de la quantité de main-d'œuvre impliquée dans l'extraction et l'étiquetage manuel de tous les types de vocalisation, ainsi que le rejet de ceux avec un chevauchement de l'expression des adultes. Malgré tout, avec les progrès rapides observés dans l'apprentissage par machine

et les réseaux nerveux profonds (DNN), il se peut que ces tâches laborieuses humaines puissent être remplacées par des machines. Cela dit, les étapes initiales dans les modèles de parole et la classification des événements audio utilisant des réseaux neuronaux profonds ont déjà été prises [72, 73] et la reconnaissance vocale par DNN du discours des enfants a déjà été proposée [74]. Les travaux futurs pourraient également explorer l'utilisation de la segmentation et de l'étiquetage basés sur DNN, ouvrant ainsi des portes à des études à grande échelle.

Chapter 1

Introduction

The American Psychiatric Association defines autism as a pervasive developmental disorder that is related to a triad of impairments: (1) atypical development in reciprocal social interaction; (2) atypical communication; and (3) restricted, stereotyped and repetitive behaviours [2]. In fact, the definition has recently been updated to include a wide spectrum of symptoms and impairment levels, thus the terminology Autism Spectrum Disorders (ASD) has been incorporated [2]. Common disorders within the spectrum can include: Autistic Disorder (AD), Asperger Syndrome (AS) and Pervasive Developmental Disorder Not Otherwise Specified (PDD-NOS). The definite (or stable) diagnosis of ASD is based on the presence of certain symptoms and their severity levels, and typically occurs around 36 months of age. Recent statistics suggest that roughly 1 in 68 children are diagnosed with autism and there is a recurrence rate of 18.7% for the biological siblings of autistic individuals [3, 4].

Recent research, however, has suggested that diagnosis can be accomplished around 18 months age [5]. Early diagnosis can allow parents to move forward with appropriate educational support [6] and clinicians to initiate intense interventions that take advantage of the young brain's plasticity properties [7]. New tools are needed, however, in

order to accurately diagnose ASD at such an early age. The Autism Diagnostic Observation Schedule (ADOS) is one of the gold-standard assessment tools used in the diagnosis of ASD. Recently, an item was added to the assessment to score a child's communication domain, thus pertaining to the existence of peculiar intonation, volume, rhythm and rate of vocalizations. It was included in the assessment following the widely noted clinical observation that many children on the autism spectrum present with atypical prosody of vocalizations and verbalizations [8, 9, 10, 11].

These studies have shown that vocalizations of autistic individuals are more difficult to recognize in terms of affective meaning and function than in their typically developing (TD) and developmentally delayed (DD) peers [12, 13]. Moreover, individuals on the spectrum have been observed to have a more hoarse, harsh, hyper-nasal voice quality, with higher recurrence of squeals, growls and yells [14]. Several terms have been used to describe prosody in ASD and include: monotonous, exaggerated, robotic, pedantic, and wooden, to name a few. Whatever the off feature might be, it is often noticed by social contact and may represent a significant barrier to peer acceptance. Atypicalities in the prosody of subjects with ASD were noted in the earliest description of autism, and studies show that impairments in vocal behaviour are evident even in pre-verbal autistic children and may represent an early feature of ASD [15, 16, 17, 12, 18].

Early studies have examined both perception and production of prosody in individuals with autism; however, an issue often addressed in the literature of ASD is the unknown prevalence of atypical prosody and the heterogeneity of the ASD population. Additionally, the lack of standardized investigative methods used to quantify vocal prosody and the shortcomings of perceptual judgment have produced inconsistent and sometimes contradictory findings [14, 19, 20]. As such, recent research has focused on incorporating computerized acoustic analysis within the diagnosis of ASD, thus overcoming some of these limitations involved in qualitative studies [9, 10, 11]. Acoustic characteristics of speech production have been measured and quantified across several studies as possible markers of ASD. Most existing studies, however, have been con-

ducted on older children or with younger children but with a wide age range of 18-36 months. Given the natural changes in vocal characteristics (e.g., growing and maturation of the vocal tract) that occur during childhood, it is important that acoustic analysis focus on individuals of roughly the same age, thus removing the potential bias from natural age-related acoustic changes and variability, and allowing the system to focus solely on disorder-related differences. The work described herein fills this gap.

1.1 Prosody in ASD

Disordered prosody has long been considered a hallmark of ASD and atypical intonation has been recognized as an important diagnostic ASD masker. Prosodic characterization of children with ASD is an under-researched area, particularly for very young and pre-verbal children, although studies suggest vocal atypicality may represent an early appearing symptom of ASD. Studies have traditionally measured and quantified parameters of intonation, volume and vocal quality to identify differences between ASD and comparison groups. The most persisting finding reported in the literature is that individuals with ASD present greater variability and a wider range in fundamental frequency (F0) measures [21, 22, 23, 14, 9, 24, 25, 26, 18]. This finding is to an extent counter-intuitive as children on the autism spectrum are often described as having monotone or robotic speech. Few studies have also suggested that impairments in the use of pitch related features in autistic individuals are linked to an abnormal processing of auditory stimuli at the brainstem level [27, 28, 29]. More recently, a higher pitch coefficient of variation (CV) per word was found for the TD group in comparison with the ASD group at school age, whilst no significant differences between the two classes could be observed at a pre-school age [11].

In addition to disordered pitch, ASD individuals have been reported to have speech that is described as “too fast” or “too slow” [14]. Most studies that have measured

sentence and word duration have reported longer duration in individuals with ASD [14, 30, 10, 31, 32]. In [10] for example, a perceptual and acoustic analysis in children with Asperger syndrome was employed. The study showed that even when children with ASD and TD performed similarly in the perception of intonation patterns and the use of prosodic features to express grammatical meaning, children with ASD showed an alteration in duration, mean and maximum pitch. These findings are in line with the results shown in [32], where a significant difference in duration for emotional speech (sad utterances having a longer duration than happy or angry utterances) was found for TD and AS groups, while subjects with high-functioning autism (HFA) did not show any difference. However, two recent studies did not find significant differences in duration between ASD and TD children [21, 33]. Again, such contradictory results may be due to a wide age-range of the study participants, thus results may be biased due to natural age-related acoustic changes and variability.

Moreover, a few studies have employed multivariate analysis using machine learning techniques. In [34], for example, four groups of features: voice quality, energy-related, spectral and cepstral features were compared for a database collected to assess children abilities in imitation of different types of prosody contours. The results showed that voice quality features improved classification performance over the other feature groups. Other studies measured voice quality related measures such as jitter, Harmonic-to-Noise Ratio (HNR) and cepstral peak prominence (CPP), which tend to increase in the children with autism with increasing ASD severity [35, 9, 25, 24].

1.2 Early vocal patterns in ASD

Mothers of children with ASD may have a difficult time recognizing the affective meaning of their infants' vocalizations, and recent evidence is emerging to indicate that vocal atypicality may be apparent in very young infants with ASD [12, 16, 36, 37, 13].

However, most of the studies have tended to focus on verbal adolescents and adults, though some have studied school and pre-school age children as well. Infant vocalizations are the earliest form of vocal communication. They play an important role in the development of the parent-child relationship and language acquisition. Infants begin to produce vegetative or reflexive sounds such as coughing or crying after birth, and through several stages the children expand their sounds and their vocalizations to become more speech-like [38]. The growth and anatomic restructuring of the vocal tract during the first half year of life and the vocal learning are the main factors that induce a change in child vocalizations [38]. Recently, a few studies have been analyzing the development of acoustic parameters on infant's vocalizations as a non-invasive tool to measure vocal-muscular maturation [39, 12, 16, 40]. Atypicalities in the production of vocal patterns could involve abnormal processing of auditory feedback or problems in the speech production mechanisms. In the next subsections we explore some studies employed in pre-verbal autistic children. Specifically crying and canonical babbling vocalizations have been studied in the ASD literature.

1.2.1 Crying

Crying is the infant's earliest form of vocal communication. It has been explored due to its relationship with the central nervous system [41, 12, 16, 42, 12, 17, 43, 44, 45, 37]. Acoustic abnormalities in infant's cries have been associated with some disorders such as asphyxiation, low birth weight, metabolic disorders, neurological symptoms, and lead exposure, amongst and others. Most of them evidence a high and variable pitch [41]. In the ASD literature, the cries of children later diagnosed with autism exhibited a higher F0 value and shorter pauses than the cries of developmentally delayed or typically developing children [12, 16, 36, 37]. Moreoever, the fundamental frequency was shown to decrease for healthy children during the first and second year of life, unlike the case with children later diagnosed with ASD [36, 16]. Sheinkopf and colleges employed a

cry acoustic analysis of 6-month old infants [46] and showed that, at 6 months, high risk children started to show a higher and variable pitch value than those with low risk. Later, when the same high-risk subjects were analyzed at the age of 36 months, they showed an even higher F0. Additionally, other studies have analyzed the variability in the fundamental frequency (pitch range) but no differences were found between the two groups [37, 46].

A more recent study in ASD employed a reaction time (RT) categorical task to analyze adults' responses of cries of children between 36 and 40 months of age with ASD [16]. They found that differences in vocal behavior in children later diagnosed with ASD caused adults to perceive the cries as distressed and more difficult to process and interpret than the cries of typically developing children, as well as mammalian animal cries and environmental noise control sounds. Esposito and colleagues, in turn, examined acoustic features of infant-cry vocalizations of a group of typically developing children and children with ASD, both aged 13 months [12], and found that the pause length was more important for the perception of distress in children with ASD than the fundamental frequency or the frequency of cry sounds per unit time across an episode of crying.

1.2.2 Canonical babbling

Babbling begins shortly after birth and is well established by 10 months old. Any delay in onset of canonical babbling is related to language delay or other developmental disabilities [17]. Just a few studies have focused on the analysis of canonical babbling in children with ASD [17, 43, 44]. Patten and colleagues studied the canonical status and vocalization frequency of a group of 37 infants with ASD compared with a typical development group at 9-12 and 15-18 months[17]. Individuals later diagnosed with autism produced significantly lower rates and had a later onset of canonical babbling than the control group. These findings are congruent with the results obtained in the

analysis of canonical syllable production of infants aged between 16-48 months [44]. However, some reported findings have been contradictory. For example, Sheinkopf and colleagues studied the nature of early vocal behaviours in young children with autism with a focus on canonical babbling and atypical vocal quality (defined as the rate of production of atypical phonation) [43] and the group with ASD did not display significant differences compared with a developmentally delayed control group in terms of rate of canonical babbling, despite their vocalizations showing atypical vocal quality [17, 44].

1.3 Thesis Objectives and Contributions

There is a growing evidence that children with ASD have a delayed developmental trajectory of speech and that early markers can be present within the first years of life. Since identification of acoustic differences in vocalizations can help in the development of earlier, more effective, targeted interventions, the objective of this thesis is to develop an automated system to discriminate between ASD and non-ASD groups. To this end, an acoustic analysis of pre-verbal vocalizations of 43 children (23 with ASD, 20 controls) during their 18-month assessment is performed. We extend previous work in automated autism detection in three ways.

First, we explore the contributions of different types of vocalizations (e.g., cry, speech, babble, laugh, shout, yell, squeal and whine) for the task at hand. Given the varying trends reported for different vocalization types, it is expected that they will contribute differently for overall ASD diagnosis performance. This knowledge has yet to be reported in the literature. Second, we propose the use of two new feature sets, namely, wavelet-based and speech modulation spectral-based features for ASD diagnosis. Wavelet features have been explored in the past for infant cry analysis (e.g., [47, 48]), thus are expected to accurately characterize vocalizations such as cries, squeals, yells,

etc. Speech modulation spectral features, in turn, have been widely used for speech emotion characterization [49] and for pathological voice analysis (e.g., [50, 51]). We explore their benefits individually, as well as combined with existing prosodic features. To the best of our knowledge, this is the first attempt at combining wavelet and modulation spectral (and prosodic) features for the purpose of autism detection. Third, by utilizing data from a routine 18-month assessment, the findings reported herein are based on data obtained from participants with the lowest and tightest age range. Since child vocalizations are known to change continuously during childhood, given the maturation of their vocal tracts, such tight age range reduces any potential biases and allows the classifiers to place focus on existing ASD discriminative features.

1.4 Thesis Outline

This rest of the thesis is organized as follows. In Chapter 2 we introduce our experimental methodology, including data collection, data pre-processing, feature extraction, classier training, fusion schemes, and the figures-of-merit used to gauge system performance. In Chapter 3, we present the experimental results while in Chapter 4 we discuss the obtained findings. Finally, conclusions and future research directions are presented in Chapter 5.

Chapter 2

Methods and materials

2.1 Preamble

In this chapter, methods and materials used throughout this study are described. First, Section 2.2 and Section 2.3 present details of the data collection and pre-processing. Next, Sections 2.4.1-2.4.5 describe the proposed wavelet-based and modulation spectral features, as well as classical prosodic features, respectively. Sections 2.5 and 2.6, in turn, describe the classification procedure and fusion schemes used in our study, respectively. Figures of merit used to gauge system performance are finally presented in Section 2.7.

2.2 Data collection

Data used in this study were extracted from a set of videotaped ADOS - Module 1 sessions, which are part of the longitudinal prospective Canadian "Infant Sibling Study" from the Autism Research Unit at Toronto's SickKids Hospital [52]. The study monitors younger siblings of probands with ASD, recruited due to a known higher risk to exhibit the disorder, estimated at an 18.7% recurrence rate [3], as well as low-risk age-

matched “control" children of families without a history of ASD. This ADOS module was designed for children whose spontaneous language is primarily in single words or simple phrases. The evaluation consists of a group of interactive tasks between the baby and a clinician, and parent interviews to evaluate social relatedness, communication skills, motor activity and play behaviours. Participants are followed during the age of 6-24 months and every 3-12 months undergo a series of (re)assessments, including the ADOS and other standardized developmental and language tests. At the 36-month follow up visit, a well experienced clinician, blinded to previous outcomes and impressions, assesses them for ASD utilizing gold-standard clinical tools, such as medical history, ADOS, and the Autism Diagnostic Interview - Revised (ADI-R) [75].

Our analysis was conducted on a subset of the Infant Sibling Database and relied on audio recordings of 43 participants during their 18-month assessment. Table 2.1 presents the participant demographics. The ASD group includes 23 children independently diagnosed with ASD at the age of 36 months and encompassed both Asperger syndrome and Autism disorder diagnoses. An age-matched comparison group of 20 low-risk TD children was used. Children in the control group received the same follow-up assessments as the ASD group and were determined at 36 months of age not to have ASD. As per the larger study, participants that were born premature or with low birth-weight are excluded from the study.

The ADOS sessions had durations ranging from 24 to 52 minutes. Audio content was extracted from the video recordings and toddler’s vocalizations were manually segmented and labelled according to vocalization type. Instances of vocalizations with overlapping adult speech (parents, clinician) were discarded from our analysis. Overall, the total audio segments extracted from *only* the toddler vocalizations resulted in 127.0 and 194.5 seconds for control and ASD classes, respectively.

Table 2.1 – Participant demographics

Group	Age (months)	Male	Female
Control	18 ± 0.23	13	7
ASD	18 ± 0.42	15	8

2.3 Pre-processing

In the literature of infant phonology, a vocalization occurs on expiration and when an inspiration occurs, it is perceived as a break that separates the vocal events [44]. Each vocal event separated by a breath time is called an utterance. In order to extract the utterances of each one of the vocalizations in our database and to avoid processing silent/noise-only intervals, an automated energy-thresholding method is used, as in [44]. The approach computes the energy of the signal every 10-ms and compares this value with a pre-defined energy threshold. Once the energy of the signal first rises to 90% of the baseline, a start point is detected and when it falls to less than 10% of the baseline, the vocalization is deemed to have ended. The energy threshold was defined empirically based on a small subset of our available recordings. Using this segmentation method, a total of 2647 utterances was obtained for both ASD and non-ASD groups. Table 2.2 summarizes the amount of vocalization utterances for each group and type of vocalization. The class labeled as "negative emotions" combines vocalization utterances such as cry, squeal, whine and shout. As shown, the numbers of vocalizations per group and per vocalization type are not balanced with babble and speech being the most prominent and laugh the least.

Table 2.2 – Summary of vocalization utterances for ASD and control groups

Group	ASD	Control	Total
Babble	451	333	784
Speech	375	527	902
Laugh	49	14	63
Negative emotions	224	96	320
Others	378	200	578
Total	1477	1170	2647

2.4 Feature Extraction

2.4.1 Wavelet Packet Decomposition (WPD)

Wavelet Packet Decomposition (WPD) is a generalization method of the Discrete Wavelet Transform (DWT) that allows a time-frequency multi-resolution analysis of an input signal. WPD has been used in previous studies for emotion recognition, speech analysis and also has shown to be useful in the analysis of pathological speech and pathological infant cry [53, 47, 55]. Due to the highly non-stationary characteristics of some vocalizations such as cry, squeal and shout, the wavelet analysis results more suitable for detection and classification than the traditional Fourier methods [76]. Unlike the DWT, both low and high frequency bands are used for further decomposition in WPD. In the signal decomposition, a signal or function $f(t)$ can be represented as a weighted sum of real-valued functions $\Psi_i(t)$ such as [54, 53]:

$$f(t) = \sum_i a_i \Psi_i(t), \tag{2.1}$$

where a_i are the real-valued expansion coefficients and i is an integer index for the finite or infinite sum. The expansion set $\Psi_i(t)$ is a basis set or basis if the expansion coefficients a_i in Equation 2.1 are unique for $f(t)$. Further, the basis set is called

orthogonal if the following holds:

$$\langle \Psi_i(t), \Psi_j(t) \rangle = \int \Psi_i(t) \cdot \Psi_j dt = 0, \qquad i \neq j. \tag{2.2}$$

The expansion coefficients are calculated using the inner product as a measure of similarity. When the function indexes are identical, the inner product is equal to a unit function and as consequence the expansion coefficients are described by equation 2.3.

$$a_i = \langle f(t), \Psi_i(t) \rangle = \int (f(t) \cdot \Psi_i) dt. \tag{2.3}$$

If the expansion coefficients described above are used to linearly combine the basis set, the original signal $f(t)$ is recovered. In wavelet expansion, all the bases are generated by simple scaling and translation of two orthogonal functions, known as the mother wavelet function $\Psi_{j,k}$ and the scaling function $\varphi_{j,k}$. They are defined by:

$$\Psi_{j,k}(t) = \sqrt{2^j} \Psi(2^j t - k), \tag{2.4}$$

$$\varphi_{j,k}(t) = \sqrt{2^j} \varphi(2^j t - k). \tag{2.5}$$

In the above equations, the factor $\sqrt{2^j}$ maintains a constant energy norm and j and k are integers. The signal $f(t)$ can be described in terms of the scaling function such as :

$$f(t) = \sum_k c_j(k) \varphi_{j,k} = \sum_k c_j(k) \sqrt{2^j} \varphi(2^j t - k). \tag{2.6}$$

The size of the subspace spanned by the scaling function increases when j increases. However, wider scaling functions can represent only coarse information, and the space spanned is smaller. The wavelet functions are introduced to model the differences

between the spaces spanned by the various scales of the scaling function as below:

$$f(t) = \sum_{k=-\infty}^{+\infty} c_k \varphi_k(t) + \sum_{j=0}^{+\infty} \sum_{k=-\infty}^{+\infty} d_{j,k} \Psi_{j,k}(t). \tag{2.7}$$

Finally, the expansion coefficients can be calculated in the jth scale on the basis the expansion coefficients of one step higher scaled as

$$c_j(k) = \sum_m h_0(m - 2k) c_{j+1}(m), \tag{2.8}$$

$$d_j(k) = \sum_m h_1(m - 2k) c_{j+1}(m). \tag{2.9}$$

In the above equations, h_0 and h_1 are the impulse response of the low-pass and high-pass filters that decompose the signal. The relation of both filters in terms of length N can be related as:

$$h_1(n) = (-1)^n h_0(N - 1 - n). \tag{2.10}$$

As can been seen, the decomposition process and multi-resolution analysis can be viewed as the application of a filter bank. More specifically, the input signal is passed through low-pass and high-pass filters, which correspond to the scaling and wavelet function [53, 54]. The lower frequency band gives the approximation coefficients and higher frequency band the detail coefficients. In wavelet packet decomposition, the process is recursively applied to both frequency sub-bands to generate the next level of decomposition. The wavelet packets can be described in terms of basic functions as follows:

$$W_{2n}(2^{P-1}x - l) = \sqrt{2^{1-P}} \sum_m h(m - 2l) \sqrt{2^P} W_n(2^P x - m), \tag{2.11}$$

$$W_{2n+1}(2^{P-1}x - l) = \sqrt{2^{1-P}} \sum_{m} g(m - 2l)\sqrt{2^P} W_n(2^P x - m), \tag{2.12}$$

where P is the scale index, l the translation index, h low-pass filter and g high-pass filter. The relation of both filters is given in terms of the K filter length as

$$g(k) = (-1)^k h(K - 1 - k). \tag{2.13}$$

Due to the orthonormal property, we can re-define the above equation as follows:

$$C^P_{n,k} = \sqrt{2^P} \sum_{m=-\infty}^{\infty} f(m) \cdot W_n(2^P m - k), \tag{2.14}$$

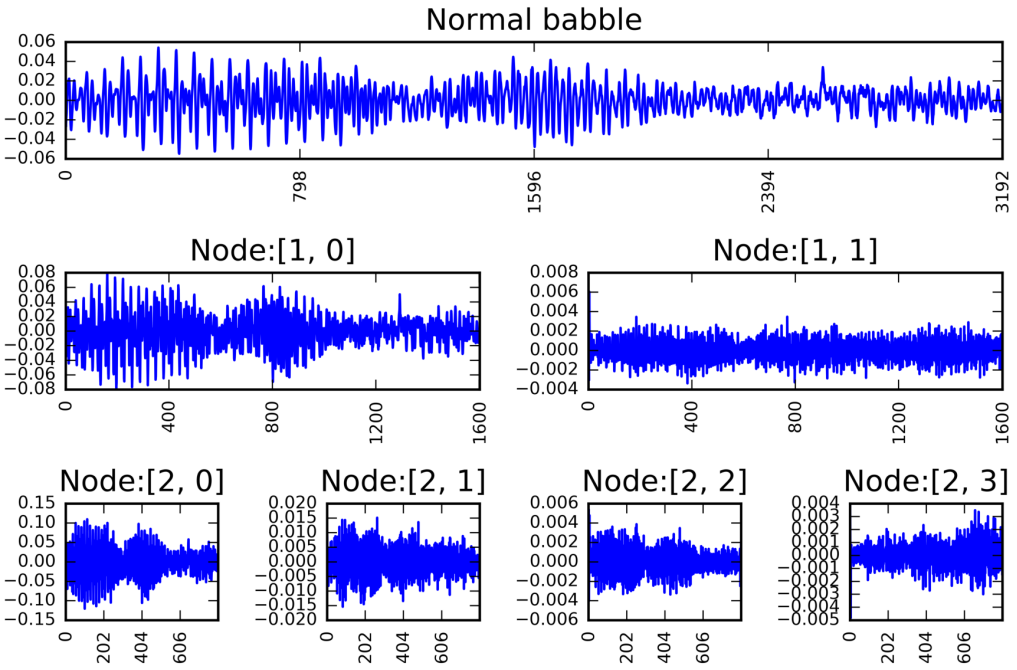
$$C^{P-1}_{2n,l} = \sum_{m} h(m - 2l) \cdot C^P_{n,m}, \tag{2.15}$$

$$C^{P-1}_{2n+1,l} = \sum_{m} g(m - 2l) \cdot C^P_{n,m}. \tag{2.16}$$

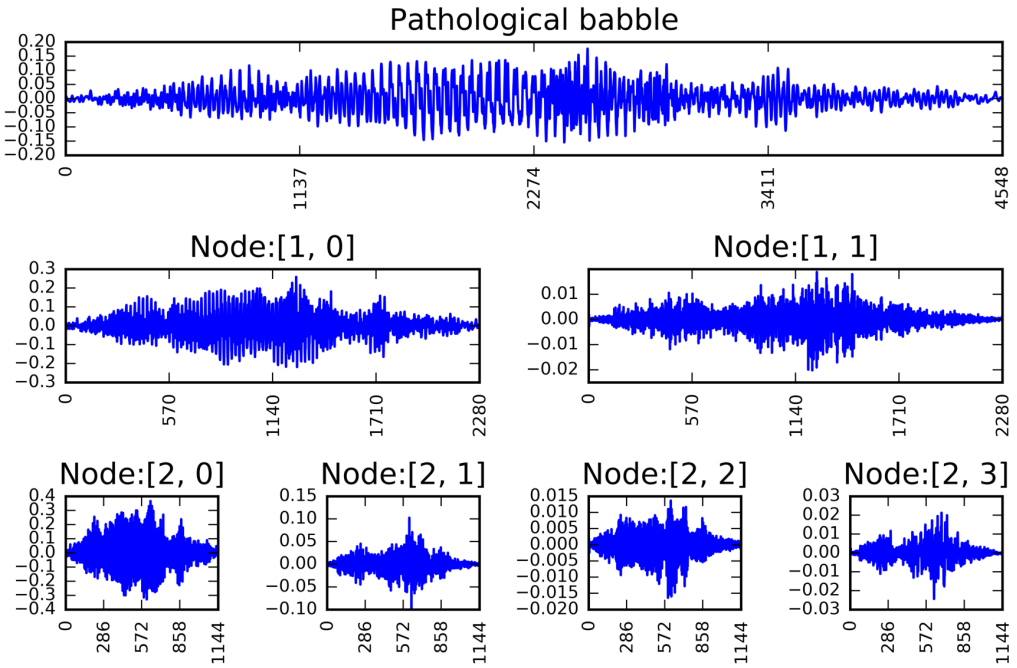
Figure 2.1 shows an example of a two-level wavelet packet decomposition for a babble signal generated from a control and an ASD participant. For n levels of decomposition, the WPD produces 2^n finer equal-width frequency sub-bands or nodes. The frequency ranges given in Hertz for the level n of decomposition are described by:

$$\left[\frac{kfs}{2^{n+1}} \frac{(k+1)fs}{2^{n+1}}\right] \quad k = 0, 1, ..2^{n-1}, \tag{2.17}$$

where f_s is the sampling frequency of the original signal [47]. Table 2.3 shows the frequency ranges for each level of decomposition for a signal with a sampling frequency of 16 kHz, as is the case in the database used herein. These frequency ranges are used in our experiments to decompose vocalization utterances.



(a) Normal babble signal



(b) Pathological babble signal

Figure 2.1 – Two-level wavelet packet decomposition of (a) control and (b) ASD babble signals with 'bior2.6' mother wavelet

Decomposition level	**Frequency band (Hz)**
1	0-4000, 4000-8000
2	0-2000, 2000-4000, 4000-8000
3	0-1000, 1000-2000, 2000-4000, 4000-8000
4	0-500, 500-1000, 1000-2000, 2000-4000, 4000-8000

Table 2.3 – Frequency bands for wavelet packet decomposition of a signal with sampling frequency of 16 kHz

2.4.2 Wavelet features

Given the WPD described above, we propose the use of energy and entropy based features for our analysis, computed from each wavelet packet coefficients $C^P_{n,k}$ as follows [53, 47, 55], respectively:

$$E_{(n,k,l)} = \sum_{l=-\infty}^{+\infty} |C^P_{n,k}(m)|^2 w(l-m), \tag{2.18}$$

$$S_{(n,k,l)} = \sum_{l=-\infty}^{+\infty} -|C^P_{n,k}(m)|^2 \log |C^P_{n,k}(m)|^2 w(l-m), \tag{2.19}$$

where l is the number of window frame, P is the scale index, n represents the decomposition level and $k = 0, 1, ..., 2^{n-1}$ is the node number. In our experiments, each sub-band wavelet packet coefficient is divided into frames of 40 ms and successive frames were overlapped by 50%. Finally, statistical measures such as mean ($\bar{X}_{(n,k)}$), standard deviation ($std_{(n,k)}$), skewness ($g_{(n,k)}$), and kurtosis ($G_{(n,k)}$) are computed for all per-frame measures over the entire vocalization utterance. For the sake of completeness, these statistical metrics are computed as follows:

$$\bar{X}_{(n,k)} = \frac{1}{N}\sum_{l=1}^{N} X_{(n,k,l)}, \tag{2.20}$$

$$std_{(n,k)} = \sqrt{\frac{1}{N}\sum_{l=1}(X_{(n,k,l)} - \bar{X}_{(n,k)})^2}, \tag{2.21}$$

$$g_{(n,k)} = \frac{\sum_{l=1}^{N}(X_{(n,k,l)} - \bar{X}_{(n,k)})^3/N}{std(n,k)^3}, \tag{2.22}$$

$$G_{(n,k)} = \frac{\sum_{l=1}^{N}(X_{(n,k,l)} - \bar{X}_{(n,k)})^4/N}{std(n,k)^4}, \tag{2.23}$$

where $X_{(n,k)}$ can represent either the per-frame energy or entropy of the coefficients.

The final feature vector is an 8-dimensional feature vector computed per node k and decomposition level i, comprised of the mean, standard deviation, skewness, and kurtosis of the energy and the entropy values. Figure 2.2 depicts a block diagram of the proposed feature extraction scheme.

2.4.3 Speech Modulation Spectral Representation

The so-called speech modulation spectral signal representation is an auditory-inspired spectro-temporal representation that captures both acoustic frequency and temporal modulation frequency properties of the analyzed signal [49]. Figure 2.3 shows the steps used in our approach to compute the spectro-temporal (ST) representation of an input signal. In the first step, the active signal level is normalized to -26 dBov (dB overload) using the P.56 speech voltmeter [56]. Next, a bank of 23 critical-band gammatone filters is used to model the frequency response of the basilar membrane. The filter bank is designed as an array of independent bandpass filters inspired in the Patterson's cochlear model [57]. The first filter is centered at 125 Hz and the last one at half of the sample frequency of the analyzed signal. In Patterson's model, the bandwidth of each filter is described by a psychoacoustic measure called Equivalent Rectangular Bandwidth

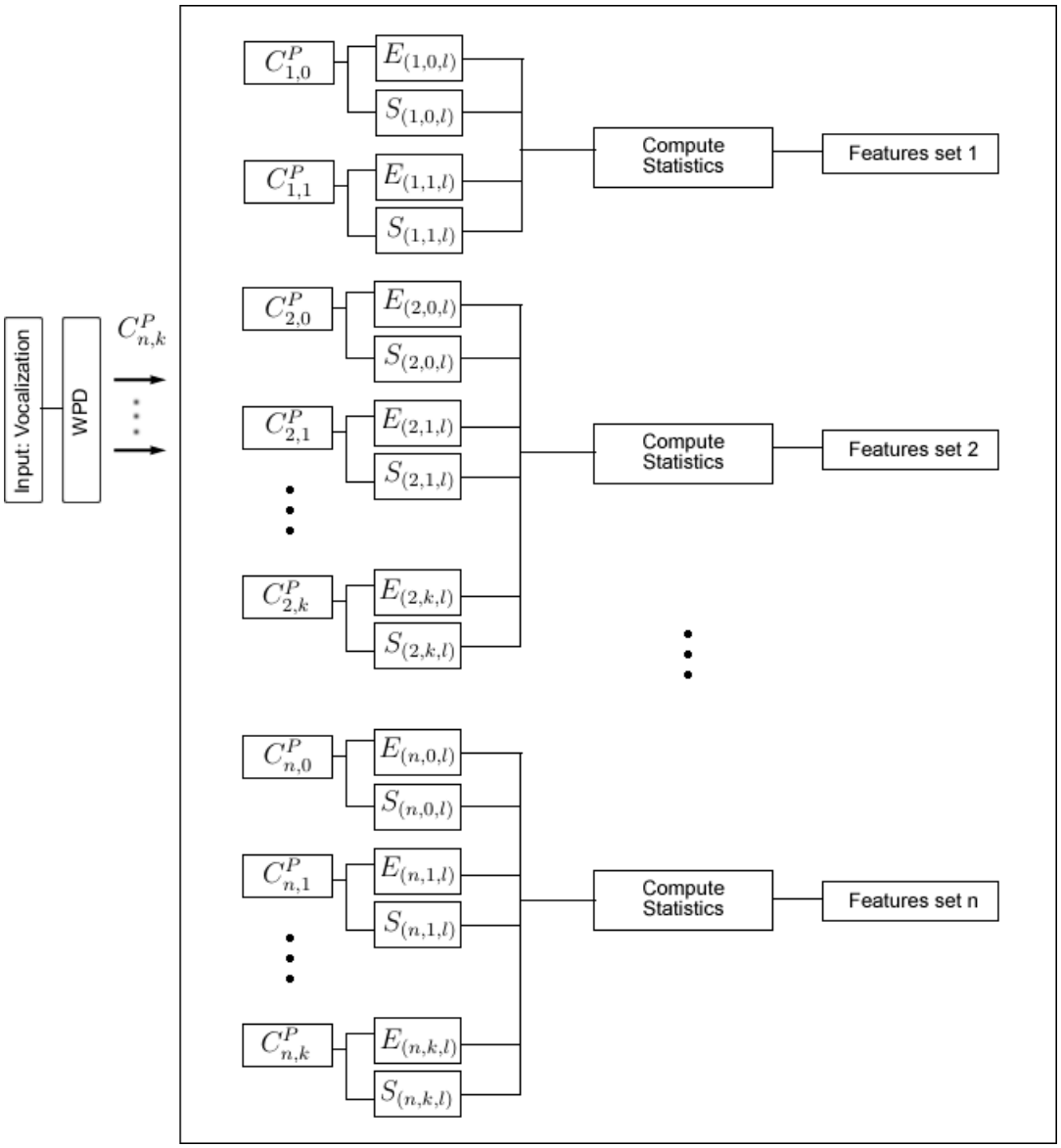


Figure 2.2 – Block diagram of the wavelet packet decomposition-based feature extraction method

(ERB) and is computed as follows:

$$ERB_i = \frac{f_i}{Q_{ear}} + B_{min}, \tag{2.24}$$

where f_i is the center frequency given in Hz of the ith critical-band filter, and Q_{ear} and B_{min} are constants set to 9.26449 and 24.7, respectively.

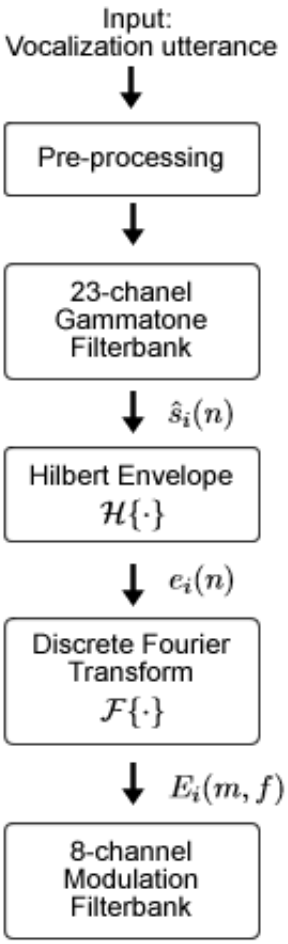


Figure 2.3 – General scheme to compute ST representation

The frequency response of the filter bank used in our experiments is depicted in Figure 2.4. The center frequencies of the gammatone filter bank correspond to the acoustic frequencies. Once the 23 critical-band gammatone filter bank is applied, the envelope $e_i(n)$ is computed for each one of the filterbank outputs $\hat{s}_i(n)$ using the Hilbert transform $\mathcal{H}\{\cdot\}$. The envelope of the ith bandpass filter signal is given by:

$$e_i(n) = \sqrt{\hat{s}_i(n)^2 + \mathcal{H}\{\hat{s}_i(n)\}^2} \qquad i = 1 \cdots, 23 \tag{2.25}$$

The modulation spectrum of the signal is computed using the Discrete Fourier Transform (DFT). Specifically, the envelope signal $e_i(n)$ is divided into frames of 256 ms every

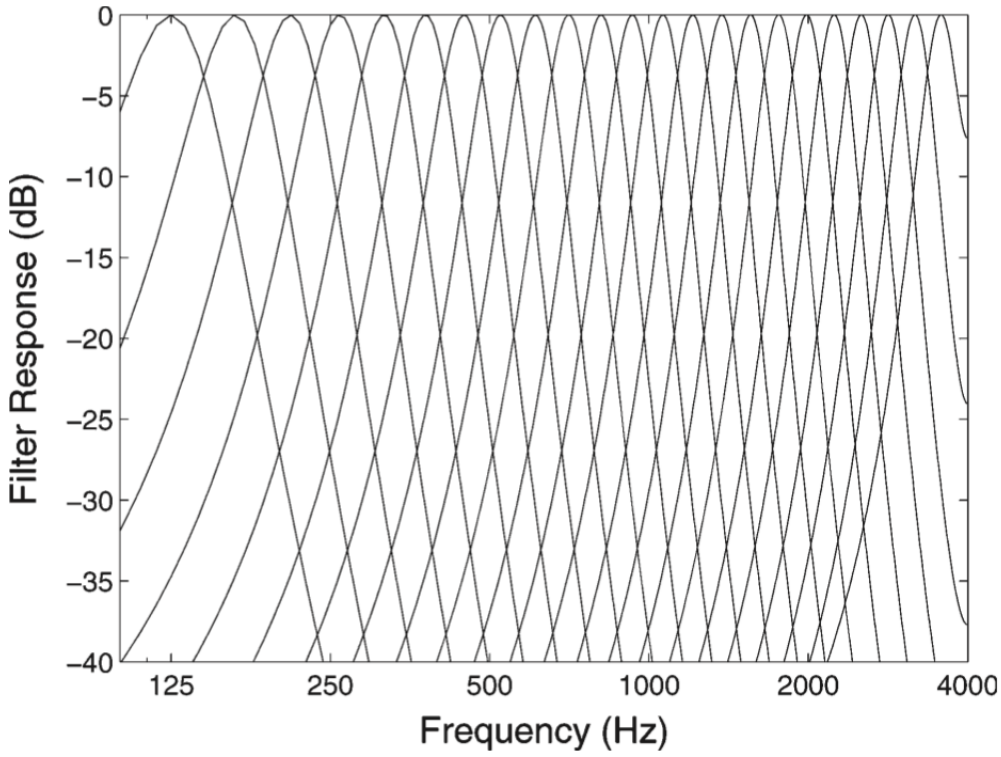


Figure 2.4 – Frequency responses of the 23-channel gammatone filterbank [1]

40 ms using a Hamming window. The notation $e_i(m)$ is used to indicate the frame m of the windowed envelope. The next step is to take the DTF $\mathcal{F}\{\cdot\}$ of each frame. Then, the modulation spectrum is defined by

$$E_i(m, f) = |\mathcal{F}(e_i(m))|, \tag{2.26}$$

where f is the modulation frequency. Finally, an auditory-inspired modulation filter bank allows us to build a representation in function of the acoustic frequency and temporal modulation frequency elements. The modulation energy of the ith critical-band signal is grouped into 8 bands, each band is denoted as $\varepsilon_{i,k}(m), k = 1, \ldots, 8$, where k is the kth modulation filter. Figure 2.5 shows the frequency responses for the modulation filter bank. The frequency responses are equally spaced in the logarithm scale from 4 to 128 Hz.

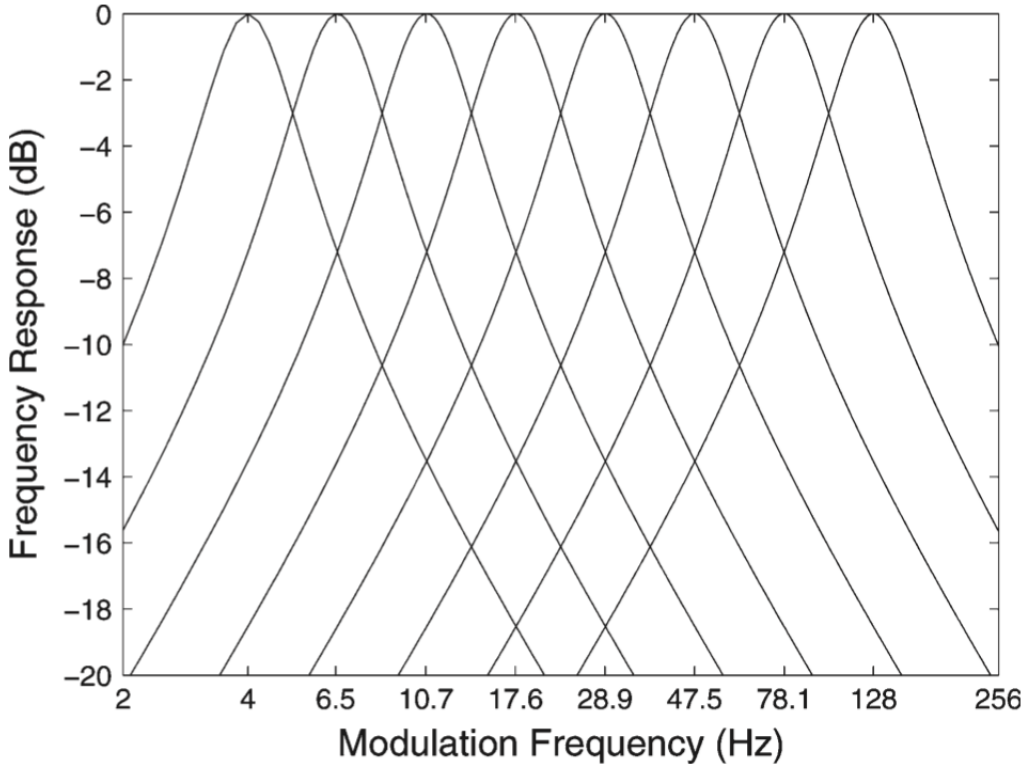


Figure 2.5 – Frequency responses of the 8-channel modulation filterbank [1]

2.4.4 Modulation spectral features

Given the modulation spectral representation above, the set of features originally proposed in [49] is extracted. The first feature set, $\Phi_{1,m}(k)$, represents the energy distribution along the modulation frequency. It is defined as the mean of the energy samples with respect to the k-th modulation channel:

$$\Phi_{1,m}(k) = \frac{\sum_{i=1}^{N} \varepsilon_{i,k}(m)}{N}. \tag{2.27}$$

The second set, $\Phi_{2,m}(k)$, is defined as the ratio of the geometric mean of a spectral energy measure and its arithmetic mean value, thus representing the spectral flatness of the spectrum. A spectral flatness value close to 1 is related with a flat spectrum, whilst a value close to 0 suggests a spectrum with high variations in its spectral amplitude.

This measure is computed as follows:

$$\Phi_{2,m}(k) = \frac{\sqrt[N]{\Pi_{i=1}^{N} \varepsilon_{i,k}(m)}}{\Phi_{1,m}(k)}. \tag{2.28}$$

The third $\Phi_{3,m}(k)$ measure corresponds to the center of mass of each modulation channel, where $f(i)$ is the index of the critical band. The spectral centroid for the i modulation channel is given by:

$$\Phi_{3,m}(k) = \frac{\sum_{i=1}^{N} f(i)\varepsilon_{i,k}(m)}{\varepsilon_{i,k}(m)}. \tag{2.29}$$

In order to measure the relationship of different modulation channels, the 23 acoustic channels are grouped in five levels: $D_1 = [1-4]$, $D_2 = [5-8]$, $D_3 = [9-12]$, $D_4 = [13-18]$ and $D_5 = [19-23]$. The modulation channels into each category are summed and used to compute the spectral centroid $\Phi_{4,m}(k)$ in the modulation frequency domain for D_l as follow:

$$\mathsf{E}_m(l,k) = \sum_{i \in D_l} \varepsilon_{i,k}(m), \tag{2.30}$$

$$\Phi_{4,m}(k) = \frac{\sum_{k=1}^{8} k\mathsf{E}_m(l,k)}{\sum_{k=1}^{8} \mathsf{E}_m(l,k)}. \tag{2.31}$$

The two final measurements capture the rate of change of each acoustic frequency region, thus providing an indication of the temporal dynamics of the utterances. The linear regression coefficient $\Phi_{5,m}(k)$ (slope) and the corresponding regression error $\Phi_{6,m}(k)$ (root mean squared error, RMSE) are computed. Those measures are associated to the first-degree polynomial model used to fit $\mathsf{E}_m(l,k)$. The final feature vector includes 184 modulation spectrum energy features $\varepsilon_{i,k}(m), k = 1, \ldots, 8$ plus the 39 features described above, thus totalling 223 features.

2.4.5 Acoustic-prosodic measures

Prosodic-acoustic features and their variations between groups have been the most widely used features in the analysis of autism spectrum disorders, as mentioned in Chapter 1. Here, vocalization utterances were acoustically analyzed using the Voice-Sauce MATLAB toolbox from the UCLA SPAP laboratory. Many of the parameters estimated by VoiceSauce depend on F0 and the formant range of the input signal. In our experiments, both measures are optimized for children's vocalizations. Pitch and formant-related parameters were computed using a fundamental frequency (F0) range of 60-1600 Hz and a nominal frequency F1 of 1250 Hz, which correspond to the nominal frequency of a 7 cm long vocal tract. The features were extracted from 25-ms frames every 10 ms.

In total, 26 acoustic parameters are extracted, as listed in Table 2.4. The final feature group includes those related to intonation (pitch), maturity of speech (first formant frequencies and amplitudes), volume (energy) and measures of vocal quality such as voice breathiness, and harshness/creakiness (harmonics, spectral tilt and cepstral peak prominence). In order to explore the variations of prosodic features between ASD and control groups, three different prosodic feature combinations are proposed. The first group (PF1) includes the mean value of the features reported in Table 2.4 for each vocalization utterance. The second group (PF2) combined the distribution mean and standard deviation, and finally, mean, standard deviation, and range are included in the third group (PF3). Such a partition was shown useful in [35].

Table 2.4 – List of extracted acoustic-prosodic parameters

Parameter	**Acronym**
Fundamental Frequency	F0
Formant Frequencies	F1, F2, F3, F4
Formant Frequency Bandwidths	BW1, BW2, BW3, BW4
Harmonic Spectra (location and magnitude)	H1, H2, H4, A1, A2, A3
Differences of Harmonic Spectra at Corrected Formant Frequencies	H1*-H2*, H2*-H4*, H1*-A1*, H1*-A2*, H1*-A3*
Volume	Energy
Cepstral Peak Prominence	CPP
Harmonic to Noise Ratio	HNR5, HNR15, HNR25, HNR35

2.5 Classifier Design

A Support Vector Machine (SVM) is a supervised learning model widely used in pattern recognition and classification problems [77]. An SVM classifies data using a hyperplane as a decision boundary to separate classes with the largest separation or margin between them. The margin is defined after a training data set is mapped into a higher dimensional space obtained via a kernel function. Three of the most widely used kernels include linear, polynomial and radial basis function (RBF). The kernel and hyperplane with the maximum margin are chosen as an optimal model. A complete description of the SVM algorithm is beyond the scope of this thesis and the interested reader is referred to [78] for more complete details. Here, an SVM classifier is used to discriminate between control and ASD classes. The next subsections provide details about the parameter estimation and validation process performed in our work.

2.5.1 Model selection and parameter estimation

In our experiments, three SVMs are trained separately on vocalization utterances from the three different feature groups, namely wavelets, modulation spectral, and acoustic-prosodic. The SVM implementation in [58] was adopted and an RBF kernel was chosen as it resulted in improved performance during our pilot experiments. In order to find the optimal parameters for each SVM, a grid search methodology was used. In this methodology, a set of models, which differ from each other in their hyper-parameters values, is trained and evaluated using k -fold cross-validation (CV). The model with the best average performance in the grid search is selected. Thus the following steps are followed to select the optimal model:

1. The feature vector group $m(1 \leq m \leq 3)$ is divided into $k = 4$ sets of approximately 11 subjects. Ideally, each set has the same number of subjects from each class (control and ASD) and approximately equal total sample size.

2. A grid space of (C, γ) with C= [4, 2, 1, 0.1, 0.01, 0.001, 0.0001] and γ = [2, 1.0, 0.5, 0.1, 0.05, 0.01, 0.005, 0.007] is defined.
3. For each hyper-parameter pair (C, γ) in the search space:
   (a) The classifier is trained $k = 4$ times, each time with a different set held out as a test set.
   (b) The estimated performance $CR\{C(i), \gamma(j)\}$ is the average of the accuracies obtained in the k folds.
4. Choose the best parameter (C, γ) that leads to the highest cross-validation classification rate.
5. Use the best parameter to create a model as predictor.
6. Repeat the steps 3-5 for features vector $m + 1$.

2.5.2 Validation process

In order to evaluate the SVMs classification system performance, stratified 10-fold cross-validation is employed using the hyper-parameters found with the process described above. In stratified 10-fold CV, the feature vector is divided into ten partitions of approximately 4 subjects, where nine sets are used for training and the remaining are left for testing only. The sets are designed as to ensure the classes are equally represented across each test fold. The process is repeated ten times, and the performance is computed on a per-participant basis. With this approach, an infant is labeled as control or ASD using a score-based scheme of the decisions made by the SVM. This method was chosen empirically due its superior performance when compared with the common method of plurality vote. More specifically, SVM outputs of the vocalization utterances for each subject are compared and the vocalization with the highest likelihood score decides the final class prediction. Thus, if $c(x_i)$ corresponds to the prediction score for

sample x_i, then the final prediction score can be computed as:

$$C = \arg\max[c(x_1) \cdots c(x_i)], \tag{2.32}$$

where $c(x_i)$ corresponds to the distance of the sample x_i to the separating hyperplane.

2.6 Fusion Schemes

2.6.1 Decision-Level Fusion

Decision-level fusion schemes combine the decisions from different classifiers in order to achieve higher robustness and to improve the performance of single-classifier systems. The combination problem consists of finding the combination function accepting N-dimensional input vectors from M classifiers and outputting N final classification decisions, where the optimal function is the function that minimize the misclassification cost. The input vector depends on the combination function and it can be probabilities, scores or labels from the classifiers. In our experiments, all the samples x_i belonging to the subject s_l from the different classifiers are used for the combination problem in order to make a per-subject diagnosis. Three different combination functions are proposed:

1. Plurality vote (PV): This is the simplest and more common-based fusion method. The subject s_l is assigned to the class c_j that obtained the highest number of votes. In this case, all the classifier weights are equal, i.e., $w_k = 1/K \ \forall K$.

2. Maximum probability vote (MPV): In this fusion scheme, the probabilities for the samples x_i belonging to the subject s_l are compared and the sample with

the highest probability decides the final prediction:

$$C = \arg\max[p(x_1) \cdots p(x_i)]. \tag{2.33}$$

3. Average probability vote (APV): The per-sample conditional probabilities per each class are averaged and the class c_j that obtained the highest average probability decides the final prediction. Thus if $p(c_j/x_i)$ is the conditional probability that x_i belong to the class c_j, the final prediction is made as follows:

$$C = \arg\max\left[\frac{\sum_i p(c_1/x_i)}{i}, \frac{\sum_i p(c_2/x_i)}{i}\right]. \tag{2.34}$$

In our experiments, the classifiers in the ensembles are comparable in the sense that they have been trained on the same data sets and using the same partitioning.

2.6.2 Feature-Level Fusion

In feature-level fusion, the feature sets from different sources are concatenated into a single feature vector before the classification process. The main advantage of this method is that correlated features within and between different feature sets can be removed via dimensionality reduction tools, thus improving the generability of the system. In our experiments, a mutual information (MI)-based algorithm was used in order to measure the degree of relatedness between the feature values [59, 60]. The MI of the feature set is computed for the algorithm using the nearest-neighbour method. The details of the method are presented in [59].

2.7 Figures-of-merit

In diagnostic tests, validity is the ability of a test to discriminate subjects with and without disease. To aid in making diagnosis, we use sensitivity (Sen) and specificity (Spec) as validate measures. Sensitivity is defined as the probability of getting a positive test result in subjects with the disease. On the other hand, specificity is a measure complementary to a sensitivity. It measures the ability of a test to correctly classify those who not have the disease. The specificity and sensitivity are reported as percentages and are computed as:

$$Spec = \frac{TP}{TP + FP}, \tag{2.35}$$

$$Sen = \frac{TP}{TP + FN}. \tag{2.36}$$

where TP, TN, FP and FN are true positive, true negative, false positive and false negative, respectively. Finally, the accuracy of the test is measured in terms of both sensitivity and specificity as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \tag{2.37}$$

2.8 Summary

This chapter has presented a general overview of the methodology used in our experiments. We introduced wavelet, modulation spectral, and prosodic features. These features are used in our experiments to classify 43 children into ASD and non-ASD groups. The details of the database an pre-processing are presented in Section 2.2 and Section 2.3, respectively. Three Support Vector Machines (SVMs) are trained separately for each feature group. The model selection and parameter estimation for each

SVMs in our experiments follow the methodology presented in Section 2.5. In the next chapter, we will present the experimental results of the proposed methodology.

Chapter 3

Experimental Results

3.1 Experiment 1: Wavelet Mother Selection

Our first aim is to investigate and compare the effectiveness of different types of mother wavelets and the distribution of information in several decomposition levels for the discrimination of Autism Spectrum Disorders. In order to do so, the extraction feature methodology proposed in Section 2.4.1 is employed using different wavelet families, such as daubechies (db), coiflet (coif), symlet (sym), biorthogonal (bior), and reverse biorthogonal (rb). Energy and entropy features are extracted from the wavelet-packet coefficient at several decomposition levels and compared between them for each kind of mother wavelet.

Tables 3.1 - 3.5 present the classification results for each wavelet family. The index number specified in each wavelet family refers to the vanishing order of the wavelet, which is related with the length of the filter. Best performances per group are highlighted in bold. From Table 3.1, it was found that; the maximum accuracy of 81.5% was obtained from the "db8" mother wavelet using the extracted features from the first level of WP decomposition. On other hand, coiflet and symlet wavelet families both

achieved a maximum performance of 76.5% at the third level of decomposition, as seen in Table 3.2 and Table 3.3. The “coif7” and “sym10” were the mother wavelets with the best performance. Finally, from Table 3.4 and Table 3.5, the maximum performance was achieved at the last level of decomposition for “bior2.8” and “rbior3.1”. An accuracy of 77.0% and 71.5% was obtained for biorthogonal and reverse biorthogonal wavelets, respectively.

Table 3.1 – Classification results at each level of WP decomposition using the daubechies wavelet family

Level of decomposition												
	1			**2**			**3**			**4**		
Wavelet	**Acc %**	**Sen %**	**Spec %**	**Acc %**	**Sen %**	**Spec %**	**Acc %**	**Sen %**	**Spec %**	**Acc %**	**Sen %**	**Spec %**
db1	65.5	85.0	45.0	62.5	95.0	25.0	65.5	80.0	50.0	67.0	78.3	55.0
db2	75.5	80.0	70.0	59.5	80.0	35.0	68.0	80.0	55.0	63.0	83.3	40.0
db3	66.5	68.3	65.0	66.0	70.0	60.0	68.0	80.0	55.0	59.5	78.3	40.0
db4	66.0	91.6	35.0	68.5	85.0	50.0	60.0	73.3	45.0	67.0	83.3	50.0
db5	67.0	86.6	45.0	62.0	95.0	25.0	58.5	80.0	35.0	62.0	81.6	40.0
db6	74.0	76.6	70.0	68.0	85.0	50.0	72.5	88.3	55.0	58.0	73.3	40.0
db7	69.5	81.6	55.0	55.0	90.0	15.0	66.0	90.0	40.0	65.0	83.3	45.0
db8	**81.5**	91.6	70.0	62.0	95.0	25.0	74.0	100	45.0	65.0	88.3	40.0
db9	76.0	81.6	70.0	75.0	90.0	60.0	52.5	75.0	30.0	60.5	78.3	40.0
db10	69.5	80.0	60.0	71.5	96.6	45.0	72.0	83.3	60.0	62.0	73.3	50.0

Table 3.2 – Classification results at each level of WP decomposition using the coiflet wavelet family

	Level of decomposition											
	1			**2**			**3**			**4**		
Wavelet	**Acc %**	**Sen %**	**Spec %**	**Acc %**	**Sen %**	**Spec %**	**Acc %**	**Sen %**	**Spec %**	**Acc %**	**Sen %**	**Spec %**
coif1	68.5	80.0	55.0	58.0	91.6	20.0	65.0	80.0	50.0	59.5	88.3	30.0
coif2	64.5	95.0	30.0	73.5	90.0	55.0	71.5	95.0	45.0	58.0	75.0	40.0
coif3	62.0	86.6	35.0	59.5	95.0	20.0	55.0	70.0	40.0	65.0	80.0	50.0
coif4	57.0	80.0	30.0	67.5	80.0	55.0	71.5	81.6	60.0	63.5	90.0	35.0
coif5	61.5	80.0	40.0	69.0	95.0	40.0	73.0	95.0	50.0	62.5	83.3	40.0
coif6	69.5	90.0	45.0	67.0	95.0	35.0	66.0	81.6	50.0	74.0	88.3	60.0
coif7	69.0	85.0	50.0	65.5	80.0	50.0	**76.5**	91.6	60.0	65.0	90.0	40.0
coif8	68.5	85.0	50.0	63.0	85.0	40.0	66.5	88.3	45.0	69.0	86.6	50.0
coif9	66.0	85.0	45.0	67.5	88.3	45.0	55.0	95.0	10.0	60.0	88.3	30.0
coif10	68.5	85.0	50.0	67.5	88.3	45.0	60.0	75.0	45.0	64.0	91.6	35.0

Table 3.3 – Classification results at each level of WP decomposition using the symlet wavelet family

	Level of decomposition											
	1			**2**			**3**			**4**		
Wavelet	**Acc %**	**Sen %**	**Spec %**	**Acc %**	**Sen %**	**Spec %**	**Acc %**	**Sen %**	**Spec %**	**Acc %**	**Sen %**	**Spec %**
sym2	75.5	80.0	70.0	59.5	80.0	35.0	68.0	80.0	55.0	63.0	83.3	40.0
sym3	66.5	68.3	65.0	66.0	70.0	60.0	68.0	80.0	55.0	59.5	78.3	40.0
sym4	71.0	81.6	60.0	76.0	95.0	55.0	68.5	71.6	65.0	58.5	80.0	35.0
sym5	56.5	75.0	35.0	59.5	95.0	20.0	73.0	85.0	60.0	71.5	90.0	50.0
sym6	57.5	95.0	15.0	61.5	85.0	35.0	66.5	76.6	55.0	70.0	88.3	50.0
sym7	73.5	81.6	65.0	64.5	68.3	60.0	67.5	70.0	65.0	59.5	73.3	45.0
sym8	64.5	86.6	40.0	71.5	90.0	50.0	71.5	85.0	55.0	68.0	81.6	55.0
sym9	59.0	80.0	35.0	62.0	78.3	45.0	67.0	83.3	50.0	62.5	80.0	45.0
sym10	62.0	85.0	35.0	62.5	95.0	25.0	**76.5**	83.3	70.0	63.0	75.0	50.0

Table 3.4 – Classification results at each level of WP decomposition using the biorthogonal wavelet family

	Level of decomposition											
	1			**2**			**3**			**4**		
Wavelet	**Acc %**	**Sen %**	**Spec %**	**Acc %**	**Sen %**	**Spec %**	**Acc %**	**Sen %**	**Spec %**	**Acc %**	**Sen %**	**Spec %**
bior1.1	65.5	85.0	45.0	62.5	95.0	25.0	65.5	80.0	50.0	67.0	78.3	55.0
bior1.3	65.5	80.0	50.0	75.5	95.0	55.0	68.0	70.0	65.0	72.5	83.3	60.0
bior1.5	66.0	85.0	45.0	66.0	65.0	65.0	67.0	100.0	30.0	63.0	83.3	40.0
bior2.2	64.5	76.6	50.0	72.0	96.6	45.0	62.5	80.0	45.0	55.0	73.3	35.0
bior2.4	71.5	81.6	60.0	69.0	90.0	45.0	55.0	65.0	45.0	62.5	78.3	45.0
bior2.6	64.0	71.6	55.0	67.0	78.3	55.0	64.5	95.0	30.0	**77.0**	91.6	60.0
bior2.8	62.0	71.6	50.0	48.5	71.6	20.0	55.0	60.0	50.0	69.0	85.0	50.0
bior3.1	74.0	91.6	55.0	67.5	73.3	60.0	53.0	68.3	35.0	65.5	83.3	45.0
bior3.3	68.5	81.6	55.0	59.5	75.0	40.0	62.0	95.0	25.0	65.0	80.0	50.0

Table 3.5 – Classification results at each level of WP decomposition using the reverse biorthogonal wavelet family

	Level of decomposition											
	1			**2**			**3**			**4**		
Wavelet	**Acc %**	**Sen %**	**Spec %**	**Acc %**	**Sen %**	**Spec %**	**Acc %**	**Sen %**	**Spec %**	**Acc %**	**Sen %**	**Spec %**
rbior1.1	65.5	85.0	45.0	62.5	95.0	25.0	65.5	80.0	50.0	67.0	78.3	55.0
rbior1.3	63.5	90.0	35.0	63.5	71.6	55.0	57.5	65.0	50.0	69.5	86.6	50.0
rbior1.5	62.0	71.6	50.0	71.5	86.6	55.0	68.0	85.0	50.0	66.5	90.0	40.0
rbior2.2	58.0	73.3	40.0	67.0	75.0	60.0	71.5	86.6	55.0	69.0	86.6	50.0
rbior2.4	56.0	75.0	35.0	69.0	85.0	50.0	62.0	95.0	25.0	62.5	70.0	55.0
rbior2.6	54.5	70.0	35.0	61.5	95.0	25.0	61.5	81.6	40.0	70.0	90.0	50.0
rbior2.8	55.0	80.0	25.0	67.0	73.3	60.0	62.5	83.3	40.0	62.5	70.0	55.0
rbior3.1	58.0	78.3	35.0	61.5	80.0	40.0	61.5	76.6	45.0	**71.5**	73.3	70.0
rbior3.3	61.5	81.6	40.0	64.5	73.3	55.0	56.0	83.3	25.0	69.0	90.0	45.0

A summary of the best performance results is presented in Table 3.6. Maximum recognition accuracy of 81.5% was achieved using a first-level decomposition and eighteenth order daubechies mother wavelet. Additionally, it was found that increasing the wavelet decomposition level improved the general performance for most mother wavelets.

Table 3.6 – Summary of best results per wavelet family

Wavelet	Acc %	Sen %	Spe %	Level of decomposition
db8	81.5	91.6	70.0	1
coif7	76.5	91.6	60.0	3
sym10	76.5	83.3	70.0	3
bior2.6	77.0	91.6	60.0	4
rbior3.1	71.5	73.3	70.0	4

3.2 Experiment 2: Feature Set Comparisons

Experimental results for the different feature sets proposed in Chapter 2 are reported in Table 3.7. As can be seen, the proposed wavelet features achieved the best accuracy, reaching up to 81.5% average recognition rate. They are followed by the modulation spectral features with a performance of 79.0%. The three benchmark prosodic feature sets achieved similar performances with PF1 achieving the best overall performance amongst the PF1-PF3 (see Section 2.4.3 for details).

Table 3.7 – Recognition results for the different features sets proposed

Features group	Acc (%)	Sen (%)	Spec (%)	AUC
Prosodic-PF1	71.5	90.0	50.0	0.71
Prosodic-PF2	65.0	86.6	40.0	0.69
Prosodic-PF3	65.0	91.6	35.0	0.56
Wavelet features "db8"	81.5	91.6	70.0	0.81
Modulation spectral features	79.0	80.0	75.0	0.72

The average recognition rates in Table 3.7 are reported on a per-participant basis as is described in Section 2.5.2. A total of 43 children (23 with ASD, 20 control) were

diagnosed for each classification model separately. The vocalization with the highest score per subject made the final diagnosis. Table 3.8 and Table 3.9 present the details of the diagnosis made between control and ASD classes for each group of features. Specifically, Table 3.8 shows the number of children correctly classified (true positives and true negatives) and the vocalization that most contributed to the final diagnosis per child. Table 3.9, in turn, follows the same methodology and shows the children that were incorrectly classified for each model (false positive and false negative).

As shown in Table 3.8, the vocalizations with the highest score among all the feature groups are the negative emotions. This group includes pain/angry related vocalizations such as cry, squeal, whine and shout. The prosodic features resulted in the highest number of ASD subjects classified correctly (n=15) through "negative emotions" vocalizations, followed by modulation spectral features (n=10). In turn, modulation spectral and wavelet features contributed mostly to correctly assigning the control label within the "others" vocalization class. On the other hand, while "negative emotions" vocalizations were shown helpful in the diagnosis of ASD, they contributed negatively to the labelling of control cases, as shown in Table 3.9.

Table 3.8 – Children correctly classified per model

	Prosodic-PF1 features		**Wavelet features "db8"**		**Modulation Spectral features**	
Vocalization	ASD	Control	ASD	Control	ASD	Control
Babble	0	1	2	2	3	1
Speech	2	0	6	0	1	4
Laugh	3	1	2	1	1	0
Others	1	4	3	7	4	6
Negative emotions	15	4	8	4	10	4
TP/TN	21	10	21	14	19	15

Table 3.9 – Children incorrectly classified per model

	Prosodic-PF1 features		**Wavelet features "db8"**		**Modulation Spectral features**	
Vocalization	ASD	Control	ASD	Control	ASD	Control
Babble	1	0	1	1	2	2
Speech	0	1	0	1	0	1
Laugh	0	2	0	0	0	0
Others	0	1	0	2	0	1
Negative emotions	1	6	1	2	2	1
FP/FN	2	10	2	6	4	5

3.3 Experiment 3: Decision- and Feature-level Fusion

First, decision-level fusion was performed by combining decisions from the classifiers that were trained and tested by prosodic, wavelet and modulation features independently. Three different decision-level fusion schemes, as described in Section 2.6, were employed. Table 3.10 presents the classification results for plurality (PV), maximum probability (MPV) and average probability (APV) fusion schemes. Also, each fusion scheme is tested under different ensembles where WF, MF and FC1 correspond to the wavelet, modulation, and (mean) prosodic features, respectively. As shown in Table 3.10, the combination of wavelet and modulation features achieved the highest performance over all methods tested. Notwithstanding, decision level fusion did not improve the performance obtained with the individual classifiers, as reported in Table 3.7.

Next, feature level fusion combined with an MI dimensionality reduction scheme was employed. In the end, an SVM classifier was trained on the top-17 features and the results are reported at the bottom of Table 3.10. As can be seen, feature-level fusion was able to improve the accuracy and specificity of the best individual ASD versus non-ASD classifier, whilst maintaining the sensitivity level at around 90%. Table 3.11 lists the top 17 features used by this classifier. As can be seen, most of the features

are from the modulation spectral class. Finally, Table 3.12 presents the details of the diagnosis made between control and ASD classes using the top 17 features selected.

Table 3.10 – Recognition results for different decision level fusion and feature-level schemes

Decision level fusion			
Plurality vote (PV)			
Features group	Acc (%)	Sen (%)	Spec (%)
WF+MF+PF1	71.5	81.6	60.0
WF+MF	74.0	91.6	55.0
WF+PF1	71.5	76.6	65.0
MF+PF1	63.0	81.6	40.0
Maximum probability vote (MPV)			
Features group	Acc (%)	Sen(%)	Spec(%)
WF+MF+PF1	69.5	90.0	45.0
WF+MF	79.0	90.0	65.0
WF+PF1	74.0	95.0	50.0
MF+PF1	67.5	95.0	35.0
Average probability vote (APV)			
Features gruop	Acc (%)	Sen(%)	Spec(%)
WF+MF+PF1	69.0	90.0	45.0
WF+MF	74.0	90.0	55.0
WF+PF1	61.5	90.0	30.0
MF+PF1	69.5	95.0	40.0
Feature-level fusion			
Features group	Acc (%)	Sen(%)	Spec(%)
WF+MF+PF1	86.5	90.0	80.0

Table 3.11 – Top 17 features chosen using the mutual information-based algorithm for classification of ASD and control groups

Feature selected	Type of feature
Entropy[1,1]_mean	Wavelet
$\varepsilon_{23,6}$	MSF
$\Phi_{2,m}(6)$	MSF
$\varepsilon_{8,5}$	MSF
Energy[1,1]_mean	Wavelet
H1A3C_mean	Prosodic
$\varepsilon_{1,8}$	MSF
$\Phi_{1,m}(3)$	MSF
$\varepsilon_{13,6}$	MSF
$\varepsilon_{4,6}$	MSF
$\varepsilon_{4,4}$	MSF
Energy[1,0]_mean	Wavelet
A1_mean	Prosodic
Energy[1,0]_std	Wavelet
$\varepsilon_{5,5}$	MSF
$\varepsilon_{4,7}$	MSF
$\varepsilon_{14,8}$	MSF

Table 3.12 – Children correctly classified after feature-level fusion

	Selected features	
Vocalization	ASD	Control
Babble	4	1
Speech	1	6
Laugh	4	1
Others	4	5
Negative emotions	8	3
FP/FN	21	16

3.4 Summary

In this chapter, results of the experimental evaluation were presented. In the first experiment, different types of mother wavelets were investigated and compared for the discrimination of ASD. Additionally, the distribution of information in several decomposition levels for five different wavelet families were included. The mother wavelet

with the best performance was compared with modulation and prosody features in Experiment 2. Finally, Experiment 3 presented the results for different fusion schemes.

Chapter 4

Discussion

Over the last decade, acoustic-prosodic characterization of children on the autism spectrum has been explored as a possible marker for very early detection. Here, we have explored two new features sets, namely features derived from a wavelet packet decomposition and features derived from an auditory-inspired spectro-temporal feature representation. We showed that on a cohort of 18-month old toddlers, we were able to accurately discriminate between the two groups with accuracies higher than those achieved with previously-proposed prosodic features [35]. Such findings are important as such early detection can allow for early interventions to commence, thus notably improving prognosis [6]. In the sections to follow, we discuss in detail the major findings of our study in light of the existing literature.

4.1 Vocalization types

Infants produce a wide variety of vocal expressions during the first years of life. Children use these vocal expressions as a communicative tool to express different emotions or different communicative functions to caregivers. From an acoustic point of view,

these vocalizations also exhibit different patterns that can be the subject of further analyses [79, 80, 81]. Several studies in language acquisition and early identification of pathologies have analyzed the specific characteristics of different vocalizations such as cries, babbles, laughter, and grunts due to their relevance during language and communicative skills development [79, 80, 81, 82, 38, 41, 12, 16, 42, 12, 17, 43, 44, 45, 37].

Vocal development in infants is considered a continuous, but non-linear process. Early vocalizations are precursors of speech and language development. In the literature, vocalizations such as crying and babbling have received significant attention. For example, babbling is likely to influence the development of spoken language due to the fact that words are composed of canonical syllables [79, 17, 83]. Infants start to produce babbling after 10 months and it has been shown that any delay in the onset of canonical babbling is a significant predictor of language delay or other disabilities [84, 85]. Given the relationship between language acquisition and babbling, few studies have explored the production of babbling in autistic individuals [17, 43, 44]. Infants diagnosed with ASD display low rates of canonical babbling, lower number of total syllables produced (volubility) and an onset in the canonical babbling stage compared with typically developing children [17, 44].

Crying, in contrast, is the first method of communication for an infant. It is used to express different needs, states, and demands. Frequency vibration of the vocal cords has been related to the dominance of laryngeal processes in early sound production [39]. Previous studies, including autism spectrum disorder analysis, have shown that low birth weight infants and infants with neurological symptoms have different acoustic patterns such as fundamental frequency (F0), vocal tract resonance frequencies, pause length, amplitude modulations, and number of utterances compared with typically developing children [41, 12, 16, 42, 12, 17, 43, 44, 45, 37]. In babies later diagnosed with ASD, cries were shown to convey high levels of distress, a factor later attributed to modulation deficits and unnatural F0 and formant values [36].

In our study, vocalizations such as cry, squeal, whine and shouts (called here "negative emotions") were grouped in order to allow for a more balanced class relative to e.g., speech and babble. Table 3.8, in fact, suggests that such vocalization types were more important than speech or babble in helping correctly classify between ASD and controls, regardless of the feature used. Similar findings could be seen with feature fusion, as reported in Table 3.12.

Laughter has also been studied within the infant population to convey information about the child's mental/affective state [86]. Autistic children have been reported as having laughter episodes without an apparent motivating stimulus or in response to specific stimuli [87, 80]. Other studies have shown that patients with neurological disorders exhibit uncontrollable episodes of pathological crying, pathological laughing or both, thus potentially related to impairment in the control and use of their emotions [88, 89]. Within our study, laughter was shown to be a useful vocalization type to help correctly detect ASD, particularly within the prosodic feature space (see Table 3.8).

4.2 Features

4.2.1 Mother wavelets

With wavelet packet decomposition, the signal is decomposed into scaled and translated versions of a mother wavelet. As each family of mother wavelets presents different characteristics such as symmetry, orthogonality, filter length, and vanishing order, different signal properties may be captured by different mother wavelets. In this investigation, Daubechies 8 ("db8") was deemed the best mother wavelet to discriminate between both control and ASD groups amongst other tested mother wavelets, including: coiflet, symlet, biorthogonal, and reverse biorthogonal. Additionally, our results showed that increasing the level of decomposition led to more detailed features and, consequently,

better classification performance. This was true for all tested mother wavelets, except db8, in which a 1-level decomposition showed to be optimal (see Table 3.6).

In the speech processing literature, the db8 mother wavelet has been shown to be widely used across numerous applications, including enhancement, compression, and recognition, to name a few [61]. Wavelet decomposition has also been used in the past for pathological cry and pathological speech analysis [47, 62, 53, 63]. This is the first time, however, that wavelet features have been explored for autism spectrum diagnosis. In [47], for example, cry signals were decomposed into five levels using four different mother wavelets from the daubechies family, namely: "db1", "db4", "db10" and "db20". The highest classification accuracy was achieved at the fifth level of decomposition using the "db20" mother wavelet.

While higher decomposition levels may have assisted with pathological cry detection, a simple one-level decomposition showed here to be optimal for the task at hand. Such decomposition likely sufficed to measure the energy differences typically reported within the ASD literature (e.g., [44]), the high frequency entropy representative of breathy and harsh sounds, as well as high frequency energies representative of squeal/cry quality [44]. As shown in Table 3.8, wavelet features were useful for the speech, other and negative emotion vocalization classes, thus likely capturing these qualities, respectively.

4.2.2 Prosodic features

According to Table 3.8, prosodic features were shown to be particularly useful in discriminating between ASD and controls within the negative emotion vocalization class, contributing to the correct classification of roughly half the participants. Such findings corroborate those previously published in the literature, which have shown cries to have different F0 and formant frequencies between ASD and controls [41, 12, 16, 42, 12, 17, 43, 44, 45, 37, 64]. The ‘others’ and ‘laughter’ categories were the second to contribute

mostly to correct classification. Such findings also corroborate those in the literature that have shown laughter to affect spectral tilt, F0, and first formant amplitudes [65]. Prosodic features have been typically explored in the literature and are used here as a benchmark to the proposed system, as well as providing complementary information to the proposed wavelet and modulation spectral features.

4.2.3 Wavelet features

As per Table 3.8, negative emotion, others and speech were the top three vocalization classes, respectively, contributing to correct classification when using only wavelet features. Wavelet features computed from the one-level decomposition basically explore energy levels and variability in high and low frequency ranges, as well as spectral entropy. In the past, such details, while not computed via WPD, were shown to discriminate between the two groups. Spectral entropy, for example, was related to rhythmic cues, breathiness, and harshness and could discriminate between autistic and typically developing children [40]. Energy variability, in turn, was shown to be a correlate of perceived prosody atypicality in ASD [66]. Wavelet features have also been shown useful in pathological cry detection and in emotion recognition from speech [67, 68]. In [67], for example, wavelet-based features were shown useful for anger detection, whereas in [68], they were shown useful in discriminating angry and disgust emotions. By computing wavelet features for different vocalization classes separately, different attributes could be measured, thus contributing positively towards ASD detection.

4.2.4 Modulation features

Auditory-inspired modulation features have been used in the past for pathological speech characterization [69, 50, 70] and speech emotion recognition [49]. High frequency modulations have also been linked to turbulence noise present in no-pain cries

[40], whereas certain cry modulation frequencies have been linked to central nervous system disorders [40]. Additionally, recent research has suggested impaired extraction of speech rhythm from temporal modulation patterns in speech in developmental dyslexia, an impaired neural representation of the sound structure of words typically observed with individuals on the spectrum [71]. The work described herein exemplifies the first attempt at using modulation features for ASD detection. Such findings corroborate the observed fact that the modulation spectral features that contributed the most towards the task at hand were computed from negative emotion, other, and speech classes (see Table 3.8).

4.2.5 Overall accuracy

Table 3.7 shows that wavelet features achieved the best overall accuracy and sensitivity, outperforming the benchmark prosodic features and proposed modulation features. The modulation features, in turn, resulted in the highest specificity. These findings suggest that different features may contribute complementary information to overall ASD detection. The next section discusses the obtained findings aimed at fusing information at the decision and feature levels.

4.3 Feature Fusion

Decision-level fusion is a widely explored method in machine learning and pattern recognition that typically improves the performance over single classifiers [90, 91]. Here, three decision-level schemes were explored. However, as reported in Table 3.10, none of the ensemble methods outperformed the results achieved with a single classifier. Plurality vote, for example, helped improve the specificity of the benchmark prosodic classifier when combined with decisions from the wavelet classifier and from the three

feature classes together, but the overall performance was below that achieved with wavelet and modulation spectral classifiers. Within the maximum probability fusion scheme, in turn, combining decisions from wavelet and modulation feature classifiers helped improve the sensitivity of the overall system, but at a cost of reduced specificity. Similar findings were observed for the average probability vote fusion scheme. From Tables 3.8 and 3.9, it can be seen that the majority of the correct decisions have been made based on negative emotion classes, regardless of the tested feature. As such, fusing decisions of individual classifiers may be overlooking the complementarity of the different feature sets and placing most weight on features that convey similar information. To overcome this potential limitation, feature fusion with feature selection has been explored.

Feature fusion combined with a mutual information-based selection algorithm should be able to sift out top features that convey complementary information from different vocalization classes. Results in Table 3.12 corroborate this hypothesis and show that, after feature fusion, the different classes are contributing to the overall accuracy in a more balanced manner. While negative emotions still play a crucial role, other classes such as babble and laughter have stood out. As expected, by attending to complementary details extracted from the different feature sets, improved overall performance is achieved. Overall, relative to using only prosodic features, improvements of 21% and 60% could be seen in accuracy and specificity, respectively. Relative to using only the wavelet features, improvements of 6.1% and 14.3% could be seen in accuracy and specificity, respectively, with a small drop in sensitivity of 1.7%. Lastly, relative to using only the modulation spectral features, gains of 9.5%, 12.5%, and 6.7% in accuracy, sensitivity, and specificity could be observed, respectively.

Close inspection of the top features reported in Table 3.11 further validate the claim that feature-level fusion has allowed different features to extract complementary information from different vocalization classes. For example, the first formant amplitude (A1_mean) and and spectral tilt (H1A3C_mean) measures have been used within

laughter research [65]. The entropy[1,1]_mean feature, in turn, conveys detail about high frequency entropy, a metric previously related to vocalizations that fall under the 'other' class. The authors in [53] also argue that the vibration of the vocal cords is reflected in the entropy of the wavelet parameters, a common finding reported in the ASD cry literature but computed via the more complex method of fundamental frequency tracking [46]. The energy[1,0]_mean and energy [1,1]_mean features, on the other hand, measure low and high frequency energy, respectively. These features, in turn, have been shown useful in speech vocalization discrimination between ASD and controls, and characterizing squeal quality [44], respectively. Energy variability, characterized by energy[1,0]_std, in turn, has shown useful for discriminating speech from ASD and control subjects from older, verbally fluent kids [66].

Interestingly, of the top 17 features chosen by the selection algorithm, 11 correspond to modulation spectral features. Of the modulation energy features $\epsilon_{i,j}$, all features are from modulation channels higher than 4, corresponding to a modulation frequency greater than 17.6 Hz. It is a well-known fact that, in running speech, modulation frequencies below 16 Hz contribute towards intelligibility [92]. In the case on pre-verbal toddlers, such information is not deemed useful and higher modulation frequencies seem to stand out. Within no-pain cries, higher modulation frequencies have been related to turbulent noise [40]. In another study looking at modulations of cries, frequencies around 10-70Hz were reported and significant differences around 40 Hz were seen for children with brain damage relative to controls [40]. Interestingly, four of the top-selected modulation features convey information about the 6th modulation channel centred near 40 Hz (47.5 Hz, to be more exact). In another recent study, modulation below 20 Hz was shown to significantly differ between children with dyslexia and controls [71]. Three top-features capture such a modulation frequency range near modulation channels 4 and 5. Lastly, the energy distribution feature $\Phi_{1,m}(3)$ has been shown for running speech to be a top discriminative feature to detect sadness and anger emotions

[49]. Such features could be detecting distress cues in cries, a finding that has been widely reported in the ASD literature [37].

Chapter 5

Conclusions and Future Research Directions

Acoustic-prosodic characterization of toddler vocalization utterances has been shown useful for autism spectrum disorder diagnosis. Existing studies have typically explored irregularities during vocal fold vibration and inappropriate use of the volume in individuals with autism[21, 22, 23, 14, 9, 24, 25, 26, 18]. Given the wide range of age of the participants, however, and the fast changing vocal tract characteristics during childhood, many of the reported findings have been contradictory [14]. Moreover, findings have typically been reported using only speech-like utterances (e.g., babble) [17, 43, 44] or cries [41, 12, 16, 42, 12, 17, 43, 44, 45, 37]; thus it is still not clear which types of vocalization contribute the most to classification. Lastly, a vast body of literature has explored the use of wavelet features for infant cry and pathological speech analysis (e.g., [53, 47, 48, 55]), as well as the speech modulation spectrum (e.g., [49, 50, 51]). To the best of the author's knowledge, however, such features have yet to be explored for ASD diagnosis. This thesis aims to fill these three gaps.

More specifically, we explore the use of wavelet, modulation spectral, and prosodic features to classify forty-three 18-month old toddlers (23 of which were diagnosed as having autism at the 36-months assessment and 20 age-matched control group) into ASD and non-ASD groups. By focusing only on 18-month old data, variability from vocal tract maturation is minimized, thus shedding light on features truly discriminative of ASD. Lastly, we explore the contributions of different vocalization types, namely babble, speech-like, laughter, negative emotions (grouping vocalizations such as cries, whines, squeals, and shouts), and others and explore the effects different features have on certain vocalization types for overall ASD diagnosis.

Overall, it was found that an accuracy of 81.5%, a sensitivity of 91.6% and a specificity of 70% could be achieved with an individual SVM classifier trained on wavelet based energy and entropy features. When trained with speech modulation spectral features, an accuracy of 79%, a sensitivity of 80% and a specificity of 75% could be achieved. These accuracies compared favourably against prosodic features previously proposed in the literature [35]. Moreover, while decision-level fusion did not improve overall performance, feature-level fusion combined with feature selection achieved an accuracy of 86.5%, a sensitivity of 90% and a specificity of 80%, thus representing a relative improvement over individual classifier of 5% and 10% in terms of accuracy and specificity, respectively. Close inspection of the top 17 features selected showed that the most important features corresponded to modulation spectral features. Interestingly, it was observed that vocalizations such as cries, squeals, whines and shouts showed to be more discriminative between classes than speech, babble or laugh vocalizations. Such findings could assist clinicians in future assessments, which currently place focus on prosodic nuances during speech-like utterances.

As for future research directions, the work presented herein can be expanded in two ways: (1) audio recordings from earlier ADOS assessments (e.g., at 12 month of age) can be explored to see if discriminative features can be found even at earlier ages and (2) data from more 18-month old toddlers can be incorporated to validate the obtained

findings, as well as explore alternate solutions. More data, for example, could allow for more balanced vocalization type classes. This would enable new fusion schemes to be explored, such as optimal combination of feature-per-vocalization class fusion. A more balanced vocalization type distribution would also allow for more accurate experiments exploring the effects of each subtype on overall classification.

Such explorations are left for future work due to the amount of labour involved in extracting and manually labelling all vocalization types, as well as discarding those with adult speech overlap. Notwithstanding, with the fast advances seen in machine learning and deep neural networks (DNN), it may be soon that such human laborious tasks may be replaced by machines. Having this said, the initial steps in speech patterns and audio event classification using deep neural networks have already been taken [72, 73] and DNN-based speech recognition of children's speech has already been proposed [74]. Future work could also explore the use of DNN-based segmentation and labelling, thus opening doors to large-scale studies.

Bibliography

[1] T. H. Falk and W.-Y. Chan, "Temporal dynamics for blind measurement of room acoustical parameters," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 4, pp. 978–989, 2010.

[2] A. P. Association, C. on Nomenclature, Statistics *et al.*, *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Association, 1960.

[3] S. Ozonoff, G. S. Young, A. Carter, D. Messinger, N. Yirmiya, L. Zwaigenbaum, S. Bryson, L. J. Carver, J. N. Constantino, K. Dobkins *et al.*, "Recurrence risk for autism spectrum disorders: a baby siblings research consortium study," *Pediatrics*, vol. 128, no. 3, pp. e488–e495, 2011.

[4] D. L. Christensen, "Prevalence and characteristics of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, united states, 2012," *MMWR. Surveillance Summaries*, vol. 65, 2016.

[5] M. Sigman, A. Dijamco, M. Gratier, and A. Rozga, "Early detection of core deficits in autism," *Developmental Disabilities Research Reviews*, vol. 10, no. 4, pp. 221–233, 2004.

[6] E. C. Fenske, S. Zalenski, P. J. Krantz, and L. E. McClannahan, "Age at intervention and treatment outcome for autistic children in a comprehensive intervention

program," *Analysis and Intervention in Developmental Disabilities*, vol. 5, no. 1-2, pp. 49–58, 1985.

[7] K. Sullivan, W. L. Stone, and G. Dawson, "Potential neural mechanisms underlying the effectiveness of early intervention for children with autism spectrum disorder," *Research in Developmental Disabilities*, vol. 35, no. 11, pp. 2921–2932, 2014.

[8] C. Lord, S. Risi, L. Lambrecht, E. H. Cook Jr, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, "The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism," *Journal of Autism and Developmental Disorders*, vol. 30, no. 3, pp. 205–223, 2000.

[9] D. Bone, S. Bishop, R. Gupta, S. Lee, and S. Narayanan, "Acoustic-prosodic and turn-taking features in interactions with children with neurodevelopmental disorders," *Interspeech 2016*, pp. 1185–1189, 2016.

[10] M. G. Filipe, S. Frota, S. L. Castro, and S. G. Vicente, "Atypical prosody in asperger syndrome: Perceptual and acoustic measurements," *Journal of Autism and Developmental Disorders*, vol. 44, no. 8, pp. 1972–1981, 2014.

[11] Y. Nakai, R. Takashima, T. Takiguchi, and S. Takada, "Speech intonation in children with autism spectrum disorder," *Brain and Development*, vol. 36, no. 6, pp. 516–522, 2014.

[12] G. Esposito, J. Nakazawa, P. Venuti, and M. H. Bornstein, "Componential deconstruction of infant distress vocalizations via tree-based models: A study of cry in autism spectrum disorder and typical development," *Research in Developmental Disabilities*, vol. 34, no. 9, pp. 2717–2724, 2013.

[13] J. S. Stephen, J. Iverson, and B. M. Lester, "Atypical cry characteristics in infants at risk for autism," 2008.

[14] R. Fusaroli, A. Lambrechts, D. Bang, D. M. Bowler, and S. B. Gaigg, "Is voice a marker for autism spectrum disorder? a systematic review and meta-analysis," *Autism Research*, 2016.

[15] L. Kanner, "Autistic disturbances of affective contact." *Acta paedopsychiatrica*, vol. 35, no. 4, pp. 100–136, 1967.

[16] M. Bornstein, K. Costlow, A. Truzzi, and G. Esposito, "Categorizing the cries of infants with asd versus typically developing infants: A study of adult accuracy and reaction time," *Research in Autism Spectrum Disorders*, vol. 31, pp. 66–72, 2016.

[17] E. Patten, K. Belardi, G. T. Baranek, L. R. Watson, J. D. Labban, and D. K. Oller, "Vocal patterns in infants with autism spectrum disorder: Canonical babbling status and vocalization frequency," *Journal of Autism and Developmental Disorders*, vol. 44, no. 10, pp. 2413–2428, 2014.

[18] C. Lord, P. C. DiLavore, A. Pickles, M. J. Elliot, C. Hellreigel, S. Arnold, and L. Tao, "Pre-linguistic vocalizations and social directedness in autistic, developmentally delayed and typical toddlers," *Infant Behavior and Development*, vol. 19, p. 61, 1996.

[19] J. McCann and S. Peppé, "Prosody in autism spectrum disorders: a critical review," *International Journal of Language & Communication Disorders*, vol. 38, no. 4, pp. 325–350, 2003.

[20] L. D. Shriberg, R. Paul, J. L. McSweeny, A. Klin, D. J. Cohen, and F. R. Volkmar, "Speech and prosody characteristics of adolescents and adults with high-functioning autism and asperger syndrome," *Journal of Speech, Language, and Hearing Research*, vol. 44, no. 5, pp. 1097–1115, 2001.

[21] J. Quigley, S. McNally, and S. Lawson, "Prosodic patterns in interaction of low-risk and at-risk-of-autism spectrum disorders infants and their mothers at 12 and 18 months," *Language Learning and Development*, vol. 12, no. 3, pp. 295–310, 2016.

[22] J. Parish-Morris, M. Liberman, N. Ryant, C. Cieri, L. Bateman, E. Ferguson, and R. T. Schultz, "Exploring autism spectrum disorders using hlt." in *CLPsych@ HLT-NAACL*, 2016, pp. 74–84.

[23] K. Kary, L. Chan, K. Carol, and S. To, "Do individuals with high-functioning autism who speak a tone language show intonation deficits?" *Journal of autism and developmental disorders*, vol. 46, no. 5, p. 1784, 2016.

[24] Y. S. Bonneh, Y. Levanon, O. Dean-Pardo, L. Lossos, and Y. Adini, "Abnormal speech spectrum and increased pitch variability in young autistic children," *Frontiers in human neuroscience*, vol. 4, p. 237, 2011.

[25] D. Bone, C.-C. Lee, M. P. Black, M. E. Williams, S. Lee, P. Levitt, and S. Narayanan, "The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 4, pp. 1162–1177, 2014.

[26] D. M. Ricks and L. Wing, "Language, communication, and the use of symbols in normal and autistic children," *Journal of autism and childhood schizophrenia*, vol. 5, no. 3, pp. 191–221, 1975.

[27] R. Paul, A. Augustyn, A. Klin, and F. R. Volkmar, "Perception and production of prosody by speakers with autism spectrum disorders," *Journal of autism and developmental disorders*, vol. 35, no. 2, pp. 205–220, 2005.

[28] S. Peppé, J. McCann, F. Gibbon, A. O'Hare, and M. Rutherford, "Receptive and expressive prosodic ability in children with high-functioning autism," *Journal of Speech, Language, and Hearing Research*, vol. 50, no. 4, pp. 1015–1028, 2007.

[29] H. Tager-Flusberg, "A psycholinguistic perspective on language development in the autistic child," *Autism: Nature, Diagnosis, and Treatment*, pp. 92–115, 1989.

[30] J. Demouy, M. Plaza, J. Xavier, F. Ringeval, M. Chetouani, D. Perisse, D. Chauvin, S. Viaux, B. Golse, D. Cohen *et al.*, "Differential language markers of pathology in autism, pervasive developmental disorder not otherwise specified and specific language impairment," *Research in Autism Spectrum Disorders*, vol. 5, no. 4, pp. 1402–1412, 2011.

[31] J. J. Diehl and R. Paul, "Acoustic and perceptual measurements of prosody production on the profiling elements of prosodic systems in children by children with autism spectrum disorders," *Applied Psycholinguistics*, vol. 34, no. 1, pp. 135–161, 2013.

[32] K. Hubbard and D. A. Trauner, "Intonation and emotion in autistic spectrum disorders," *Journal of Psycholinguistic Research*, vol. 36, no. 2, pp. 159–173, 2007.

[33] L. M. Morett, K. O'Hearn, B. Luna, and A. S. Ghuman, "Altered gesture and speech production in asd detract from in-person communicative quality," *Journal of Autism and Developmental Disorders*, vol. 46, no. 3, p. 998, 2016.

[34] M. Asgari, A. Bayestehtashk, and I. Shafran, "Robust and accurate features for detecting and diagnosing autism spectrum disorders." in *Interspeech*, 2013, pp. 191–194.

[35] J. F. Santos, N. Brosh, T. H. Falk, L. Zwaigenbaum, S. E. Bryson, W. Roberts, I. M. Smith, P. Szatmari, and J. A. Brian, "Very early detection of autism spectrum disorders based on acoustic analysis of pre-verbal vocalizations of 18-month old toddlers," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7567–7571.

[36] G. Esposito and P. Venuti, "Understanding early communication signals in autism: a study of the perception of infants' cry," *Journal of Intellectual Disability Research*, vol. 54, no. 3, pp. 216–223, 2010.

[37] G. Esposito, M. del Carmen Rostagno, P. Venuti, J. D. Haltigan, and D. S. Messinger, "Brief report: Atypical expression of distress during the separation phase of the strange situation procedure in infant siblings at high risk for asd," *Journal of Autism and Developmental Disorders*, vol. 44, no. 4, pp. 975–980, 2014.

[38] P. K. Kuhl and A. N. Meltzoff, "Infant vocalizations in response to speech: Vocal imitation and developmental change," *The journal of the Acoustical Society of America*, vol. 100, no. 4, pp. 2425–2438, 1996.

[39] K. Wermke, W. Mende, C. Manfredi, and P. Bruscaglioni, "Developmental aspects of infant's cry melody and formants," *Medical Engineering & Physics*, vol. 24, no. 7, pp. 501–514, 2002.

[40] W. Mende, K. Wermke, S. Schindler, K. Wilzopolski, and S. Hock, "Variability of the cry melody and the melody spectrum as indicators for certain cns disorders," *Early Child Development and Care*, vol. 65, no. 1, pp. 95–107, 1990.

[41] F. B. Furlow, "Human neonatal cry quality as an honest signal of fitness," *Evolution and Human Behavior*, vol. 18, no. 3, pp. 175–193, 1997.

[42] G. Esposito and P. Venuti, "Comparative analysis of crying in children with autism, developmental delays, and typical development," *Focus on Autism and Other Developmental Disabilities*, vol. 24, no. 4, pp. 240–247, 2009.

[43] S. J. Sheinkopf, P. Mundy, D. K. Oller, and M. Steffens, "Vocal atypicalities of preverbal autistic children," *Journal of Autism and Developmental Disorders*, vol. 30, no. 4, pp. 345–354, 2000.

[44] D. K. Oller, P. Niyogi, S. Gray, J. Richards, J. Gilkerson, D. Xu, U. Yapanel, and S. Warren, "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development," *Proceedings of the National Academy of Sciences*, vol. 107, no. 30, pp. 13 354–13 359, 2010.

[45] G. Esposito and P. Venuti, "Developmental changes in the fundamental frequency (f0) of infants' cries: A study of children with autism spectrum disorder," *Early Child Development and Care*, vol. 180, no. 8, pp. 1093–1102, 2010.

[46] S. J. Sheinkopf, J. M. Iverson, M. L. Rinaldi, and B. M. Lester, "Atypical cry acoustics in 6-month-old infants at risk for autism spectrum disorder," *Autism Research*, vol. 5, no. 5, pp. 331–339, 2012.

[47] M. Hariharan, S. Yaacob, and S. A. Awang, "Pathological infant cry analysis using wavelet packet transform and probabilistic neural network," *Expert Systems with Applications*, vol. 38, no. 12, pp. 15 377–15 382, 2011.

[48] J. Saraswathy, M. Hariharan, V. Vijean, S. Yaacob, and W. Khairunizam, "Performance comparison of daubechies wavelet family in infant cry classification," in *2012 IEEE 8th International Colloquium on Signal Processing and its Applications (CSPA)*. IEEE, 2012, pp. 451–455.

[49] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication*, vol. 53, no. 5, pp. 768–785, 2011.

[50] M. Markaki and Y. Stylianou, "Voice pathology detection and discrimination based on modulation spectral features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1938–1948, 2011.

[51] N. Malyska, T. F. Quatieri, and D. Sturim, "Automatic dysphonia recognition using biologically-inspired amplitude-modulation features," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05)*, vol. 1. IEEE, 2005, pp. I–873.

[52] S. Georgiades, P. Szatmari, L. Zwaigenbaum, S. Bryson, J. Brian, W. Roberts, I. Smith, T. Vaillancourt, C. Roncadin, and N. Garon, "A prospective study of

autistic-like traits in unaffected siblings of probands with autism spectrum disorder," *JAMA Psychiatry*, vol. 70, no. 1, pp. 42–48, 2013.

[53] R. Behroozmand and F. Almasganj, "Optimal selection of wavelet-packet-based features using genetic algorithm in pathological assessment of patients' speech signal with unilateral vocal fold paralysis," *Computers in Biology and Medicine*, vol. 37, no. 4, pp. 474–485, 2007.

[54] C. S. Burrus, R. A. Gopinath, and H. Guo, "Introduction to wavelets and wavelet transforms: a primer," 1997.

[55] Y. Huang, A. Wu, G. Zhang, and Y. Li, "Speech emotion recognition based on coiflet wavelet packet cepstral coefficients," in *Chinese Conference on Pattern Recognition*. Springer, 2014, pp. 436–443.

[56] I. Ree, "P. 56: Objective measurement of active speech level," 1993.

[57] M. Slaney *et al.*, "An efficient implementation of the patterson-holdsworth auditory filter bank," *Apple Computer, Perception Group, Tech. Rep*, vol. 35, no. 8, 1993.

[58] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[59] B. C. Ross, "Mutual information between discrete and continuous data sets," *PloS one*, vol. 9, no. 2, p. e87357, 2014.

[60] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, p. 066138, 2004.

[61] M. H. Farouk, *Application of wavelets in speech processing*. Springer, 2014.

[62] J. Saraswathy, M. Hariharan, T. Nadarajaw, W. Khairunizam, and S. Yaacob, "Optimal selection of mother wavelet for accurate infant cry classification," *Australasian Physical & Engineering Sciences in Medicine*, vol. 37, no. 2, pp. 439–456, 2014.

[63] Y. Long, L. Gang, and G. Jun, "Selection of the best wavelet base for speech signal," in *Intelligent Multimedia, Video and Speech Processing, 2004. Proceedings of 2004 International Symposium on.* IEEE, 2004, pp. 218–221.

[64] S. Orlandi, L. Bocchi, C. Manfredi, M. Puopolo, A. Guzzetta, S. Vicari, and M. L. Scattoni, "Study of cry patterns in infants at high risk for autism." in *MAVEBA*, 2011, pp. 7–10.

[65] C. Menezes and Y. Igarashi, "The speech laugh spectrum," *Proc. Speech Production, Brazil*, pp. 157–164, 2006.

[66] D. Bone, M. P. Black, C.-C. Lee, M. E. Williams, P. Levitt, S. Lee, and S. Narayanan, "Spontaneous-speech acoustic-prosodic features of children with autism and the interacting psychologist." in *InterSpeech*, 2012, pp. 1043–1046.

[67] V. N. Degaonkar and S. D. Apte, "Emotion modeling from speech signal based on wavelet packet transform," *International Journal of Speech Technology*, vol. 16, no. 1, pp. 1–5, 2013.

[68] H. K. Palo and M. N. Mohanty, "Wavelet based feature combination for recognition of emotions," *Ain Shams Engineering Journal*, 2017.

[69] T. H. Falk, W.-Y. Chan, and F. Shein, "Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility," *Speech Communication*, vol. 54, no. 5, pp. 622–631, 2012.

[70] M. Markaki and Y. Stylianou, "Normalized modulation spectral features for cross-database voice pathology detection," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[71] V. Leong and U. Goswami, "Impaired extraction of speech rhythm from temporal modulation patterns in speech in developmental dyslexia," *Frontiers in human neuroscience*, vol. 8, 2014.

[72] L. Deng and D. Yu, "Deep convex net: A scalable architecture for speech pattern classification," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[73] Z. Kons and O. Toledo-Ronen, "Audio event classification using deep neural networks." in *INTERSPEECH*, 2013, pp. 1482–1486.

[74] M. Matassoni, D. Falavigna, and D. Giuliani, "Dnn adaptation for recognition of children speech through automatic utterance selection," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 644–651.

[75] C. Lord, M. Rutter, and A. Le Couteur, "Autism diagnostic interview-revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders," *Journal of autism and developmental disorders*, vol. 24, no. 5, pp. 659–685, 1994.

[76] S. Mallat, *A wavelet tour of signal processing*. Academic press, 1999.

[77] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[78] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.

[79] H. Rothgänger, "Analysis of the sounds of the child in the first year of age and a comparison to the language," *Early Human Development*, vol. 75, no. 1, pp. 55–69, 2003.

[80] E. Schoen, R. Paul, and K. Chawarska, "Phonology and vocal behavior in toddlers with autism spectrum disorders," *Autism Research*, vol. 4, no. 3, pp. 177–188, 2011.

[81] H.-C. Hsu, A. Fogel, and R. B. Cooper, "Infant vocal development during the first 6 months: Speech quality and melodic complexity," *Infant and Child Development*, vol. 9, no. 1, pp. 1–16, 2000.

[82] C. Papaeliou, G. Minadakis, and D. Cavouras, "Acoustic patterns of infant vocalizations expressing emotions and communicative functions," *Journal of Speech, Language, and Hearing Research*, vol. 45, no. 2, pp. 311–317, 2002.

[83] S. Nakazima, "A comparative study of the speech developments of japanese and american english in childhood (1): A comparison of the developments of voices at the prelinguistic period," 1962.

[84] I. Van den Dikkenberg-Pot, F. Koopmans-van Beinum, and C. Clement, "Influence of lack of auditory speech perception on sound productions of deaf infants," in *Proceedings of the Institute of Phonetic Sciences Amsterdam*, vol. 22, 1998, pp. 47–60.

[85] N. Masataka, "Why early linguistic milestones are delayed in children with williams syndrome: late onset of hand banging as a possible rate–limiting constraint on the emergence of canonical babbling," *Developmental Science*, vol. 4, no. 2, pp. 158–164, 2001.

[86] H. Rao, J. C. Kim, A. Rozga, and M. A. Clements, "Detection of laughter in children's speech using spectral and prosodic acoustic features." in *INTERSPEECH*, 2013, pp. 1399–1403.

[87] T. J. Runyon, "Function of laughter from a student with autism," Ph.D. dissertation, San Francisco State University, 2015.

[88] J. Parvizi, S. W. Anderson, C. O. Martin, H. Damasio, and A. R. Damasio, "Pathological laughter and crying: a link to the cerebellum," *Brain*, vol. 124, no. 9, pp. 1708–1719, 2001.

[89] J. Parvizi, D. B. Arciniegas, G. L. Bernardini, M. W. Hoffmann, J. P. Mohr, M. J. Rapoport, J. D. Schmahmann, J. M. Silver, and S. Tuhrim, "Diagnosis and management of pathological laughter and crying," in *Mayo Clinic Proceedings*, vol. 81, no. 11. Elsevier, 2006, pp. 1482–1486.

[90] D. W. Opitz and R. Maclin, "Popular ensemble methods: An empirical study," *J. Artif. Intell. Res.(JAIR)*, vol. 11, pp. 169–198, 1999.

[91] N. M. Baba, M. Makhtar, S. A. Fadzli, and M. K. Awang, "Current issues in ensemble methods and its applications," *Journal of Theoretical and Applied Information Technology*, vol. 81, no. 2, p. 266, 2015.

[92] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, no. 5, pp. 2670–2680, 1994.