

Novel paradigms in the processing of speech and its disorders

Khalid Daoudi

▶ **To cite this version:**

Khalid Daoudi. Novel paradigms in the processing of speech and its disorders. Computer Science [cs]. Université de bordeaux, 2021. tel-03884101

HAL Id: tel-03884101
https://inria.hal.science/tel-03884101

Submitted on 4 Dec 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

université de BORDEAUX

université de BORDEAUX

MÉMOIRE POUR OBTENIR L'

HABILITATION À DIRIGER LES RECHERCHES

délivrée par **UNIVERSITÉ DE BORDEAUX**

ÉCOLE DOCTORALE MATHÉMATIQUES ET INFORMATIQUE

SPÉCIALITÉ : **INFORMATIQUE**

présenté par **Khalid DAOUDI**

Nouveaux paradigmes en traitement de la parole et de ses troubles

Soutenue le ??/03/2021

Membres du jury :

VALÉE, Nathalie	Directrice de Recherche	Rapporteure
LAPRIE, Yves	Directeur de Recherche	Rapporteur
SICARD, Etienne	Professeur	Rapporteur
ANDRÉ-OBRECHT Régine	Professeure	Examinatrice
BONASTRE, Jean-François	Professeur	Examinateur
ZIOU, Djemel	Professeur	Examinateur

Table des matières

1 Résumé étendu de mes activités de recherche **10**

1.1 Parcours professionnel . . . 11

1.2 Thématiques de recherche . . . 12

1.2.1 Thématiques avant mon recrutement à Inria . . . 13

1.2.2 Thématiques à Inria-Lorraine . . . 13

1.2.3 Thématiques à IRIT . . . 14

1.2.4 Thématiques à Inria Bordeaux Sud-Ouest . . . 15

1.2.4.1 Analyse non-linéaire et parole pathologique . . . 15

1.2.4.2 Imagerie satellitaire océanographique : Upwelling côtier . . . 17

1.3 Encadrement de doctorants . . . 18

1.3.1 Thèse de Murat Deviren . . . 18

1.3.1.1 Impact . . . 19

1.3.1.2 Parcours après la thèse . . . 19

1.3.2 Thèse de Jérôme Louradour . . . 19

1.3.2.1 Impact . . . 20

1.3.2.2 Parcours après la thèse . . . 20

1.3.3 Thèse de Reda Jourani . . . 20

1.3.3.1 Impact . . . 21

1.3.3.2 Parcours après la thèse . . . 21

1.3.4 Thèse de Vahid Khanagha . . . 22

1.3.4.1 Impact . . . 22

1.3.4.2 Parcours après la thèse . . . 23

1.3.5 Thèse de Biswajit Das . . . 23

1.3.6 Thèse de Ayoub Tamim . . . 24

1.3.6.1 Impact . . . 25
1.3.6.2 Parcours après la thèse . . . 25
1.3.7 Thèse de Anass El Aouni . . . 25
1.3.7.1 Impact . . . 26
1.3.7.2 Parcours après la thèse . . . 26
1.4 Autres encadrements . . . 27
1.4.1 Encadrement de postdoc . . . 27
1.4.2 Encadrement de stages . . . 27
1.5 Management . . . 28
1.5.1 Création de l'équipe de recherche Inria GEOSTAT . . . 28
1.5.2 Projets de recherche . . . 29
1.6 Transferts technologiques . . . 30
1.6.1 Débruitage d'images de grands scanners . . . 30
1.6.2 Caractérisation de l'upwelling . . . 31
1.6.3 Reconnaissance automatique des émotions dans la parole . . . 31
1.7 Prix et distinctions . . . 31
1.8 Responsabilités collectives . . . 32
1.9 Enseignement . . . 32
1.10 Organisation du mémoire . . . 32

2 Dynamic Bayesian networks for speech recognition 34
2.1 Introduction . . . 35
2.2 Bayesian networks . . . 37
2.3 Inference algorithms for Bayesian networks . . . 40
2.3.1 Construction of the junction tree . . . 40
2.3.2 Propagation of evidence in the junction tree . . . 42
2.4 Multi-band speech recognition : the classical approach . . . 45
2.5 Multi-band speech recognition : the DBNs perspective . . . 47
2.5.1 Model definition . . . 47
2.5.2 Construction of the junction tree . . . 50
2.5.3 Model parameters estimation . . . 51
2.6 Application to isolated speech recognition . . . 54
2.7 Application to continuous speech recognition . . . 57

2.7.1 Decoding algorithm . . . 58
2.7.2 Experiments . . . 59

3 Sequence kernels for SVM speaker recognition 64
3.1 Introduction . . . 65
3.2 On sequence kernels . . . 67
3.2.1 Sequence kernels based on generative models . . . 67
3.2.2 Other sequence kernels . . . 68
3.2.3 Overview of the GLDS kernel . . . 69
3.3 FSNS kernels . . . 70
3.3.1 Definition . . . 70
3.3.2 Dual/Kernelized form of FSNS kernels . . . 73
3.3.2.1 Relationship with Gaussian processes . . . 73
3.3.2.2 Factorized form . . . 74
3.3.2.3 Computational complexity . . . 75
3.4 Incomplete Cholesky Decomposition of the Gram Matrix . . . 77
3.5 Mahalanobis kernels in the feature space . . . 80
3.5.1 Definition . . . 80
3.5.2 Dual/Kernelized form of the FSMS kernels . . . 82
3.6 Experiments . . . 83
3.6.1 Database and front-end processing . . . 83
3.6.2 System implementation . . . 84
3.6.3 Development of SVM system with FSNS/FSMS kernels . . . 84
3.6.3.1 Kernel normalization and regularization strategies . . . 85
3.6.3.2 Choice of the codebook . . . 85
3.6.3.3 Choice of the vector kernel k . . . 86
3.6.3.4 Score normalization . . . 88
3.6.4 Evaluation . . . 89

4 Nonlinear speech processing 92
4.1 The production mechanism of the speech signal . . . 93
4.2 Non-linear character of the speech signal . . . 95
4.3 Speech as a realization of a non-linear dynamical system . . . 97
4.4 The Microcanonical Multiscale Formalism . . . 100

4.4.1 Singularity exponents . . . 101
4.4.2 The Most Singular Manifold . . . 103
4.4.3 Estimation of singularity exponents . . . 104
4.4.3.1 The choice of $\Gamma_r(\cdot)$ . . . 104
4.4.3.2 Estimation of $h(t)$ . . . 105
4.5 Application to glottal closure instants detection . . . 106
4.5.1 The significant excitation of the vocal tract . . . 106
4.5.2 Review of existing methods . . . 107
4.5.3 The relationship between MSM and GCIs . . . 111
4.5.4 MSM-based GCI detection . . . 113
4.5.5 Experimental results . . . 116
4.5.5.1 Performance measures . . . 118
4.5.5.2 Clean speech . . . 120
4.5.5.3 Noisy speech . . . 120
4.5.5.4 Computational complexity . . . 124
4.6 Application to sparse linear prediction . . . 125
4.6.1 Sparse linear prediction . . . 126
4.6.2 Approximation of the l_0-norm . . . 127
4.6.3 The weighted l_2-norm solution . . . 130
4.6.3.1 Optimization algorithm . . . 131
4.6.3.2 The weighting function . . . 132
4.6.4 Experimental results . . . 133
4.6.4.1 Voiced sound analysis . . . 134
4.6.4.2 Estimation of the all-pole vocal-tract filter . . . 136
4.6.4.3 Multi-pulse excitation estimation . . . 136

5 Dysarthric speech processing **139**
5.1 Introduction . . . 140
5.2 Parkinsonism . . . 141
5.2.1 Parkinson's disease . . . 141
5.2.2 Multiple system atrophy . . . 141
5.2.3 Progressive supranuclear palsy . . . 142
5.3 The challenge of differential diagnosis . . . 143

5.4 Dysarthria . . . 143
5.4.1 Hypokinetic dysarthria . . . 145
5.4.1.1 Perceptual evaluation . . . 145
5.4.1.2 Objective evaluation . . . 146
5.4.2 Spastic dysarthria . . . 147
5.4.2.1 Perceptual evaluation . . . 148
5.4.2.2 Objective evaluation . . . 148
5.4.3 Ataxic dysarthria . . . 149
5.4.3.1 Perceptual evaluation . . . 150
5.4.3.2 Objective evaluation . . . 150
5.5 Dysarthria-based differential diagnosis . . . 151
5.6 The Voice4PD-MSA project . . . 154
5.6.1 The consortium . . . 155
5.6.2 Beyond Voice4PD-MSA . . . 157
5.6.3 The challenge of data collection . . . 159
5.7 Differential diagnosis between MSA and PSP . . . 159
5.7.1 Dataset . . . 160
5.7.2 Acoustic features . . . 161
5.7.3 Methodology and experiments . . . 161
5.7.3.1 Univariate statistical analysis . . . 163
5.7.3.2 Learning a new speech feature . . . 163
5.7.4 Discussion and conclusion . . . 166
5.8 Distortion of voiced obstruents for differential diagnosis between PD and MSA-P . . . 167
5.8.1 Introduction . . . 167
5.8.2 Database . . . 169
5.8.3 Method and results . . . 169
5.8.3.1 Devoicing analysis . . . 170
5.8.3.2 VOT analysis of voiced plosives . . . 171
5.8.3.3 Classification . . . 173
5.8.4 Discussion and conclusion . . . 174

Conclusion et perspectives **177**

I . Parole et troubles respiratoites . . . 178
I . I . Le projet VocaPnée . . . 178
I . II . Justification scientifique . . . 180
I . II ..1 La respiration . . . 180
I . II ..2 Parole et respiration . . . 180
I . II ..3 La dyspnée . . . 182
I . II ..4 Déficit de perception de la dyspnée . . . 182
I . II ..5 Biomarqueurs physiologiques de la dyspnée, biomarqueurs vocaux . . . 183
I . III . Méthodologie . . . 183
II . Étude la dysarthrie . . . 185
II . I . Poursuite de Voice4PD-MSA . . . 185
II . II . Passage à la médecine de ville . . . 185
II . III . Méthodologie . . . 186

Bibliographie générale **189**

Bibliographie personnelle **216**

A . Revues internationales . . . 217
A . I . Réseaux Bayésiens et RAP . . . 217
A . II . Reconnaissance automatique du locuteur . . . 217
A . III . Analyse non-linéaire de la parole . . . 218
A . IV . Upwelling . . . 218
A . V . Analyse multi-fractale . . . 219
A . VI . Divers . . . 220
B . Conférences internationales avec comité de lecture . . . 220
B . I . Réseaux Bayésiens et RAP . . . 220
B . II . Reconnaissance automatique du locuteur . . . 223
B . III . Analyse non-linéaire de la parole . . . 225
B . IV . Upwelling . . . 227
B . V . Analyse multi-fractale . . . 228
B . VI . Divers . . . 229
C . Chapitres de livres . . . 230

D . Autres publications internationales (posters, short papers) 231
E . Conférences nationales avec comité de lecture 232

Chapitre 1

Résumé étendu de mes activités de recherche

Sommaire

1.1 Parcours professionnel 11

1.2 Thématiques de recherche 12

1.2.1 Thématiques avant mon recrutement à Inria 13

1.2.2 Thématiques à Inria-Lorraine 13

1.2.3 Thématiques à IRIT 14

1.2.4 Thématiques à Inria Bordeaux Sud-Ouest 15

1.3 Encadrement de doctorants 18

1.3.1 Thèse de Murat Deviren 18

1.3.2 Thèse de Jérôme Louradour 19

1.3.3 Thèse de Reda Jourani 20

1.3.4 Thèse de Vahid Khanagha 22

1.3.5 Thèse de Biswajit Das 23

1.3.6 Thèse de Ayoub Tamim 24

1.3.7 Thèse de Anass El Aouni 25

1.4 Autres encadrements 27

1.4.1 Encadrement de postdoc 27
1.4.2 Encadrement de stages 27
1.5 Management 28
1.5.1 Création de l'équipe de recherche Inria GEOSTAT 28
1.5.2 Projets de recherche 29
1.6 Transferts technologiques 30
1.6.1 Débruitage d'images de grands scanners 30
1.6.2 Caractérisation de l'upwelling 31
1.6.3 Reconnaissance automatique des émotions dans la parole . . . 31
1.7 Prix et distinctions 31
1.8 Responsabilités collectives 32
1.9 Enseignement 32
1.10 Organisation du mémoire 32

Ce chapitre présente une description générale de mes principales activités de recherche depuis ma thèse. Je commence par décrire mon parcours professionnel puis je présente une description résumée des principales thématiques de recherche que j'ai abordées durant ma carrière. Je présente ensuite les différents encadrements que j'ai effectués puis mes activités de management et de gestion de projets scientifiques. Je termine par mes actions de transfert technologique.

1.1 Parcours professionnel

Je commence par présenter en quelques dates charnières les principales étapes de mon parcours professionnel.

— Nov 1996 : Thèse de doctorat de l'Université Paris 9 Dauphine, intitulée "Généralisations des systèmes de fonctions itérées et Applications au Traitement du Signal", sous la direction de Jacques Lévy-Véhél et Yves Meyer. J'ai préparée ma thèse au sein de l'équipe "Fractales" d'Inria-Rocquencourt.

— Déc 1996 - Nov 1997 : Séjour postdoctoral avec Claude Tricot au département de mathématiques de l'école polytechnique de Montréal.

— Déc 1997 - Sep 1999 : Séjour postdoctoral avec Alan Willsky au Stochastic Systems Group (SSG) du Laboratory for Information and Decision Systems (LIDS), Massachusetts Institute of Technology (MIT), USA.

— Oct 1999 : Recrutement CR2 à l'institut National de Recherche en Informatique et Automatique (INRIA).

— Oct 1999 - Fév 2004 : Membre permanent de l'équipe Parole d'Inria-Lorraine.

— Mars 2004 - Fév 2009 : Mise à disposition puis détachement au CNRS au sein de l'équipe Samova de l'Institut de Recherche en Informatique de Toulouse (IRIT).

— Mar 2009 : Co-création, avec Hussein Yahia, de l'équipe GeoStat à Inria Bordeaux Sud-Ouest.

— Depuis Mars 2009 : Membre permanent de l'équipe GeoStat.

Chacune de ces dates correspond non seulement une mobilité géographique mais aussi à une mobilité thématique importante. Ces mobilités ont été pour moi une source de renouveaux, parfois aussi de rupture, mais que je juge aujourd'hui formatrice et souhaitable à l'apprentissage, puis à la vivification d'une carrière de chercheur. J'ai donc vécu l'ensemble de ce parcours comme une formidable opportunité de satisfaire une curiosité scientifique en diversifiant mon travail, ses enjeux, et à chaque fois, comme le besoin de remettre en question les objectifs de ma mission de chercheur.

1.2 Thématiques de recherche

Depuis mon recrutement à Inria, le fil conducteur de mes activités est un quête scientifique, hors sentiers battus, à la recherche de nouveaux paradigmes et formalismes capables de faire une percée majeure en traitement automatique de la parole (TAP). Cette quête a été motivé par le constat que plusieurs domaines du TAP étaient arrivés à saturation et qu'il fallait passer à une nouvelle génération d'outils de traitement. Ma mobilité

thématique correspond aux différentes phases de cette quête ainsi qu'à ma mobilité géographique.

Note : Ma bibliographie personnelle est organisée en sections et sous-sections correspondants aux différentes thématiques que je décris dans cette section.

1.2.1 Thématiques avant mon recrutement à Inria

Mes recherches lors de ma thèse et de mon premier postodoc ont porté sur l'analyse multi-fractale et ses applications en traitement du signal. Je me suis ensuite intéressé aux processus stochastiques et au traitement statistique du signal. J'ai ainsi rejoint le Stochastic Systems Group du MIT où j'ai travaillé sur les modèles auto-régressifs multi-échelles (MAR) pour lesquels j'ai développé le chaînon manquant pour établir un lien avec l'analyse en ondelettes. Les MAR étant des réseaux Bayésiens sur des arbres, ce travail m'a permis de mûrir l'idée d'introduire les réseaux Bayésiens dynamiques (DBN) en modélisation de la parole, qui est devenu ensuite mon programme de recherche pour le concours Inria.

1.2.2 Thématiques à Inria-Lorraine

En terme de modélisation, les modèles de Markov cachés (HMM) régnaient sans partage sur la technologie en reconnaissance automatique de la parole (RAP). Il se trouve que les HMM sont un cas très particulier des DBN. J'ai ainsi proposé de franchir un cap en RAP en utilisant les DBN pour s'affranchir de certaines hypothèses imposées par les HMM et pour pouvoir modéliser les différents aspects multi-échelles présents dans le signal de parole. Ceci impliquait implicitement la nécessité de franchir un cap aussi en paramétrisation du signal de parole, c'est-à-dire, passer outre la représentation cepstrale.

Je me suis ainsi lancé dans le projet ambitieux de s'attaquer à toutes les composantes qui constituent un système de RAP : la modélisation acoustique, la paramétrisation, la robustesse au bruit et la modélisation du langage. Avec mes collaborateurs, nous avons pu proposer de nouvelles approches novatrices et originales dans ces 4 composantes, non

seulement d'un point de vue fondamentale mais aussi en tenant toujours compte des contraintes de la complexité algorithmique. Nous avons ainsi proposé une nouvelle vision pour faire la RAP qui permet, dans certains cas, d'atteindre des performances de reconnaissance supérieures aux systèmes de RAP classiques.

Cependant, les nouvelles méthodologies que nous avons proposées n'étaient pas suffisantes pour prétendre lever le verrou technologie HMM-Cepstre. Elles pouvaient conduire à certaines améliorations mais leur complexité algorithmique restait relativement élevée. Au bout de 4 années de recherche dans ce domaine, je suis arrivé à la conviction que cette approche n'était pas la bonne pour atteindre mon objectif initial : poser les fondations d'une nouvelle génération de systèmes de RAP. Le futur m'a donné raison car c'est tout récemment que les réseaux de neurones profonds ont réalisé une percé majeure en RAP.

1.2.3 Thématiques à IRIT

Partant de cette conviction, j'ai rejoint IRIT pour travailler sur un autre domaine du TAP, la reconnaissance automatique du locuteur (RAL). La méthode de classification qui régnait dans ce domaine était les SVM (Support Vector Machines). Cependant, cette approche ne permettait pas de d'intégrer directement/naturellement l'aspect dynamique du signal. J'ai ainsi travaillé sur le développement de noyaux entre séquences de vecteurs pour la classification SVM et son application en RAL. D'un point de vue théorique, ces nouveaux noyaux ont eu un écho très positif dans la communauté Machine Learning. Du point de vue pratique, les évaluations conduites sur des tâches benchmark (compagnes NIST) ont montré une réelle amélioration par rapport aux systèmes SVM classiques, tant sur le plan de la précision que sur le plan de l'efficacité algorithmique. Après ces travaux, je me suis intéressé à une approche émergente de classification discriminante qui semblait être prometteuse, les mélanges Gaussiens à large marge (LM-GMM). Contrairement aux SVM, un LM-GMM construit une frontière non-linéaire entre les classes directement dans l'espace des données. Ces travaux ont été intéressants d'un point de vue théorique mais n'ont pas conduit à des améliorations significatives par rapport aux SVM, dans le cadre

de la RAL.

1.2.4 Thématiques à Inria Bordeaux Sud-Ouest

1.2.4.1 Analyse non-linéaire et parole pathologique

Toutes ces années de recherche m'ont m'ont montré qu'il était difficile de faire des avancées majeures (du type levage de verrou) en RAP, en RAL et en traitement de la parole en général. Je suis arrivé à la conviction qu'une des raisons principale est la quasi homogénéité des méthodes d'extraction de l'information du signal de parole (la paramétrisation). En effet, cette dernière est presque toujours basée sur le modèle linéaire source-filtre de la parole. Ainsi la majorités des méthodes de paramétrisation sont essentiellement linéaires. Ce succès peut s'expliquer par la simplicité et l'efficacité algorithmique de ces méthodes. Cependant, il est établit que plusieurs phénomènes non-linéaire sont présents dans la production et perception de la parole, notamment pathologique, que les méthodes linéaires ne peuvent pas décrire. Je me suis ainsi intéressé à l'analyse non-linéaire de la parole pour tenter d'aller outre le modèle source-filtre. Je ne suis évidemment pas le premier à avoir eu ce type d'intérêt, j'ai donc tenté de comprendre l'absence d'impact des approches non-linéaires proposées dans littérature. J'ai conclu qu'une des raison principale est leur difficulté théorique et algorithmique ainsi que le manque de consensus sur la validité des résultats obtenus par ces méthodes, essentiellement basées sur la théorie des systèmes dynamiques non-linéaires. Cette réflexion est en fait à l'origine (en partie) de ma collaboration avec H. Yahia pour créer l'équipe GeoStat. Je me suis ainsi lancé dans une nouvelle aventure pour la création d'une équipe de recherche Inria dédiée à l'analyse de signaux naturels complexes (comme la parole) par de nouvelles approches issus de la physique statistique.

Je me suis ensuite lancé sur le projet de recherche exploratoire suivant : utiliser le cadre des représentations multi-échelles et parcimonieuses du signal pour développer des méthodes d'analyse algorithmiquement efficaces et faciles à interpréter. Avec mes collaborateurs, nous avons pu montrer la pertinence de ce cadre pour l'analyse non-linéaire de la

parole et développer plusieurs applications dans lesquelles nous avons apporté des avancées considérables comparé aux méthodes linéaires classiques : une nouvelle méthode de segmentation phonétique indépendante du texte ; un nouveau codeur de la forme d'onde ; un algorithme robuste et précis pour la détection des instants de fermeture glottale ; une solution analytique et stable au problème de la prédiction linéaire parcimonieuse ; un algorithme efficace d'estimation de la source d'excitation du signal de la parole. En outre, les méthodes que nous avons développées ont le grand avantage d'être théoriquement et algorithmiquement beaucoup plus lisibles que celles issues des systèmes dynamiques. D'un point de vue global, nous avons montré qu'il était possible de développer des algorithmes simples et performants hors du cadre dominant, le modèle source-filtre linéaire.

Ces résultats ont été très prometteurs et ont suscité un intérêt considérable, cependant la tâche d'aller plus loin et tenter de proposer une théorie « générale » à la communauté était trop ambitieuse et trop difficile à réaliser à mon échelle. Je me suis alors orienté vers un domaine « niche » dans lequel je pouvais apporter une contribution majeure à mon échelle, la parole pathologique. Cette dernière est en effet un terrain fertile pour l'analyse non-linéaire de la parole en raison des différentes perturbations qui peuvent s'y produire et qui ne peuvent pas être décrites/captées par les approches standards. Je suis ainsi entré en interaction avec le tissu médical locale et régionale, en particulier les neurologues et les phoniatres, pour montrer le potentiel de l'analyse de la parole dans l'aide au diagnostic de maladies neurodégénératives. J'ai pu ainsi former un consortium multidisciplinaire entre Bordeaux et Toulouse : IRI, IMT, CHU-Bordeaux et CHU-Toulouse, avec pour ces 2 derniers des chercheurs neurologues qui sont des sommités mondiales dans le domaine. Le but de cette collaboration est de développer des outils numériques basées sur le traitement de la parole pour l'aide au diagnostic des syndromes Parkinsonniens. J'ai pu obtenir en 2017 un financement conséquent de l'ANR (appel générique) pour mener ce projet de recherche ambitieux (étude toujours en cous). Des résultats préliminaires intéressants ont été obtenus, mais les principaux résultats sont encore à venir, dans le cadre du développement d'un biomarqueur vocal pour l'aide au diagnostic différentiel entre la maladie de Parkinson et la variante Parkinsonniene d'un maladie rare, l'atrophie

multisytématisée.

1.2.4.2 Imagerie satellitaire océanographique : Upwelling côtier

Peu de temps après la création de GeoStat, j'ai ajouté à mes activités une nouvelle thématique de recherche hors du cadre du traitement de la parole, le traitement d'images satellitaires. Grâce à ma collaboration historique avec le Maroc, j'ai pu identifier le potentiel applicatif d'un axe de recherche principal de GeoStat, l'imagerie satellitaire océanographique, pour le secteur de la pêche au Maroc. En effet la côte atlantique marocaine, avec un potentiel halieutique considérable, est parmi les côtes les plus riches en ressources biologiques exploitables. Une bonne gestion de ces ressources nécessite une bonne connaissance des phénomènes physiques et biologiques qui interagissent pour gouverner la dynamique des populations de poissons. En effet, l'évolution de l'écosystème pélagique de cette région est influencée en grande partie par la variabilité spatio-temporelle du phénomène d'upwelling côtier, la remontée d'eaux froides des profondeurs vers la surface de l'océan et près de la côte, principal moteur de la fertilisation de la couche euphotique par son apport en eaux froides et riches en matières minérales nécessaires pour le démarrage de la productivité de l'écosystème marin. J'ai proposé d'étudier le phénomène d'upwelling par une analyse des images staellitaires de la côté atlantique marocaine.

J'ai ensuite initié, mis en place et dirigé un projet R&D à long terme avec les acteurs locaux clés, académiques (Univ. de Rabat www.fsr.ac.ma/lrit et le CNRST www.cnrst.ma), opérationnels (CRTS, crts.gov.ma) et décisionnels (Ministère de la pêche, www.mpm.gov.ma), ainsi qu'en établissant un partenariat avec des acteurs Français importants dans le domaine, académique (LEGO-CNRS, (www.legos.obs-mip.fr) et opérationnel (Mercator-Océan, www.mercator-ocean.fr/). Un tel projet était novateur et ambitieux étant donné que l'importance de l'upwelling et de son étude par imagerie satellitaire n'étaient pas encore solidement encrés localement. La principale difficulté d'ailleurs a été de convaincre les acteurs locaux opérationnels puis décisionnels de la pertinence et du potentiel du projet. Du point de vue scientifique, pour mener à bien ce projet, j'ai dû élargir

mon champ de compétences en acquérant des connaissances (théoriques et pratiques) qui sortent du champ de mes recherches habituelles (la parole).

Cette entreprise m'a conduit à obtenir et diriger 2 projets PHC (Partenariat Hubert Curien) entre la France et le Maroc sur la caractérisation multi-capteurs et le suivi spatio-temporel par imagerie satellitaire de l'upwelling sur la côte atlantique marocaine. Dans ces projets nous avons exploré l'utilisation de l'imagerie satellitaire, physique et biologique, pour appréhender la dynamique spatiale et temporelle de l'upwelling et comprendre ces effets sur les ressources halieutiques. Nous avons obtenu des résultats remarquables qui nous ont permis d'avoir récemment un financement local conséquent du CNRST pour passer à l'étape finale : l'étude de la corrélation entre les données de pêche et la dynamique de l'upwelling, caractérisée par les outils d'analyse que nous avons développés. Ces 2 projets ont aussi permis le recrutement de 2 doctorants qui ont eu chacun un prix de thèse récompensant la qualité de leurs travaux.

Note : Je ne vais pas détailler ces travaux dans la partie scientifique du mémoire de la thèse, pour garder une structuration centrée autour du traitement de la parole qui est ma thématique de recherche principale passée et future.

1.3 Encadrement de doctorants

1.3.1 Thèse de Murat Deviren

Entre 2001 et 2004, j'ai encadré la thèse de Murat Deviren, avec Jean-Paul Haton comme directeur officiel. Cette thèse a porté sur l'introduction et le développement d'une approche novatrice en reconnaissance automatique de la parole (RAP), les réseaux bayésiens dynamiques ou DBN (pour Dynamic Bayesian networks). Ce formalisme permet de généraliser plusieurs techniques probabilistes utilisées en RAP. Nous avons élaboré sur quatre composantes fondamentales d'un système de RAP : la modélisation acoustique, la modélisation du langage, la paramétrisation du signal acoustique et la compensation du bruit. Nous avons proposé des techniques nouvelles dans chacune de ces composante, et

nous avons apporté des perspectives novatrices. La reformulation des modules de modélisation dans ce formalisme, notamment les HMMs, nous a ouvert de nouvelles perspectives inexploitées auparavant. En plus des nouvelles approches pour la modélisation, nous avons également proposé de nouvelles stratégies pour l'extraction des paramètres acoustiques robustes au bruit. Nous avons aussi abordé le problème de robustesse au bruit par adaptation des modèles acoustiques et nous avons proposé une nouvelle méthode de compensation prédictive supervisée.

1.3.1.1 Impact

Les premiers travaux sur les DBN en RAP ont été (indépendamment) initiés par Jeffry Bilmes (à l'époque à University of California in Berkeley), Gregory Zweig (à l'époque à IBM) et moi-même. Il s'en est suivi une grande vague de recherche sur cette thématique. La thèse de M. Deviren a beaucoup contribué à cet essor.

1.3.1.2 Parcours après la thèse

M. Deviren a soutenu sa thèse en octobre 2004. Il a été recruté dès la fin de sa thèse par Nuance Communications, un des leader mondiaux en technologies de la communication parlée. Il y travaille toujours.

URL de la thèse : www.sudoc.abes.fr/cbs/DB=2.1/SRCH ?IKT=12&TRM=082847274

1.3.2 Thèse de Jérôme Louradour

Entre 2004 et 2007, j'ai encadré la thèse de Jérôme Louradour, sous la direction de Régine André-Obrecht. Cette thèse a porté sur la reconnaissance automatique du locuteur (RAL) par des approches de classification discriminantes. La méthode de classification qui régnait dans ce domaine était les SVM (Support Vector Machines). Cependant, cette approche ne permettait pas d'intégrer directement/naturellement l'aspect dynamique du signal. La thèse s'est penchée sur l'exploration et le développement des noyaux de séquences pour la classification SVM du locuteur. Nous avons proposé une nouvelle famille

de noyaux en se basant sur une généralisation d'un noyau qui a fait ses preuves en RAL, le noyau GLDS. Nous avons effectué l'analyse théorique et algorithmique de cette nouvelle famille avant de l'appliquer à la RAL par SVM. Après la mise en œuvre des systèmes SVM à base des différents noyaux que nous avons étudiés, nous avons comparé leurs performances dans le cadre de la campagne d'évaluation NIST-SRE 2005. Enfin, nous avons introduit un nouveau concept pour aborder le problème de RAL, dont le principe est de déterminer si deux séquences ont été prononcées par le même locuteur. L'utilisation des SVMs pour exploiter ce concept nous a amené a définir une nouvelle catégorie de noyaux : les noyaux entre paires de séquences.

1.3.2.1 Impact

Les travaux de la thèse de J. Louradour ont suscité un intérêt non négligeable dans la communauté Machine Learning. J'ai par exemple été invité en 2007 par Sami Bengio à Google (CA, USA) pour présenter ces travaux. En RAL, l'intérêt a été de courte durée car ils ont vite été dépassés par l'approche qui allait devenir dominante, les i-vectors.

1.3.2.2 Parcours après la thèse

J. Louradour a soutenu sa thèse en janvier 2007. Il a ensuite effectué un séjour post-doctoral à l'Université de Montréal sous la direction de Yoshua Bengio, un des pionniers du Deep Learning. Il travaille actuellement à Wolfram Research.

URL de la thèse : www.afcp-parole.org/doc/theses/theseJL07.pdf

1.3.3 Thèse de Reda Jourani

Entre 2008 et 2012, j'ai encadré la thèse en co-tutelle de Reda Jourani, sous la direction de Régine André-Obrecht et Driss Aboutajdine. La plupart des systèmes de reconnaissance étaient basées sur l'apprentissage génératif de modèles de mélange Gaussien (GMM). L'entraînement génératif du GMM n'optimise cependant pas directement les performances de classification. Il était donc intéressant de développer des approches dis-

criminatoires alternatives qui abordent directement le problème de classification car elles conduisent généralement à de meilleures performances que les méthodes génératives. Par exemple, les SVM (super vector machines) associées aux paramètres GMM faisaient partie des approches de pointe en matière de vérification du locuteur. La thèse de R. Jourani à proposé une alternative à ces deux approches en se basant sur une autre approche discriminante émergente pour la classification multi-classes, les GMM à grande marge (LM-GMM). Comme les SVM, les paramètres des LM-GMM sont estimés en résolvant un problème d'optimisation convexe. Cependant, ils diffèrent des SVM en utilisant des ellipsoïdes pour modéliser les classes directement dans l'espace des données, au lieu d'un espace de grande dimension comme le font les SVM. Bien que les LM-GMM aient été utilisé en reconnaissance vocale, ils ne l'étaient pas en reconnaissance du locuteur. Nous avons ainsi proposé des versions simplifiées, rapides et plus efficaces des LM-GMM qui exploitent les propriétés et caractéristiques spécifiques aux systèmes de reconnaissance du locuteur. Cette approche a été validée et évaluée sur des tâches d'identification et de vérification du locuteur dans le cadre de la campagne d'évaluation NIST-SRE'2006.

1.3.3.1 Impact

Les travaux de la thèse de R. Jourani ont été intéressants d'un point de vue théorique mais n'ont pas conduit à des améliorations significatives par rapport aux SVM, dans le cadre de la RAL.

1.3.3.2 Parcours après la thèse

R. Jourani a soutenu sa thèse en septembre 2012. Il a ensuite été recruté comme professeur assistant à l'Université de Tétouan au Maroc.

URL de la thèse : tel.archives-ouvertes.fr/tel-00807563/document

1.3.4 Thèse de Vahid Khanagha

Entre 2008 et 2012, j'ai encadré la thèse de Vahid Khanagha, avec Hussein Yahia comme directeur officiel. Cette thèse a porté sur l'application du Formalisme Microcanonique Multiéchelles (FMM) à l'analyse de la parole. Dérivé de principes issus en physique statistique, le FMM peut permettre une analyse géométrique précise de la dynamique non-linéaire des signaux complexes. Il est fondé sur l'estimation des paramètres géométriques locaux, appelés les exposants de singularité, qui quantifient le degré de prédictibilité à chaque point du domaine du signal. Si correctement définis est estimés, ils fournissent des informations précieuses sur la dynamique locale de signaux complexes et ont été utilisés dans plusieurs applications allant de la représentation des signaux à l'inférence ou la prédiction. Nous avons démontré la pertinence du FMM en analyse de la parole et développé plusieurs applications qui montrent le potentiel et l'efficacité du FMM dans ce domaine. Ainsi, nous avons introduit : un nouvel algorithme performant pour la segmentation phonétique indépendante du texte, un nouveau codeur du signal de parole, un algorithme robuste pour la détection précise des instants de fermeture glottale, un algorithme rapide pour l'analyse par prédiction linéaire parcimonieuse et une solution efficace pour l'approximation multi-pulse du signal source d'excitation.

1.3.4.1 Impact

Les travaux de cette thèse ont eu un impact considérable dans la communauté. J'ai été ainsi invité en 2012 par Jasha Droppo à Microsoft Research (WA, USA) pour présenter ces travaux. En outre, le papier "A novel text-independant phonetic segmentation algorithm based on the microcanonical multiscale formalism" a été nominé pour le Best Paper Award à la conférence majeure Interspeech 2010. Le papier "An efficient solution to sparse linear prediction analysis of speech" a été parmi les top 10 des téléchargements en 2012 de EURASIP Journal on Audio, Speech, and Music Processing. L'outil de détection des GCI a été mis à la disposition de la communauté et est largement utilisé, il fait encore partie à ce jour de l'état de l'art du domaine (geostat.bordeaux.inria.fr/index.php/downloads.html).

1.3.4.2 Parcours après la thèse

V. Khanagha a soutenu sa thèse en janvier 2013. Il a été ensuite recruté comme ingénieur de recherche à l'Université du Maryland-College Park. Il est actuellement ingénieur R&D chez Motorola Solutions.

URL de la thèse : geostat.bordeaux.inria.fr/images/these-v-khanagha.pdf

1.3.5 Thèse de Biswajit Das

Depuis 2018, j'encadre la thèse de Biswajit Das, avec Hussein Yahia comme directeur officiel. Cette thèse est toujours en cours, la soutenance est prévue fin juin 2021. La thématique générale de cette thèse est l'analyse de la parole pour l'aide au diagnostic différentiel entre 2 maladies neurodégénératives, la maladie de Parkinson (MP) et l'atrophie multisystématisée (AMS). Dans les premiers stades de la maladie, les symptômes de MP et AMS sont en effet très similaires, surtout pour la variante AMS-P où le syndrome parkinsonien prédomine. Le diagnostic différentiel entre AMP-P et MP peut être très difficile dans les stades précoces de la maladie, tandis que la certitude de diagnostic précoce est important pour le patient en raison du pronostic divergent. Malgré des efforts récents, aucun marqueur objectif valide n'est actuellement disponible pour guider le clinicien dans ce diagnostic différentiel. La nécessité de ces marqueurs est donc très élevé dans la communauté de la neurologie, en particulier compte tenu de la gravité du pronostic de AMS-P.

Durant la première partie de cette, nous ne disposions pas d'assez de données pour nous attaquer directement à cette tâche. La collecte des données aux CHU de Bordeaux et Toulouse a pris en effet du retard pour différentes raisons. Nous avons ainsi commencé par explorer une tâche similaire, le diagnostic différentiel AMS et la paralysie supranucléaire progressive (PSP), en utilisant des données fournies par des collaborateurs de l'université technique de Prague. La PSP est une autre maladie neurodégénérative dont les symptômes peuvent aussi être similaires à AMS et à MP aux stades précoces.

Durant la deuxième partie de cette, une fois nous avons pu collecté assez de données,

nous nous sommes concentrés sur l'étude de la réalisation acoustique des consonnes dans des pseudo-mots isolés. Cette direction a été motivé par nos investigations dans la première partie. Les résultats obtenus sont très prometteurs. Nous projetons actuellement de capitaliser sur ces résultats pour étudier a réalisation acoustique des consonnes dans un texte lu.

1.3.6 Thèse de Ayoub Tamim

Entre 2012 et 2015, j'ai encadré la thèse en co-tutelle de Ayoub Tamim, sous la direction de Hussein Yahia et Driss Aboutajdine. La thèse a porté sur l'étude des systèmes d'upwelling qui est d'un grand intérêt pour nombre d'applications géophysiques et océanographiques à sub-mésoéchelle. Les images satellites acquises en temps réel permettent d'avoir une surveillance continue de l'espace maritime, notamment dans le domaine thermique et visible délivrant des informations respectivement sur la distribution des températures à la surface de la mer et sur la couleur de l'eau. Nous nous sommes intéressés à la caractérisation et l'étude du phénomène d'upwelling, à partir des images de température de surface de la mer, qui est à l'origine de la formation de nombreuses structures océaniques, telles que les fronts thermiques, les filaments, les méandres, et les tourbillons. Le but est d'améliorer les interprétations visuelles des images par des océanographes qui restent souvent subjectives. La démarche proposée consiste à relier les problèmes de détection des structures thermiques à sub-mésoéchelle dans les eaux côtières d'upwelling à des concepts théoriques de traitement d'images et de vision par ordinateur. Il s'agit notamment d'utiliser des méthodes de segmentation. Par ailleurs, des méthodes de traitement d'images, basées sur des concepts de physique statistique et de thermodynamique ont été utilisées pour la mise en évidence de la turbulence océanographique caractérisant le régime chaotique des phénomènes complexes marins. Ainsi, les points de singularités détectés dans les images infrarouges, en utilisant le formalisme multi-échelle microcanonique, contiennent des informations clés à la compréhension de ces notions de turbulence et d'intermittence océanique. Ensuite, un algorithme de calcul d'indices d'upwelling, (re-

prenant et améliorant une première version développée préalablement par le CRTS et l'INRH), a été développé afin d'extraire et de suivre l'évolution de l'intensité et de l'extension d'upwelling sur la côte marocaine. Cet algorithme permet de suivre à la fois en temps réel et d'analyser les séries temporelles des variations de ce phénomène. Enfin, une étude statistique a permis de faire un suivi de la variabilité saisonnière de l'intensité de l'upwelling sur la côte atlantique marocaine tout en mettant en évidence les caractéristiques spatio-temporelles de ce phénomène via des indices établis à cet effet. Cette analyse a permis de mettre en évidence les zones de forte activité d'upwelling.

1.3.6.1 Impact

La qualité des travaux de la thèse de A. Tamim a été reconnue par la communauté. Il a reçu la **Médaille d'or du prix de thèse PHC 2017** (www.inria.fr/fr/ayoub-tamimmedaille-dor-du-prix-de-these-2017-hubert-curie).

1.3.6.2 Parcours après la thèse

A. Tamim a soutenu sa thèse en septembre 2015. Il a été ensuite recruté par l'institut de pêche maritime au Maroc.

URL de la thèse : hal.inria.fr/tel-01242495

1.3.7 Thèse de Anass El Aouni

Entre 2016 et 2019, j'ai encadré la thèse en co-tutelle de Anass El Aouni, sous la direction de Hussein Yahia et Khalid Minaoui. La thèse a porté sur l'étude des processus physiques du système d'upwelling qui sont essentiels pour comprendre sa variabilité actuelle et ses changements passés et futurs. Cette thèse a présenté une étude interdisciplinaire du système d'upwelling côtier à partir de différentes données acquises par satellite, l'accent étant mis principalement sur le système d'upwelling d'Afrique du Nord-Ouest (NWA). Cette étude interdisciplinaire aborde (1) le problème de l'identification et de l'extraction automatiques du phénomène d'upwelling à partir d'observations satellitaires biologiques

et physiques. (2) Une étude statistique de la variation spatio-temporelle de l'upwelling de la NWA tout au long de son extension et de ses différents indices d'upwelling. (3) Une étude des relations non linéaires entre le mélange de surface et l'activité biologique dans les régions d'upwelling. (4) études lagrangiennes de tourbillons cohérents ; leurs propriétés physiques et identification automatique. (5) L'étude des transports effectués par les tourbillons lagrangiens de la NWA Upwelling et leur impact sur l'océan.

1.3.7.1 Impact

— La qualité des travaux de la thèse de A. El Aouni a été reconnue par la communauté. Il a reçu prix de thèse prestigieux **Prix de thèse Systèmes complexes CNRS ISC-PIF 2020** (iscpif.fr/recherche/prixde-these/)

— Le logiciel Upwelling-Seg-Index, développé par Anass El Aouni sous ma supervision, a été transféré en 2018 au CRTS (crts.gov.ma) puis mis à disposition de la communauté (github.com/aelaouni/Upwelling-segmentation-and-index). Ce logiciel permet l'identification des zones d'upwelling sur la côte nord-ouest africaine et calcule un indice d'intensité de l'upwelling. Il permet aussi de calculer sa variation spatio-temporelle pour une période donnée.

— Des expérimentations basées sur ce travail ont été menées à Mercator-Océan (www.mercator-ocean.fr/) pour valider le résultat des données du modèle océanique opérationnel et des produits de ré-analyse.

1.3.7.2 Parcours après la thèse

A. El Aouni a soutenu sa thèse en septembre 2019. Il effectue actuellement un séjour postdoctoral à Inria Grenoble Rhône-Alpes.

URL de la thèse : tel.archives-ouvertes.fr/tel-02506913/document

1.4 Autres encadrements

1.4.1 Encadrement de postdoc

— Juin 2005 – Nov 2006 : Eduardo Sanchez-Soto
Subject : Multimodal speech recognition
Co-encadré avec Alex Potamianos
Financé par le réseau d'excellence européen MUSCLE

1.4.2 Encadrement de stages

— March – Aug 2017 : Gongfeng LI, Télécom Paris-Tech
Master2 Internship funded by INRIA
Subject : Discrimination between atypical Parkinsonian syndromes using speech analysis

— Mai – July 2017 : Quentin Robin, INP-Grenoble
Master1 Internship funded by INRIA
Subject : Automatic detection of GCI on pathological voices

— June - July 2014 : Jiri Mekyska, Technical University of Brno, Czech republic
Ph.D Internship funded by Institut Joseph Fourier
Subject : Analysis of Parkinsonian speech

— Juin – Aug 2013 : Blaise Bertrac, University Bordeaux 1
Master1 Internship funded by INRIA
Subject : Matching pursuit for speech coding and classification

— March – Aug 2013 : Safa Mrad, National engineering school of Tunis (ENIT)
Master2 Internship funded by INRIA
Subject : Nonlinear speech analysis of pathological speech

— Oct 2012 – Apr 2013 : Nicolas Vinuesa, National University of Rosario, Argentina
Master2 Internship funded by INRIA

Subject : Biologically realistic coding efficiency in the auditory cortex versus wavelet analysis

— Apr – Sept 2011 : Joshua Winebarger, Georgia Tech, USA and Supelec Paris
Master2 Internship funded by INRIA
Subject : Speech segmentation

— May – Sept 2010 : Joshua Winebarger, Georgia Tech, USA
Master1 Internship funded by INRIA
Subject : Speaker segmentation methods for phonetic segmentation

— Apr – June 2007 : Andrey Temko, Technical University of Catalonia (UPC), Spain
Ph.D Internship funded by UPC
Subject : SVM acoustic events detection with temporal overlaps

— March – Aug 2003 : Sanaa Ghouzali, University of Rabat, Morocco
Master2 Internship funded by INRIA
Subject : Speech modeling using HMMs on wavelet-trees

1.5 Management

1.5.1 Création de l'équipe de recherche Inria GEOSTAT

En 2009, j'ai crée avec Hussein Yahia l'équipe-projet Inria GEOSTAT au centre de recherche Inria Bordeaux Sud-Ouest (geostat.bordeaux.inria.fr/). GEOSTAT est une équipe de recherche multidiscplinaire sur l'analyse et caractérisation de certaines classes de signaux naturels complexes (séries chronologiques physiologiques, données d'observation de la terre et de l'univers) en étudiant, dans les signaux acquis, les propriétés prédites par les modèles physiques décrivant le mieux les systèmes sous-jacents.

1.5.2 Projets de recherche

Je liste dans ce qui suit, par orde chronologique, les projets de recherche que j'ai dirigés ou que je dirige, seul ou en tandem.

— En mars 2020 (dès le début de la crise Covid-19), j'ai proposé de développer un outil pour aider au suivi à domicile de patients atteints du Covid. Cette proposition a suscité l'intérêt d'Inria, du MESRI et du corps médical. Je dirige depuis, avec Thomas Similowski responsable du service de pneumologie et de réanimation de la Pitié-Salpêtrière et de l'UMR-S 1158, le projet VocaPnée (dream.inria.fr/vocapnee/). L'objectif ambitieux de ce projet d'envergure est le développement d'un biomarqueur vocal de la fonction respiratoire et de son évolution au cours du télé-suivi à domicile après une affection respiratoire (Covid ou autre).

Partenaires : service de pneumologie et de réanimation de la Pitié-Salpêtrière ; UMR-S 1158.

Participants Inria : Direction scientifique, équipe TAU et potentiellement d'autres équipes.

Partenaire européen : Équipe SAMI de l'université technique de Prague.

— Depuis mars 2017 : Coordinateur du projet ANR Voice4PD-MSA (voice4pd-msa.inria.fr/)

Thématique : Diagnostic différentiel précoce entre la maladie de Parkinson et l'atrophie multisystématisée par analyse de la parole.

Partenaires : Services de neurologue et d'ORL des CHU Bordeaux et Toulouse ; Équipe Samova de l'IRIT, Institut de mathématiques de Toulouse.

— 2016-2018 : Coordinateur d'un projet PHC-Toubkal

Thématique : Caractérisation multi-capteurs et suivi spatio-temporel de l'upwelling sur la côte atlantique marocaine par imagerie satellitaire

Partenaire Français : Mercator-Océan (www.mercator-ocean.fr/)

Partenaires Marocains : LRIT, université de Rabat (www.fsr.ac.ma/lrit) ; Centre Royal de Télédétection Spatiale (CRTS, crts.gov.ma) ; Ministère de l'Agriculture et

de la Pêche Maritime (www.mpm.gov.ma)

— 2015-2016 : Coordinateur d'un projet Carnot-Inria
Thématique : Reconnaissance automatique des émotions dans la parole
Partenaire : Start-up Batvoice (www.batvoice.com)

— 2011-2013 : Coordinateur d'un projet PHC-Volubilis (version antérieure de Toubkal)
Thématique : Étude par imagerie satellitaire de l'Upwelling sur la côte atlantique marocaine
Partenaires Français : LEGO-CNRS (www.legos.obs-mip.fr)
Partenaires Marocains : LRIT ; CRTS.

— 2009 : Coordinateur d'un projet RTRA-STAE (www.fondation-stae.net)
Thématique : Geometrical and multiscale approaches for predictability and analysis of complex data in astrophysics and geophysics

— 2004-2008 : Membre du comité de pilotage du réseau d'excellence européen MUSCLE (muscle.ercim.eu) et responsable de l'axe « traitement de la parole ».
Thématique : Multimedia Understanding through Semantics, Computation and Learning

— 2003 – 2004 : Co-coordinateur (avec Laurent Younes, ENS Cachan) d'un projet national MathStic
Thématique : Modèles probabilistes graphiques pour la reconnaissance la parole

1.6 Transferts technologiques

Dans ce qui suit, de donne une brève description de 3 actions de transfert auxquelles j'ai participé. J'ai été membre du COPIL de la première et j'ai dirigé les deux dernières.

1.6.1 Débruitage d'images de grands scanners

Entre 2018 et 2020, j'ai été membre du comité de pilotage de Inria Innovation Lab I2S-GeoStat (www.inria.fr/fr/i2s-geostat-un-innovation-lab-en-imagerie-numerique). Ce der-

nier a permis le transfert à la société I2S (www.i2s.fr) une technologie de débruitage haute qualité d'images de grands scanners. Cette technologie, à la pointe au niveau mondial, est basée sur les résultats des travaux exceptionnels de la thèse de Hicham Badri (hal.inria.fr/tel-01239958) dont j'ai permis le recrutement à GEOSTAT grâce à ma collaboration avec le Maroc. Elle a permis à I2S de gagner des parts de marché internationaux hautement compétitifs. La thèse de H. Badri a reçu le prix de thèse prestigieux AFRIF 2016 (www.afrif.asso.fr/).

1.6.2 Caractérisation de l'upwelling

Le logiciel Upwelling-Seg-Index, développé par Anass El Aouni sous ma supervision, a été transféré gracieusement en 2018 au CRTS (crts.gov.ma) puis mis à disposition de la communauté (github.com/aelaouni/Upwelling-segmentation-and-index). Ce logiciel permet l'identification des zones d'upwelling sur la côte nord-ouest africaine et calcule un indice d'intensité de l'upwelling. Il permet aussi de calculer sa variation spatio-temporelle pour une période donnée.

1.6.3 Reconnaissance automatique des émotions dans la parole

Le logiciel prototype Classif'Emo, développé par Nicolas Brodu sous ma supervision, a été transféré à la start-up Batvoice (www.batvoice.com) en 2016. Ce logiciel est un démonstrateur de la reconnaissance de certaines émotions à partir d'enregistrements de parole spontanée. Ce transfert a eu un impact important en terme de visibilité de GeoStat. J'ai en effet effectué une demo du logiciel lors de l'évènement « Tremplin Entreprises 2016 » qui a eu lieu au Sénat en juin 2016 et auquel Batvoice a été lauréat.

1.7 Prix et distinctions

"Best paper award" à IEEE International Conference on Industrial and Engineering Application of Artificial Intelligence and Expert Systems (IEA/AIE), Cairns, Australia,

June 17-20th, 2002, pour mon papier "Automatic Speech Recognition : the New Millennium".

Grâce à ce travail j'ai reçu la Médaille INRIA-LORIA 2002.

1.8 Responsabilités collectives

Je suis reviewer régulier de la revue IEEE/ACM Transactions on Audio, Speech and Language Processing et des 2 conférences majeures en parole, ICASSP et Interspeech. De façon occasionnelle, je fais des reviews dans d'autres revues et conférences. J'ai aussi expertisé quelques projets ANR et internationaux.

J'ai été membre du comité d'organisation de IEEE-ISIVC'2018 et des JEP'2002 (Journées d'Études sur la Paroles) ainsi que membre du comité de programme de plusieurs conférences.

1.9 Enseignement

— 2019 : Université de Lorraine ; Master1 MIAGE-ACSI.
  20h de cours : Data mining

— 2014 à 2016 : Université de Lorraine ; Master2 MIAGE.
  20h de cours chaque année : Mathématiques financières

— 2012 à 2014 : Université de Rabat, Master2 InfoTelecom
  20h de cours chaque année : Traitement automatique de la parole

— 2001-2002 : Université Nancy 1, DEA Informatique
  20h de cours : Apprentissage et inférence des réseaux Bayésiens

1.10 Organisation du mémoire

Le reste du document sera consacré à la description scientifique de mes principales recherches en traitement de la parole, les autres thématiques ne seront pas décrites. Il

est organisé de la manière suivante. Le prochain chapitre présente l'une de mes principales contributions en reconnaissance de la parole. Il est basé sur l'article [P2] de ma bibliographie personnelle. Le chapitre 3 présente l'une de mes principales contributions en reconnaissance du locuteur. Il est basé sur l'article [P4]. Les chapitres suivants seront plus élaborés car leur contenu fait toujours partie de l'état de l'art et est toujours d'actualité pour mes recherches actuelles et futures. Le chapitre 4 présente mes principales contributions en analyse non-linéaire de la parole. Il s'appuie sur les articles [P7] et [P10]. Le chapitre 5 présente mes recherches et contributions en analyse de la dysarthrie. Il s'appuie sur mes articles ICASSP-2018, 2019 et 2021 (soumis). Les chapitres 2 à 5 sont rédigés en anglais. Enfin, je termine par une description des perspectives pour mes recherches futures.

Chapitre 2

Dynamic Bayesian networks for speech recognition

Sommaire

2.1 Introduction 35

2.2 Bayesian networks 37

2.3 Inference algorithms for Bayesian networks 40

2.3.1 Construction of the junction tree 40

2.3.2 Propagation of evidence in the junction tree 42

2.4 Multi-band speech recognition : the classical approach 45

2.5 Multi-band speech recognition : the DBNs perspective 47

2.5.1 Model definition 47

2.5.2 Construction of the junction tree 50

2.5.3 Model parameters estimation 51

2.6 Application to isolated speech recognition 54

2.7 Application to continuous speech recognition 57

2.7.1 Decoding algorithm 58

2.7.2 Experiments 59

2.1 Introduction

Automatic speech recognition (ASR) systems have been based on probabilistic modeling of the speech signal using Hidden Markov Models (HMMs). These models lead to the best recognition performances in ideal "lab" conditions or for easy tasks. However, in real word conditions of speech processing (noisy environment, spontaneous speech, non-native speakers...), the performance of HMM-based ASR systems can decrease drastically and their use becomes limited. One of the major reasons for this discrepancy is the fact that classical HMM's parameterization and modeling fail to capture some acoustic phenomena which are specific to speech. For instance, while speech temporal dynamics are well captured by HMMs, the frequency dynamics (which are phonetically very informative) are weakly modeled in classical HMM-based systems.

A speech modeling approach, known as *multi-band* speech recognition, has been proposed [1, 2, 3]. This approach takes its origin in an extensive study done by Harvey Fletcher [4] on how humans process and recognize speech. Basically speaking, this study (reviewed by Jont B. Allen in [5]) suggests that the human auditory system processes speech *locally* in the time-frequency domain before recognition. The general approach to multi-band speech recognition is to divide the time-frequency domain into several sub-bands, then each sub-band is independently modeled by a HMM. The recognition scores in the sub-bands are then fused with some recombination module. This approach has also been motivated by the desire to improve robustness to additive noise, particularly band-limited noise. Indeed, in classical systems the full frequency band is *globally* processed in order to extract speech features, thus the resulting acoustic vectors are all corrupted even if the noise covers only a small frequency sub-band. Using the multi-band local frequency processing, only the information extracted from the noisy sub-band will be corrupted, the remaining non-corrupted information can be then exploited for recognition.

Definitely the multi-band principle (i.e. local processing in the time-frequency domain) is very attractive because it attempts to mimic the behavior of the human auditory system and it can lead to noise-robust systems. However, the classical approach (described above)

to exploit this principle is far from being optimal. For instance, the sub-bands are assumed mutually independent which is an unrealistic hypothesis. Moreover, the information contained in one sub-band is not discriminative in general. In addition, the recombination step can be a very difficult task, particularly in continuous speech recognition.

The scope of this work was to propose a new approach to multi-band speech recognition which has the advantage to overcome *all* the limitations (mentioned above) of the classical multi-band (CMB) approach. In the latter, the fundamental weakness is the fact that sub-bands modeling is *independent*, the basic idea behind our approach is to render *dependent* such modeling. A way to do so is to create "communications" or "interactions" between the different HMMs that model the different sub-bands. For this propose, we used the formalism of *Bayesian networks* (BNs) which was an appropriate framework for our goal for two reasons. First, through meaningful graphical representations, Bayesian networks has the advantage to provide a natural tool to represent interactions and dependencies between variables of a given system. Second, by exploiting conditional independence between system variables, they introduce some "modularity" in large-scale problems in order to split them up into small and tractable problems. Consequently, Bayesian networks not only provide an attractive tool for modeling complex systems, but also lead to efficient general-purpose algorithms.

After Judea Pearl's pioneering work [6], Bayesian networks have emerged as a powerful formalism unifying many concepts of probabilistic modeling widely used in statistics, artificial intelligence, signal processing and other fields. For example, HMMs, mixture models, and Kalman filters are all particular instances of the more general BNs formalism. BNs have then become a very popular framework for reasoning under uncertainty and have been widely used in expert systems design and decision making systems. However, the use of BNs in automatic speech recognition has started to gain attention very recently back then [7, 8, 9, 10, 11, 12, 13, 14, 15, 16]. This chapter presents a multi-band system which relies on a "uniform" time-frequency speech model. Namely, instead of considering an independent HMM for each sub-band (as in the CMB approach), we built a more complex but unique dynamic Bayesian network on the time-frequency domain by "coupling"

all the HMMs associated with the different sub-bands. We developed the learning and decoding algorithms corresponding to this new speech model and carried out illustrative experiments to show the potential of this multi-band approach. The chapter is organized as follows. In the next section, a brief introduction to Bayesian networks is given. In order to make the chapter self-contained, section 2.3 presents the inference algorithms that we used in learning and decoding. Section 2.4 presents the classical approach to multi-band ASR. Section 2.5 presents the principle and the algorithmic details of our approach to multi-band ASR. Section 2.6 and 2.7 show the application of our approach in isolated and continuous speech recognition and presents *illustrative* experiments on an isolated and connected digit recognition task.

2.2 Bayesian networks

During the last decades, Bayesian networks (and probabilistic graphical models in general) have become very popular in artificial intelligence (and other fields) due to many breakthroughs in several aspects of inference and learning. The literature is now extremely rich in papers and books dealing with the theory and applications of BNs, among which we refer to [17, 18] for a very good introduction. The formalism of probabilistic graphical models (PGMs) is well summarized in the following quotation by M. Jordan [19] :

"Graphical models are a marriage between probability theory and graph theory. They provide a natural tool for dealing with two problems that occur throughout applied mathematics and engineering - uncertainty and complexity - and in particular they are playing an increasingly important role in the design of machine learning algorithms. Fundamental to the idea of a graphical model is the notion of modularity - a complex system is built by combining simpler parts. Probability theory provides the glue whereby the parts are combined, ensuring that the system as whole is consistent, and providing ways to interface models to data. The graph theoretic side of graphical models provides both an intuitively appealing interface by which humans can model highly-interacting sets of variables as well as a data structure that lends itself naturally to the design of efficient general-purpose

algorithms."

More precisely, given a system of random variables (r.v.), a PGM consists in associating a graphical structure to the joint probability distribution of this system. The nodes of this graph represent the r.v., while the edges encode the (in)dependencies which exist between these variables. One distinguishes three types of graphs : directed, undirected and those for which the edges are a mixture of both. In first case, one talks about *Bayesian networks*, in the second case, one talks about *Markov random fields*, and in the third case one talks about *chain networks*. PGMs have two major advantages :

— They provide a natural and intuitive tool to illustrate the dependencies which exist between variables. In particular, the graphical structure of a PGM clarifies the conditional independencies embedded in the associated joint probability distribution.

— By exploiting these conditional independencies, they provide a powerful setting to specify efficient inference algorithms. Moreover, these algorithms may be specified automatically once the initial structure of the graph is determined.

So far, the conditional independencies semantics (or Markov properties) embedded in a PGM are well-understood for Bayesian networks and Markov random fields. For chain networks, these are still not well-understood. In our current research, given the causal and dynamic aspects of speech, Bayesian networks (BNs) are of particular interest to us. Indeed, thanks to their structure and Markov properties, BNs are well-adapted to interpret causality between variables and to model temporal data and dynamic systems. In addition, not only HMMs are a particular instance of (dynamic) BNs, but also the Viterbi and Forward-Backward algorithms (which made the success of HMMs in speech) are particular instances of generic inference algorithms associated to BNs [20]. This shows that BNs provide a more general and flexible framework than the HMMs paradigm which has ruled ASR for the last three decades.

Formally, a (static) Bayesian network has two components : a directed acyclic graph S and a numerical parameterization Θ. Given a set of random variables $X = \{X_1, ..., X_N\}$ and $P(X)$ its joint probability distribution (JPD), the graph S encodes the conditional

independencies which (are supposed to) exist in the JPD. The parameterization Θ is given in term of conditional probabilities of variables given their parents. Once S and Θ are specified, the JPD can be expressed in a factored way as [1]

$$P(x) = \prod_{i=1}^{N} P(x_i|pa(x_i)) \tag{2.1}$$

where $pa(x_i)$ denotes an outcome of the parents of X_i. The conditional independence semantics (or Markov properties) of a BN imply that, conditioned on its parents, a variable is independent of all the other variables except its descendants.

Dynamic Bayesian networks (DBNs) extend the BN representation to dynamic processes. This representation encodes the beliefs about possible trajectories of the process. Consider a time evolving set $X[t] = \{X_1[t], \ldots, X_N[t]\}$ of variables. A DBN encodes the joint probability distribution of these variables in a finite time interval $[0, T]$. In general, this JPD can be encoded in a huge static BN with $T \times N$ variables with (possibly) different structure and parameters for each time slice. If the underlying process is stationary, then the independence assertions and the associated conditional probabilities are identical for each time slice t. In this case, the repeating structure and parameters can be encoded with a static BN in a single time slice. From this point of view, it is obvious that a HMM is a particular DBN as shown in Figure 2.1. Contrarily to the usual state transition diagram, in the DBN representation each node H_t (resp. O_t) is a random variable whose outcome indicates the state occupied (resp. the observation vector) at time t. Time is thus made explicit and arrows linking the H_t must be understood as "causal influences" (not as state transitions). It is this representation of HMMs that we shall use in the rest of the paper.

The next section presents the algorithms we used to infer our acoustic models. This section is introduced to make the paper self-contained. Thus, it can be skipped by readers who are not interested in the algorithmic details. On the other hand, those who are interested in these aspects may find the description too short. We advise them the very nice tutorials on Bayesian networks [21, 20] and also the very interesting thesis [22].

1. In the whole paper, upper-case (resp. lower-case) letters are used for random variables (resp. outcomes).

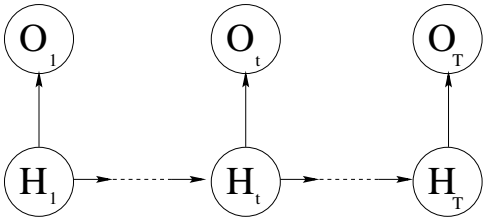


FIGURE 2.1 – a HMM represented as a dynamic Bayesian network

2.3 Inference algorithms for Bayesian networks

Once a Bayesian network is specified, i.e., its graph and numerical parameterization are given, the most common problem is *inference*. Namely, one is interested in the calculation of marginal or conditional probabilities of subsets of variables, such as the likelihood of observed evidence or the probability of some variables given evidence, or one is interested in identifying the most likely outcome of unobserved (hidden) variables given observed evidence. In the last decade, major progress has been accomplished in the theory of BNs. In particular, fast and exact inference algorithms has been developed when all the variables are discrete, all Gaussian or mixed discrete-Gaussian [17]. In this section, we present the JLO algorithm [23] as well as the David algorithm [24] which are the most popular algorithms for exact inference of discrete BNs. We recall here that the JLO and Dawid algorithms correspond respectively to the Forward-Backward and Viterbi algorithms when applied to the particular case of HMMs [20].

The JLO and Dawid algorithms proceeds in two steps. The first one consists in using graph-theoretic tools to transform the initial graphical structure of the BN into a specific graphical entity called the *junction tree*. In the second step, the junction tree is used as a channel to transmit and propagate the effect of observations (or evidence).

2.3.1 Construction of the junction tree

The first operation in the construction of the junction tree for BNs is the *moralization*. It consists in adding an extra undirected edge between any two nodes with a common child and subsequently removing directions. The undirected graph obtained this way is

called the *moral* graph. The second operation consists in adding sufficient edges to the moral graph to make it *triangulated*. An undirected graph is triangulated (or chordal) if all cycles containing four or more nodes have a chord, i.e., an undirected edge between two non-consecutive nodes in the cycle. There are several ways to add a chord in a cycle with length greater than four. Hence, the triangulation process is not unique. In general, it is desired to obtain a triangulation with a minimum number of additional edges. Unfortunately, the problem of automatically obtaining a minimal triangulation is NP-complete [25]. However, there exists some heuristic algorithms which work well in practice. For instance, the Maximum Cardinality Search Fill-In algorithm [26] can be used to obtain a triangulation of a given undirected graph. This algorithm can be implemented in linear time $O(N+l)$, where N is the number of nodes and l is the number of links in the graph. In the final operation, one identifies the set $\mathcal{C}$ of cliques [2] in the triangulated graph and forms a tree with these cliques in such way that resulting tree, the junction tree, satisfies the *running intersection property*. This property states that each variable which appears in two different cliques has to appear in all the cliques on the path between these two cliques. Figure 2.2 shows an example illustrating these different steps in the junction tree construction.

Attached with each edge linking two cliques C_1 and C_2 in the junction tree is a *separator* $S \triangleq C_1 \cap C_2$. We denote the set of separators by $\mathcal{S}$. The main advantage of the junction tree representation is the fact that, as shown in [6], the joint probability distribution $P(X)$ can be factored as the product of clique marginals over separator marginals :

$$P(x) = \frac{\prod_{C \in \mathcal{C}} P(x_C)}{\prod_{S \in \mathcal{S}} P(x_S)} \tag{2.2}$$

where $P(x_C)$ and $P(x_S)$ are the marginal distributions over the variables in C and S respectively. Thus, probability calculations on X can be carried out locally and efficiently if the cliques are relatively small. The next subsection presents the message passing scheme of the JLO and Dawid algorithms leading to such local factorization of the JPD, in the

2. A clique is a subset of nodes which are fully connected and maximal, i.e, if a node is added to the subset, the latter does not remain fully connected.

light of observed evidence.

2.3.2 Propagation of evidence in the junction tree

Given the junction tree, the JPD $P(X)$ can be factored as

$$P(x) = \frac{\prod_{C \in \mathcal{C}} \phi_C(x_C)}{\prod_{S \in \mathcal{S}} \phi_S(x_S)} \tag{2.3}$$

where $\phi_C(x_C)$ (resp. $\phi_S(x_S)$) is a non-negative potential function on the clique $\mathcal{C}$ (resp. the separator $\mathcal{S}$). The collection of potentials $\Phi = \{\{\phi_C, C \in \mathcal{C}\}, \{\phi_S, \mathcal{S} \in \mathcal{S}\}\}$ is termed a *representation* of $P(X)$. A factorizable distribution $P(X)$ may have many different representations, i.e., many collections of potentials which satisfy (2.3). For BNs, an *initial* representation is obtained from (2.1) in the following way. First, assign each X_i to just one clique. Second, for each clique C, define the potential ϕ_C to be either the product of $P(X_i|pa(X_i))$ over all X_i assigned to C, or 1 if no variable is assigned to C. Then, if ϕ_S is set to be 1 for each separator S, one obtains a representation of $P(X)$.

To propagate the effect of an observed evidence e, the JLO algorithm operates by transforming one representation to another, starting from the initial one modified by the incorporation of the evidence. The algorithm finishes with the *marginal* representation in which, for each clique C (resp. separator S), the potential ϕ_C (resp. ϕ_S) is equal to the marginal (joint) probability distribution for the variables in C (resp. S) and the evidence. The incorporation of evidence in the initial representation is done simply by setting $\phi_C(x_C)$ to 0 for any clique C containing an observed variable and for any configuration x_c involving a different state of the one observed. After this incorporation, the algorithm proceeds by passing a sequence of *flows* along the edges of the junction tree. Each flow from clique C_1 to C_2 updates the potentials of C_2 and the separator $S = C_1 \cap C_2$ in the following manner. Suppose that, prior to this flow, we have a representation Φ. Then, the activation of the

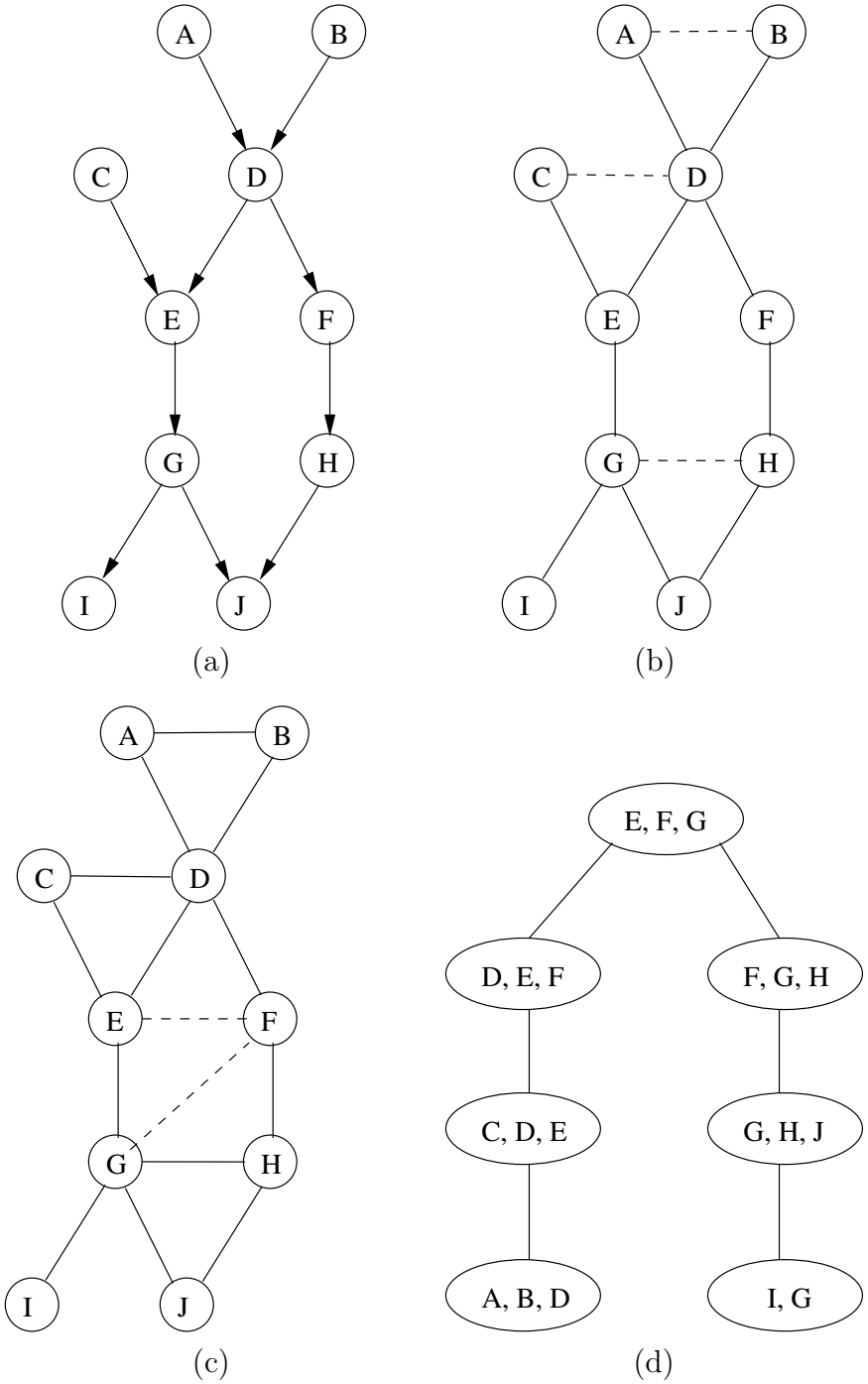


FIGURE 2.2 – Illustration of the junction tree construction algorithm. (a) A directed acyclic graph. (b) Corresponding moral graph. (c) A triangulation of the moral graph. (d) A junction tree associated with the triangulated graph.

flow yields a new representation Φ^* where the new potentials of C and S are [3]

$$\phi^*_S = \sum_{C_1 \setminus S} \phi_{C_1} \;\; ; \;\; \phi^*_{C_2} = \frac{\phi^*_S}{\phi_S} \phi_{C_2} \tag{2.4}$$

and all the other potentials being unchanged. A *schedule* of such flows consists in updating all the cliques using the available information. This is done by choosing a clique C_r to be the *root* clique and, then, operating a recursive two-phase propagation scheme : *collecting* evidence and *distributing* evidence. In the collection phase, flows are activated along all the edges of the junction tree toward C_r. In the distribution phase, flows are activated out from C_r in the reverse direction. Once a schedule is complete, one obtains a new (final) representation Φ^f in which the potentials ϕ^f_C and ϕ^f_S of each clique C and separator S equal $P(x^h_C, e)$ and $P(x^h_S, e)$ respectively, where x^h_C (resp. x^h_S) is a configuration of the hidden variables in C (resp. S) :

$$P(x) = \frac{\prod_{C \in \mathcal{C}} P(x^h_C, e)}{\prod_{S \in \mathcal{S}} P(x^h_S, e)} \tag{2.5}$$

Therefore, by marginalizing over the unobserved variables in any clique or separator, on gets the likelihood of observations $P(e)$. Also, by normalizing the potential at a clique C to sum 1, one get the posterior conditional probability $P(x^h_C|e)$ of the hidden variables in C given the evidence e. At this point, it is easy to note that the complexity of the JLO algorithm scales as the sum of the clique state-spaces [4].

The Dawid algorithm [24] is a slightly modified version of the JLO algorithm and allows the identification, with the same time complexity, of the most likely sequence of hidden states given observations [24]. There are only two modifications to operate (w.r.t. the JLO algorithm) and both are in the propagation scheme phase. The first one is to replace the sum-marginalization by a max-marginalization in the definition of a flow, i.e., to replace the summation by a maximization in (2.4). The second modification is in the distribution phase : once the potential of a clique is computed, one finds a configuration of its variables that maximizes the potential, this configuration is then considered as a

3. The summation $\sum_{C_1 \setminus S}$ is over the state-space of variables that are in C_1 but not in S.

4. A clique state-space is the product over each variable in the clique of the number of states of each variable

new evidence when flows are activated. The running intersection property guarantees that when a variable outcome is fixed in one clique, that variable has the same outcome in all other cliques.

2.4 Multi-band speech recognition : the classical approach

The multi-band principle was originally motivated by an extensive research on the way humans process and recognize speech. This research, conducted by Harvey Fletcher [4] during the first half of the 20th century, suggests that the human auditory system recognizes speech using partial information across frequency, probably in the form of speech features that are localized in the frequency domain. However, Fletcher's work has been little known until 1994 when Jont B. Allen published a paper [5] in which he reviewed the work of Fletcher and also proposed to adapt a multi-band paradigm to automatic speech recognition. Many researchers have then studied this principle to build multi-band ASR systems [1, 2, 3, 27, 28].

When applied to automatic speech recognition, the multi-band principle can be viewed as a new architecture for ASR systems. In general, this architecture consists in dividing the frequency domain of the speech signal into several frequency sub-bands, then independent processing is applied in each sub-band. The application of such a principle generally leads to a multi-band ASR system which has the architecture represented in Figure 2.3.

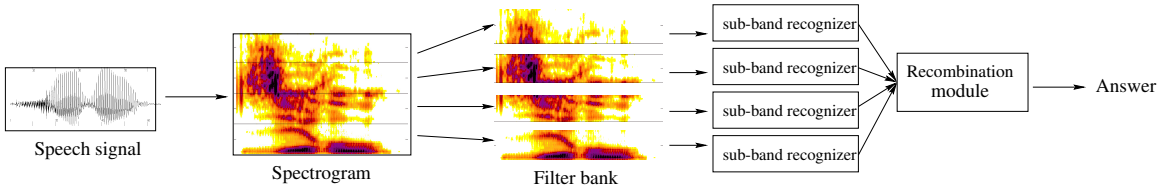


FIGURE 2.3 – Classical multi-band architecture

In such a system, the speech signal is first passed to a filter-bank which splits it into several frequency bands. The signal in each band is then encoded into a stream of acoustic vectors, which are passed to a modeling or recognition stage. This stage is

usually composed of Hidden Markov Models (HMM). During recognition, the HMMs scores are given to a recombination module, whose role is to deliver a unique answer to the recognition task. The inputs of the recombination module may either be the likelihoods that each HMM has generated the speech segment, or the label of the winning model in each band, or an ordered list of the concurrent HMMs.

Besides the motivation to mimic the human auditory system, this multi-band architecture has been also motivated by the following aspects. First, speech is characterized by *asynchrony* between frequency sub-bands, in the sense that stationary segments transition may occur at different times in the time-frequency domain. Thus, asynchrony may be taken into account by local frequency processing, this in turn could lead to higher fidelity speech modeling than traditional HMM modeling. Indeed, in the latter, speech segments are implicitly assumed to contain phoneme-synchronous information given that features extraction uses the whole frequency band. Second, the information contained in some sub-bands may be more relevant than in the other sub-bands. Thus, an "appropriate" weighting of each sub-band could improve recognition accuracy. Finally, this multi-band architecture could improve the recognition robustness to band-limited noise w.r.t. standard HMM-based systems. Indeed, even when the noise covers only one frequency sub-band, the latter would yield bad performances since the acoustic features (MFCC coefficients in general) are calculated on the whole spectrum and are then all corrupted. Using this architecture, only the acoustic features corresponding to the noisy sub-band would be corrupted. One can then exploit the non-corrupted information in the other sub-bands for recognition.

While the ideas leading to multi-band speech recognition are attractive, the classical architecture described above has many drawbacks however. For instance, the sub-bands are assumed mutually independent which is an unrealistic hypothesis. Moreover, the information contained in one sub-band is not discriminative in general. In addition, it is not easy to deal with asynchrony, particularly in continuous speech recognition. As a consequence, the recombination step can be a very difficult task.

The next section presents our approach to multi-band speech recognition which has the

advantage to overcome *all* the limitations (mentioned above) of the classical multi-band systems.

2.5 Multi-band speech recognition : the DBNs perspective

Let's assume that we are given a vocabulary V of $|V|$ words. The basic idea behind our approach is the following : for each word $v \in V$, instead of considering an independent HMM for each sub-band (as in the classical multi-band approach), we build a more complex but uniform DBN on the time-frequency domain by "coupling" all the HMMs associated with the different sub-bands. By coupling we mean adding (directed) links between the variables in order to capture the dependency between sub-bands. A natural question is : what are the "appropriate" links to add ? Probably the best answer is to learn the graphical structure (i.e., the dependencies between variables) from data. However this approach, known as *structural learning* [29], is a difficult problem [12, 30]. Instead, our philosophy was to (first) impose a "reasonable" graphical structure (for all words) and then see whether the principle of the new multi-band approach is promising. If yes, this "reasonable" structure could be used as *prior knowledge* [31] in a structural learning procedure.

2.5.1 Model definition

We built such "reasonable" structure using the following computational and physical criteria. We wanted a model where no continuous variable has discrete children in order to apply an exact inference algorithm. Indeed, only approximate inference is possible in networks where continuous variables have discrete children[32]. We also wanted a model with a relatively small number of parameters and for which the (exact) inference algorithms are tractable. Finally, we wanted to have links between the hidden variables along the frequency axis in order to capture the asynchrony between sub-bands. A simple model which satisfies these criteria is shown in Figure 2.4. In this BN, the hidden variables of

sub-band n are linked to those of sub-band $n+1$ in such way that the state of a hidden variable in sub-band $n+1$ at time t is conditioned by the state of two hidden variables : at time $t-1$ in the same sub-band and at time t in sub-band n. Each $H_t^{(n)}\left(=H_t^{(n)}(v)\right)$

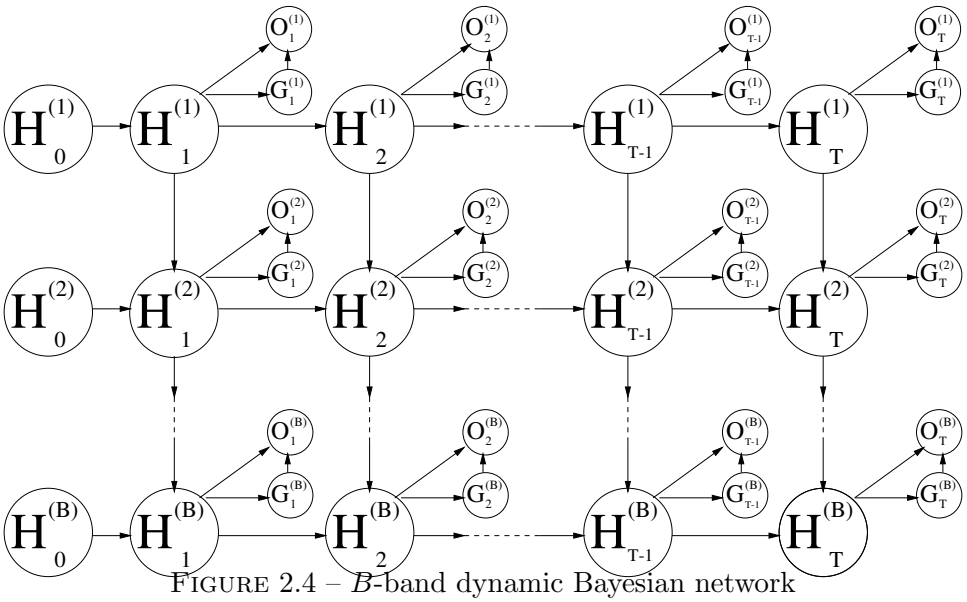


FIGURE 2.4 – B-band dynamic Bayesian network

is a discrete variable taking its values in the set of ordered labels $I_v = \{1_v, ..., m_v\}$, $|I_v|$ is the number of hidden states. Each $O_t^{(n)}\left(=O_t^{(n)}(v)\right)$ is a continuous variable with a Gaussian-mixture distribution (given an outcome of the corresponding hidden variable $H_t^{(n)}$) representing the observation vector at time t in sub-band n ($n = 1, ..., B$), B is the number of sub-bands. Each $G_t^{(n)}\left(=G_t^{(n)}(v)\right)$ is a discrete variable taking its values in the set $J = \{1, ..., M\}$, M is the number of Gaussian components in each mixture [5]. We impose a left-to-right topology on each sub-band and assume that the model parameters are stationary. Therefore, given a word $v \in V$ (and for each $(i, j, k, p) \in I_v^3 \times J$) , the

5. The use of the variables $G_t^{(n)}$ is not necessary to define the model. We use them only to have a consistency with the fact that exact inference is possible in mixed discrete-continuous BNs only if the continuous variables are Gaussian [17].

numerical parameterization Θ_v of its B-band DBN model is :

$$\begin{cases} a_{ij}(v) \triangleq P(H_t^{(1)}(v) = j | H_{t-1}^{(1)}(v) = i) \\ u_{ijk}^{(n)}(v) \triangleq P(H_t^{(n)}(v) = k | H_t^{(n-1)}(v) = i, H_{t-1}^{(n)}(v) = j) \quad \text{for} \quad n = 2, ..., B \\ w_{ip}^{(n)}(v) \triangleq P(G_t^{(n)}(v) = p | H_t^{(n)}(v) = i) \quad \text{for} \quad n = 1, ..., B \\ b_{i,p}^{(n)}(v, \cdot) \triangleq P(O_t^{(n)}(v) = \cdot | H_t^{(n)}(v) = i, G_t^{(n)}(v) = p) \quad \text{for} \quad n = 1, ..., B \end{cases} \tag{2.6}$$

where $b_{i,p}^{(n)}(v, \cdot)$ is a Gaussian with mean $\mu_{i,p}^{(n)}(v)$ and covariance $\Sigma_{i,p}^{(n)}(v)$. The asynchrony between sub-bands is taken into account by allowing all the $u_{ijk}^{(n)}(v)$ to be non-zero, except when $k < j$ or $k > j + 1$ because of the left-to-right topology.

Note that our model is a mixed discrete-Gaussian BN. Therefore, in principle, inference should be done using the Lauritzen algorithm [32]. However, in our setting, inference will always involve a *complete* observations set of the continuous variables. In other words, *all* the $O_t^{(n)}$ are observed when our model is infered. Thus, even tough the B-band DBN is mixed discrete-Gaussian, the JLO and Dawid algorithms (which apply to discrete networks) are enough to perform inference in this setting. Note also that our model is a special case of the so called *tree structured HMMs* [21]. However, we did not use the variational approach described in [21] to infer this model because we were interested in exact inference.

The following stretch the advantages of such approach to multi-band ASR and describe briefly some related work. Unlike HMMs, our multi-band DBN provides *a* modeling of the frequency dynamics of speech. Unlike to the classical multi-band approach, our DBN allows interaction between sub-bands and the possible asynchrony between them easily handeled. Moreover, our model uses the information contained in all sub-bands and no recombination step is needed. A related work has been proposed in [33, 34] where a multi-band Markov random field is analyzed by mean of Gibbs distributions. This approach (unlike ours) does not lead however to exact nor fast inference algorithms and assumes a linear model for asynchrony between sub-bands. In our approach, the asynchrony is learned from data. In term of introducing frequency dynamics in the modeling process, a related work has been proposed in [35, 36, 37]. In this work, the authors propose a

new model, called HMM2, which is an HMM "mixture" consisting in a primary (classical) HMM, modeling the temporal properties of the speech signal, and a secondary HMM modeling its frequency properties. This secondary HMM is inserted at the level of each state of the primary HMM, estimating local emission probabilities of acoustic feature vectors (consisting in spectral features). Consequently, the components of an acoustic vector are assumed to be generated by the secondary HMM, the goal being to perform a (state dependent) dynamical spectral warping to complement the (classical) time warping done by the primary HMM. This HMM2 has then been used as a decoder as well as feature extractor and tested in various conditions. In the case of clean speech, performances were showed to be comparable to classical MFCC-based HMM systems. However, in the case of noisy speech, the performances are so far still limited by the choice of the spectral features, which are less robust to noise than MFCCs. Besides the introduction of some modeling of the frequency dynamics of speech, HMM2 is different from our model in all aspects : topology and parameterization. Indeed, our model is a "pure" multi-band model in the sense that the frequency axis is divided in a static way, we use MFCCs in the parameterization and obtain some good performances in the presence of noise as compared to classical HMM systems (as we will see later).

2.5.2 Construction of the junction tree

Now that the graphical structure and the numerical parametrization of our multi-band DBN are specified, inference can be performed using the JLO or/and the Dawid algorithms. The only step remaining is to associate a "minimal" junction tree to our network, in the sense that no other junction tree can lead to faster inference. This is of particular importance given that the decoding efficiency requires a junction tree with "small" clique state-spaces. As explained in section 2.3, the first step in this construction is the moral graph. Figure 2.5 displays the moral graph of our multi-band BN. In the one-band (i.e. HMMs) or two-band cases, finding a minimal junction tree is obvious [10] because the moral graph is triangulated as it is (consider $B = 1$ or $B = 2$ in Figure 2.5).

This is not true any more when $B > 2$. Since the problem of automatically finding minimal junction trees for arbitrary BNs is NP-complete [25], we needed to find an appropriate (analytical) technique to derive a minimal junction tree for our particular B-band BN. We did so as follows :

First, it is clear from the moral graph that the only cliques which contain the variables $O_t^{(n)}$ and $G_t^{(n)}$ are of the form $H_t^{(n)}G_t^{(n)}O_t^{(n)}$. For the remaining cliques (which all contain only the hidden variables $H_t^{(n)}$), we proceed by induction. If $B = 1$, it is obvious that the clique which has to be linked to $H_t^{(1)}G_t^{(1)}O_t^{(1)}$ is $H_t^{(1)}H_{t-1}^{(1)}$. If $B = 2$ it is easy to check from the moral graph (which is triangulated as it is) that the clique which has to be linked to $H_t^{(1)}G_t^{(1)}O_t^{(1)}$ (resp. $H_t^{(2)}G_t^{(2)}O_t^{(2)}$) is $H_t^{(1)}H_{t-1}^{(1)}H_{t-1}^{(2)}$ (resp. $H_t^{(1)}H_t^{(2)}H_{t-1}^{(2)}$). Then, by induction one can prove that the clique which has to be linked to $H_t^{(n)}G_t^{(n)}O_t^{(n)}$ is $H_t^{(1)}\cdots H_t^{(n)}H_{t-1}^{(n)}\cdots H_{t-1}^{(B)}$. The time slices $t = 1$ and $t = T$ are then treated separately to remove the variables which are not necessary to satisfy the runing intersection property. The resulting junction tree is shown in Figure 2.6. We thus have a computationally optimal tree to propagate the effect of observed evidence.

The complexity of the JLO and Dawid algorithms scales as the sum of clique state-spaces. Therefore, given the asynchrony assumptions, the left-to-right topology and our junction tree, the total complexity to infer this B-band DBN is $O(MBD|I_v|^B T)$ where D is the dimension of the acoustic vectors.

2.5.3 Model parameters estimation

So far, we have assumed that parameters of the B-band DBN are known for each word. This section presents an algorithm for parameters estimation. In all the experiments we carried out, we learned the model of each word independently of the others, i.e., we did not perform embedded training. Note however that this is not a constraint of our system. Indeed, embedded training can be performed exactly as in the HMMs setting given that all graphical structures of the models are the same. Also, the use of words as acoustic units in not a constraint neither, any kind of acoustic units (phonemes, diphones...) can

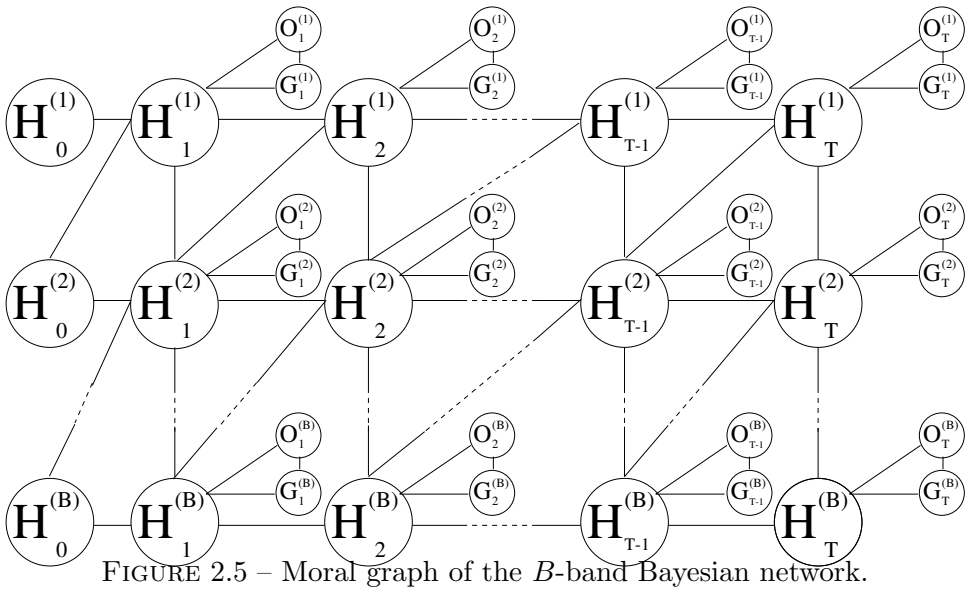


FIGURE 2.5 – Moral graph of the B-band Bayesian network.

be used without any particular change in our methodology.

In order to simplify the notation in the formulae below, we drop the reference to the word under consideration. Suppose that we have (for a given word v) an observation vector $o = (o_1^{(1)}, ..., o_T^{(1)}, ..., o_1^{(B)}, ..., o_T^{(B)})$. Let

$$\mathcal{H} = \left\{ h = (h_0^{(1)}, ..., h_T^{(1)}, ..., h_0^{(B)}, ..., h_T^{(B)}) : h_t^{(n)} \in \{1, ..., m\} \right\}$$

be the set of all possible trajectories of the hidden process defined by the $H_t^{(n)}$. Then, from (2.1) and by marginalization over $\mathcal{H}$, the likelihood is given by :

$$P(o) = \sum_{h \in \mathcal{H}} \left\{ \prod_{t=1}^{T} a_{h_{t-1}^{(1)} h_t^{(1)}} \prod_{n=2}^{B} u_{h_t^{(n-1)} h_{t-1}^{(n)} h_t^{(n)}} \right\} \left\{ \prod_{t=1}^{T} \prod_{n=1}^{B} \left(\sum_{p=1}^{M} w_{h_t^{(n)} p} b^{(n)}_{h_t^{(n)} p}(o_t^{(n)}) \right) \right\}.$$

Therefore, by aplying the EM algorithm, the auxiliary function can be decomposed as the sum of terms depending each on one component of the parameters set. Thus, solving the parameters estimation problem comes back to a simple generalization of the Baum-Welch algorithm. This is made possible essentially because we have imposed that continuous variables are conditioned by discrete ones. The re-estimation formulae are obtained as follows : suppose that we have estimated the parameters at iteration l and define for $(i, j, k, p) \in \{1, ..., m\}^3 \times \{1, ..., M\}$ (here we assume that the number of hidden states is the same for all words and equals some integer m, i.e., $|I_v| = m, \forall v$) :

$$\begin{cases} \psi_t^{(1)}(i,j) \triangleq P(H_{t-1}^{(1)} = i, H_t^{(1)} = j|o) \\ \psi_t^{(n)}(i,j,k) \triangleq P(H_t^{(n-1)} = i, H_{t-1}^{(n)} = j, H_t^{(n)} = k|o) \quad \text{for} \quad n = 2, ..., B \\ \psi^{(1)}(i,j) \triangleq \sum_{t=1}^{T} \psi_t^{(1)}(i,j) \\ \psi^{(n)}(i,j,k) \triangleq \sum_{t=1}^{T} \psi_t^{(n)}(i,j,k) \quad \text{for} \quad n = 2, ..., B \\ \phi_t^{(n)}(i,p) \triangleq P(H_t^{(n)} = i, G_t^{(n)} = p|o) \quad \text{for} \quad n = 1, ..., B \\ \phi^{(n)}(i,p) \triangleq \sum_{t=1}^{T} \phi_t^{(n)}(i,p) \quad \text{for} \quad n = 1, ..., B. \end{cases} \tag{2.7}$$

Then, the new parameters at iteration $l+1$ are given by [6]

$$a_{ij} = \frac{\psi^{(1)}(i,j)}{\sum_{k=1}^{m} \psi^{(1)}(i,k)}$$

$$u_{ijk}^{(n)} = \frac{\psi^{(n)}(i,j,k)}{\sum_{l=1}^{m} \psi^{(n)}(i,j,l)}$$

$$w_{ip}^{(n)} = \frac{\phi^{(n)}(i,p)}{\sum_{q=1}^{M} \phi^{(n)}(i,q)}$$

$$\mu_{i,p}^{(n)} = \frac{\sum_{t=1}^{T} \phi_t^{(n)}(i,p) o_t^{(n)}}{\phi^{(n)}(i,p)}$$

$$\Sigma_{i,p}^{(n)} = \frac{\sum_{t=1}^{T} \phi_t^{(n)}(i,p)(o_t^{(n)} - \mu_{i,p}^{(n)})(o_t^{(n)} - \mu_{i,p}^{(n)})^*}{\phi^{(n)}(i,p)}$$

What remains is to efficiently compute the posterior probabilities $\psi_t^{(1)}(i,j)$, $\psi_t^{(n)}(i,j,k)$ and $\phi_t^{(n)}(i,p)$ defined in equation (2.7). All these posterior probabilities can be efficiently computed using the JLO algorithm which allows the computation of marginal and conditional probabilities of clique variables. Note also that in the full-band case ($B = 1$) which collaps to an HMM, the computation of $\psi_t^{(1)}(i,j)$ and $\phi_t^{(1)}(i,p)$ using the JLO algorithm is (exactly) equivalent to the Forward-Backward algorithm.

6. For sake of notational simplicity, we drop the iteration index.

2.6 Application to isolated speech recognition

For an isolated speech recognition task, the decoding algorithm is readily given by the material described in section 2.3. Indeed, once we have learned a B-band DBN model Θ_v for each $v \in V$, then given a speaker utterance o, the likelihood $P(o|\Theta_v)$ for each $v \in V$ is computed using the JLO algorithm and the word v^* is chosen such that

$$v^* = \text{argmax}_v \; P(o|\Theta_v)$$

to be the pronounced word. The computational complexity of this decoding algorithm is $O(MBm^BT)$, where that M is the number of Gaussian components in each mixture and m is the number of hidden states.

Experiments

We evaluated the performance of the B-band DBN on an isolated digit recognition task. We compared our model to HMMs, a classical multi-band (CMB) system and a synchronous "multi-band" Bayesian network. The experiments were carried out on the isolated part of the Tidigits database [7] in which 112 (resp. 113) speakers were used for training (resp. test). Each speaker utters 11 digits twice. The parameterization for the classical full-band HMM was done as follows : 25ms frames with a frame shift of 10ms, each frame is passed through a set of 24 triangular filters resulting in a vector of 35 features, namely, 11 static MFCC (the energy is dropped), 12 Δ and 12 $\Delta\Delta$. For our model, we present experiments in the case of 2, 3 and 4-band BN. The parameterization for the 2-band DBN was done as follows : each frame was passed through the 14 first (resp. last 10) filters resulting in the acoustic vector of sub-band 1 (resp. sub-band 2). Each vector contains 17 features : 5 static MFCC, 6 Δ and 6 $\Delta\Delta$. The resulting bandwidths of sub-bands 1 and 2 were $[0..1467Hz]$ and $[1211Hz..10000Hz]$ respectively. For the 3-band BN, each frame was passed through the first 8, second 8 and last 8 filters resulting in the

7. We emphasize here that our purpose in this paper is not to perform benchmark tests, i.e., our goal here is not to tune the parameters in order to achieve the highest performances. Rather, we provide comparisons using baseline systems in order to *illustrate* the capabilities of each system and have a fair judgment on their potentail.

Model	HMM	2-band	3-band	4-band
Score	93.4%	**97.4**%	**97.3**%	**95.4**%

TABLE 2.1 – Recognition scores of the HMM and the B-band DBN ($B = 2, 3, 4$) on clean speech.

acoustic vector of sub-band 1, 2 and 3 respectively. Each vector contained 11 features : 3 static MFCC, 4 Δ and 4 $\Delta\Delta$. The resulting bandwidths of sub-bands 1, 2 and 3 were $[0..692Hz]$, $[615Hz..2152Hz]$ and $[1777Hz..10000Hz]$ respectively. Similarly, for the 4-band BN, each frame was passed through the first 6, second 6, third 6 and last 6 filters. Each resulting vector contained 8 features. The resulting bandwidths of sub-bands 1, 2, 3 and 4 were $[0..538Hz]$, $[461Hz..1000Hz]$, $[923Hz..3158Hz]$ and $[2607Hz..10000Hz]$ respectively. In all the experiments, for every digit and all models, the number of hidden states was six ($m = 6$) and we had a single Gaussian per state with a diagonal covariance matrix. Table 2.1 shows the recognition scores obtained using the 2, 3 and 4-band BNs and also the score with a classical full-band HMM. In this experiment, both train and test were on clean speech.

These results show that the three B-band BNs all outperform the HMM recognizer that we tested. Thus, taking into account the frequency dynamics leads to a higher fidelity speech modeling. One can notice however (in this experiment) that when the number of sub-bands increases the accuracy decreases. This should not be understood as a characteristic of our multi-band system. Probably one explanation is the fact that we are using the same amount of data to estimate models with an increasing number of parameters. We believe however that this behavior is mainly due to the ad-hoc choice of sub-bands bandwidths. For instance, sub-band n is more relevant than sub-band n' in the B-band DBN if $n' > n$, in the sense that it governs the behaviour of sub-band n'. Thus, for example, sub-band 1 is more relevant than all the others and in the parameterization we chose, when the number of sub-bands increases the amount of information contained in sub-band 1 decreases. The optimization of the sub-bands bandwidths was not our major concern in this work because we did not perform benchmark tests. What should be retained from

these results is that, even with such ad-hoc choice of sub-bands frequencies, the B-band DBN outperform HMMs. This was in fact a major advantage since the only multi-band systems which out-performed HMMs back then in clean conditions use the full-band parameterization as an additional "sub-band". Such a manipulation is conceptually unrealistic in our opinion and penalizes the systems in noisy conditions.

In the following experiments, training is done on clean speech and test is done on noisy speech. We show the performances of our model when the noise is additive and corrupt one spectral sub-band with two different bandwidths. We compared a 2-band DBN to HMMs and two other models. The first one is a standard 2-band model where the recombination is performed by a multi-layer perceptron (MLP), we term this model CMB-MLP. The topology of this MLP is : 22 input, 15 hidden and 11 output neurons. In the second model that we term Sync, for each frame, we concatenated the acoustic vectors of sub-band 1 and 2 and used the resulting vector (34 features) as an input for the HMM-based system. It is important to note that, since we use diagonal covariances, Sync is equivalent to a 2-band DBN where a complete synchrony is imposed between the two bands. Indeed, given that covariances are diagonal, the HMM representing Sync can be viewed as the DBN shown in Figure 2.7, where $B = 2$ and each variable H_t (resp. G_t) takes its values in the set $\{1, ..., m\}$ (resp.$\{1, ..., M\}$). The model of Figure 2.7 corresponds in turn to a completely synchronous B-band DBN. Therefore, the comparison between Sync and our 2-band DBN is a good indication about the importance of asynchrony. The issue of asynchrony in multi-band ASR has been studied by many researchers. For instance, in [38] the authors conclude that considering asynchrony in multi-band ASR may improve the acoustic modeling. They failed however, in [39], to improve the recognition performances when relaxing the synchrony constraints in a multi-band system, they then conclude that asynchrony is not advantageous. We believe that if this argument is true, it is only in the sense of incorporating asynchrony assumptions in a *classical* multi-band system. In other words, even if this argument is true, it does not mean that asynchrony does not exist or is irrelevant. It only means that it is difficult to exploit and deal with asynchrony in classical multi-band systems. As we will see below, The DBNs perspective to multi-band

Noise A : 5-10 KHz	HMM	Sync	CMB-MLP	2-band DBN
SNR 26db	53.4%	43.5%	59.5%	**84.9**%
SNR 20db	38.8%	27.2%	39.0%	**77.9**%
SNR 14db	26.3%	18.8%	23.4%	**70.8**%
SNR 8db	18.4%	11.6%	14.7%	**65.5**%
SNR 2db	13.7%	9.3%	10.4%	**63.2**%

TABLE 2.2 – Recognition scores for Noise A using the different models

ASR suggests that asynchrony is in fact advantageous.

The noisy speech (in test) is obtained by adding to the clean one, at different SNRs, band-pass filtered white noises with different bandwidths : $[5000Hz..10000Hz]$ for Noise A and $[2000Hz..7000Hz]$ for Noise B, the SNR being estimated as :

$$SNR = 10\log_{10}\left(\frac{\text{Signal Energy}}{\text{Noise Energy}}\right).$$

Therefore, in both cases, only sub-band 2 is corrupted [8]. Table 2.2 (resp. Table 2.3) gives the recognition rates obtained for Noise A (resp. Noise B) using the four models (HMM, Sync, CMB-MLP and the 2-band DBN). For both noises, the 2-band DBN largely outperformed all the other models. Even when the scores of the latter are extremely low (sometimes close to random decision), our model still yielded relatively good recognition rates. This indicates that our model is well adapted to the case of band-limited noisy speech. It is also remarkable how significant are the differences between the scores of our model and those of Sync. We believe this illustrates the importance of asynchrony in multi-band speech modeling.

2.7 Application to continuous speech recognition

In a continuous speech recognition task, given a B-band DBN model of each word in the vocabulary and a speaker utterance, the goal is to identify the most likely sequence

8. Obviously, in most real world applications one does not know a priori which sub-bands are corrupted, these has to be detected using some noise estimation algorithm.

Noise B : 2-7 KHz	HMM	Sync	CMB-MLP	2-band DBN
SNR 26db	52.3%	45.8%	54.0%	**82.7**%
SNR 20db	46.2%	32.1%	42.8%	**71.2**%
SNR 14db	39.0%	20.1%	37.3%	**60.8**%
SNR 8db	32.5%	11.3%	33.7%	**54.4**%
SNR 2db	28.6%	9.5%	26.9%	**49.4**%

TABLE 2.3 – Recognition scores for Noise B using the different models

of words given the observation. A naive solution would be to use a B-dimensional Viterbi algorithm [40] which is computationally very expensive. This section presents an efficient decoding algorithm which relies essentially on a state-augmented B-band DBN model, it then presents experiments on a connected digits recognition task.

2.7.1 Decoding algorithm

The basic idea is to build a new B-band DBN model which represents all the words in the vocabulary, decoding is then performed by inferring this new DBN. Precisely, the graphical structure of this new B-band DBN is the same as the one of Figure 2.3, the difference is that the variables do not depend any more on the word under consideration, and each variable $H_t^{(n)}$ takes now its values in the set $I = \bigcup_{v \in V} I_v$. To complete the definition of this new DBN we need to specify the conditional probabilities of the hidden and the observed variables. Let $(i, j, k, p) \in I^3 \times J$ such that $(i, j, k) \in {I_v}^3$ for some $v \in V$. Then, the observation's conditional probabilities are simply given by those corresponding to each word, namely :

$$P(O_t^{(n)} = \cdot | H_t^{(n)} = i, G_t^{(n)} = p) \stackrel{\triangle}{=} b_{i,p}^{(n)}(v, \cdot).$$

To specify the hidden process parameterization we need to include the language model, we also make some (a)synchrony assumptions : we still allow complete asynchrony *inside* a word, but we impose a full synchrony of all sub-bands when *transiting* between words. Precisely, since we have a left-to-right topology, the only non-zero conditional probabilities

are the following :

• The synchronous transition between two (not necessarily different) words v and v' :

$$P(H_t^{(1)} = 1_v | H_{t-1}^{(1)} = m_{v'}) \stackrel{\triangle}{=} P(v|v')$$

$$P(H_t^{(n)} = 1_v | H_t^{(n-1)} = 1_v, H_{t-1}^{(n)} = m_{v'}) \stackrel{\triangle}{=} P(v|v')$$

where $P(v|v')$ is given by the language model.

• The inside-word conditional probabilities :

$$P(H_t^{(1)} = j | H_{t-1}^{(1)} = i) \stackrel{\triangle}{=} a_{ij}(v)$$

$$P(H_t^{(n)} = k | H_t^{(n-1)} = i, H_{t-1}^{(n)} = j) \stackrel{\triangle}{=} u_{ijk}^{(n)}(v).$$

Now we have a completely defined B-band model on which decoding can be performed. To do so, we use the Dawid algorithm [24] which allows the identification (with the same time complexity as the JLO algorithm [23]) of the most likely sequence of hidden states given observations.

Given the (a)synchrony assumptions and the left-to-right topology, the total complexity of this decoding algorithm is $O(MBm^B T + |V|^2 T)$. Note also that in the 1-band case (i.e. HMMs), this algorithm is equivalent to Viterbi decoding.

2.7.2 Experiments

The experiments are carried out on the connected part of the Tidigits database in which 112 (resp. 113) speakers were used for training (resp. test). Each speaker utters 77 sentences resulting in 8642 sentences for training and 8701 for test, each sentence contains between 1 and 7 digits. We show comparisons [9] of the performances of a 2-band DBN with a single Gaussian per state to HMMs with a different number of Gaussian components in

9. At the time of writing we do not have a continuous version of the CMB system to show its performances. However, we expect our system to have even better performances than such a system as compared to the results obtained in the isolated task. Indeed, the latter is the ideal setting for a CMB system because there is no word-asynchrony to deal with, and it is well known that recombination is a more difficult task in the continuous setting than in the isolated one. Our system does not have such discrepancy. Also, we do not show the results of the Sync model because it always yields the lowest performances.

each mixture. For every digit and the silence model, the number of hidden states is seven ($m = 7$) and all the covariance matrices are diagonal. We used a uniform language model, i.e., $P(v|v') = \frac{1}{12}$ (eleven digits + silence). The parameterization of the classical full-band HMM was done as in the isolated task (see the previous section). The parameterization of the 2-band DBN was done as follows : each frame is passed through the 16 first (resp. last 8) filters resulting in the acoustic vector of sub-band 1 (resp. sub-band 2). Each vector contained 17 features : 5 static MFCC, 6 Δ and 6 $\Delta\Delta$. The resulting bandwidths of sub-bands 1 and 2 were $[0..2152Hz]$ and $[1777Hz..10000Hz]$ respectively. The training of all models was done on clean speech only. The test however was performed on noisy speech which was obtained by adding, at different SNRs, a band-pass filtered white noise with a bandwidth of $[3000Hz..6000Hz]$. Table 2.4 and 2.5 show respectively the digit and phrase accuracy that we obtained using both models. If one compares the 2-band DBN with HMM-1G which both have a single Gaussian per state, one sees that our model largely outperforms the HMM-1G model. One can argue that this may be due to the fact that our model uses (slightly) more parameters than the other model. The comparison between our model and the other HMMs (which have more than 2 Gaussian components per state) shows the opposite of this argument. Indeed, all these HMMs use much more parameters than the 2-band DBN, still our model yielded the best performances. Particularly, the more the SNR is low the higher is the accuracy of the 2-band DBN as compared to the HMMs. These results show the potential of our approach in exploiting the information contained in the non-corrupted sub-band. In summary, the behavior of our system in the continuous task remains consistent as compared to the isolated task and its performances on this illustrative experiment were impressive. This showed that the DBNs perspective to multi-band speech recognition was very promising, back then.

Noise 3-6 KHz	HMM-1G	HMM-2G	HMM-4G	HMM-8G	2-band DBN (1G)
SNR 26 db	89.95%	92.69%	97.20%	96.82%	**96.16**%
SNR 20 db	82.17%	85.17%	94.19%	93.59%	**94.89**%
SNR 14 db	73.27%	75.33%	87.44%	86.64%	**90.81**%
SNR 8 db	62.57%	59.57%	73.85%	72.91%	**82.27**%
SNR 2 db	58.86%	40.82%	54.60%	53.48%	**75.51**%

TABLE 2.4 – Digit accuracy rates(nG means n Gaussian components per state)

Noise 3-6 KHz	HMM-1G	HMM-2G	HMM-4G	HMM-8G	2-band DBN (1G)
SNR 26 db	71.47%	79.05%	92.00%	90.77%	**89.42**%
SNR 20 db	52.49%	59.38%	84.09%	82.65%	**85.90**%
SNR 14 db	35.69%	40.13%	69.22%	67.29%	**74.67**%
SNR 8 db	20.90%	20.55%	46.00%	43.17%	**53.86**%
SNR 2 db	10.97%	9.696%	23.82%	22.52%	**39.87**%

TABLE 2.5 – Phrase accuracy rates (nG means n Gaussian components per state)

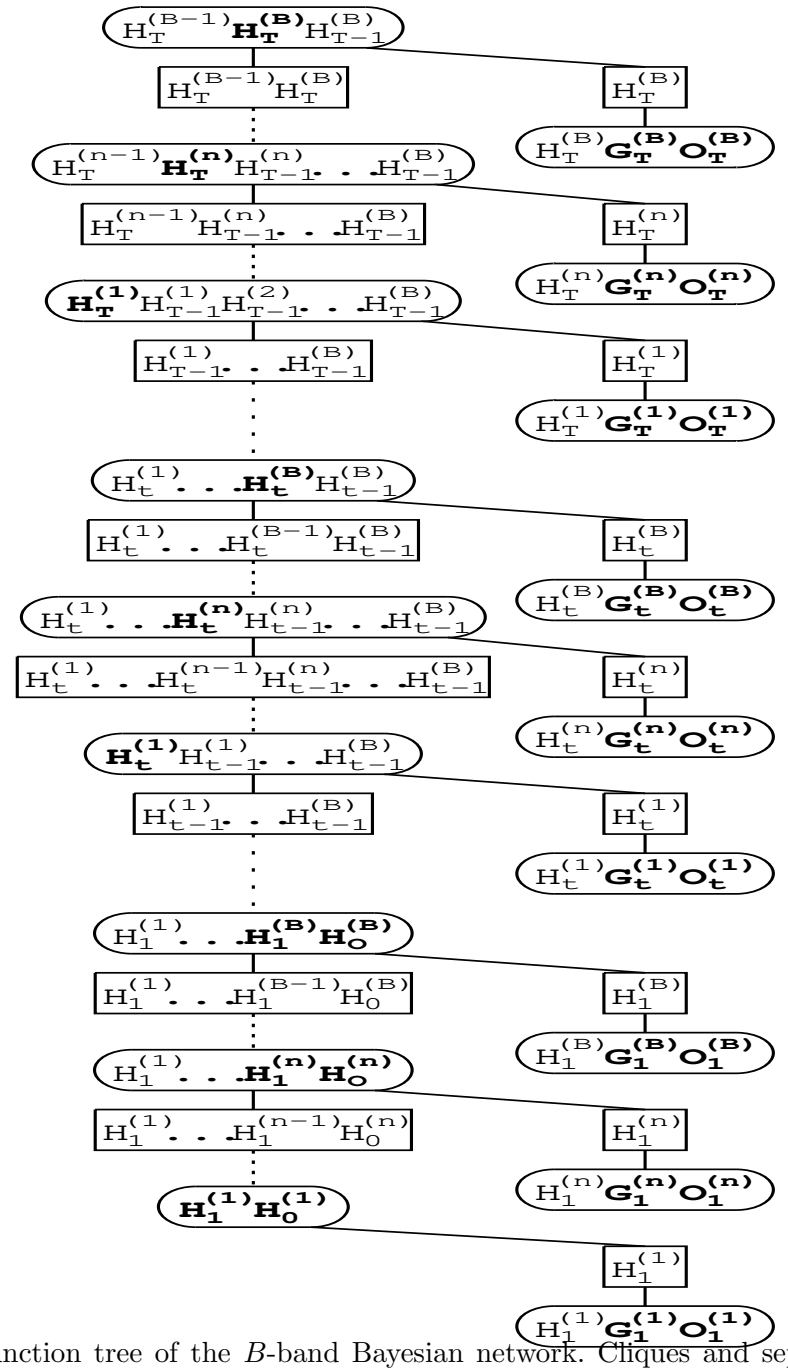


FIGURE 2.6 – Junction tree of the B-band Bayesian network. Cliques and separators are respectively represented by ellipsoids and rectangles. For each clique, the variables in bold are those which are *assigned* to that clique in order to obtain an *initial representation* of the JPD (see section 2.3).

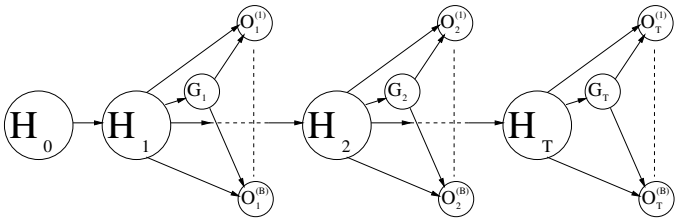


FIGURE 2.7 – Synchronous B-band dynamic Bayesian network

Chapitre 3

Sequence kernels for SVM speaker recognition

Sommaire

3.1 Introduction 65

3.2 On sequence kernels 67

3.2.1 Sequence kernels based on generative models 67

3.2.2 Other sequence kernels 68

3.2.3 Overview of the GLDS kernel 69

3.3 FSNS kernels 70

3.3.1 Definition 70

3.3.2 Dual/Kernelized form of FSNS kernels 73

3.4 Incomplete Cholesky Decomposition of the Gram Matrix 77

3.5 Mahalanobis kernels in the feature space 80

3.5.1 Definition 80

3.5.2 Dual/Kernelized form of the FSMS kernels 82

3.6 Experiments 83

3.6.1 Database and front-end processing 83

3.6.2 System implementation 84

3.6.3 Development of SVM system with FSNS/FSMS kernels 84
3.6.4 Evaluation 89

This chapter presents the sequence kernel that we had developed for SVM speaker verification. Implicitly, this kernel projects all input frames in a high/infinite-dimensional vector space, then operates a normalization in this feature space so as to induce a Mahalanobis distance, and finally computes the average of dot products between all inter-sequence pairs of frames. An existing kernel which *explicitly* performs this is the Generalized Linear Discriminant Sequence (GLDS) kernel, which has been shown to provide very good performance and efficiency at NIST Speaker Recognition Evaluations (SRE) in the last few years. The GLDS kernel is based on an explicit map of polynomial expansions which, because of practical limitations, have to be of a degree less or equal to three. The kernel presented here generalizes the GLDS kernel, and allows not only any polynomial degree but also any embedding, including infinite dimensional associated with Mercer kernels (such as Gaussian kernels). The exact "kernelized" form of this sequence kernel involves the computation of the Gram matrix on background data, and may be intractable when the background corpus is very large (which is the case in speaker verification). To overcome this problem, we used low-rank approximation of the Gram matrix to provide an approximate but tractable form of this sequence kernel. We carried out comparative experiments on NIST SRE 2005. The results showed that the this kernel outperforms the GLDS one, and gives similar (individual) performances to the traditional UBM-GMM system. As expected, the fusion of both improved the scores.

3.1 Introduction

Support Vector Machines are an interesting alternative to Gaussian Mixture Models (GMMs) for speaker verification systems based on spectral feature extraction [41], as they are well suited to separating rather complex regions in binary classification problems, in particular through the use of non linear kernels for non linear decision boundaries. A

challenge however in applying SVM to monitor conversations in a communication network, such as in NIST Speaker Recognition Evaluations (SRE), is to deal with the huge amount of data available. Thus, in order to exploit a rich database involving various types of (low quality) cell phone conversations with an SVM training algorithm, the frame-based approach such as the one in [42] needs to be adapted to make a sufficiently efficient training and testing procedure. A solution could be to use clustering methods to reduce the size of the training corpus as it was done in [43]. But an elegant solution is to use an appropriate sequence kernel. As well as significantly reducing the number of inputs in the training algorithm (a 2 minute long audio sequence is equivalent to about 6,000 acoustic vectors), it makes the training criterion more suitable as the goal in speaker verification is to classify sequences and not isolated speech frames.

The use of sequence kernels and SVM in text-independent speaker verification has gained considerable attention in recent years [44, 45, 46, 47]. A sequence kernel that has shown very good results so far in NIST SRE is the GLDS kernel [48]. It consists basically of an explicit map of each sequence to a single "supervector" in a feature space using polynomial expansions of input vectors. Then an SVM with a linear kernel is used in this feature space.

The GLDS kernel has however a practical and a theoretical limitation because of the explicit map computation. The former is due to the fact that only expansions with (relatively) low dimensions can be used in practice (thus encompassing only a limited number of non-linear correlations between input variables). The latter is due to the fact that it does not readily generalize to infinite expansions such as the ones encoding the Gaussian kernel [49]. The approach we proposed overcomes these two limitations. After a quick review of sequence kernels for text-independent speaker verification and a description of the GLDS kernel which inspired this work, a class of sequence kernels is defined in section 3.3, which amounts to

— projecting all frames in a high/infinite-dimensional vector space so as to make the data more separable, then

— processing a normalization in this feature space so as to induce a Mahalanobis distance, and finally

— computing the average of dot products between all inter-sequence pairs of frames.

We then expressed them in terms of frame kernel evaluations, so as to provide a simple finite analytic form and apply the kernel trick. An *ad hoc* approximation technique to drastically reduce the complexity of the "kernelized" form is presented in section 3.4. In section 3.5, an alternative theoretical interpretation of our kernel is presented. We focus on the application of text-independent speaker verification, where the goal is to determine whether a sequence was pronounced or not by a target speaker whatever the phonetic context. Section 3.6 shows some experiments on NIST SRE, which give an insight into the tuning of the frame kernel, and exhibit encouraging performance of this SVM system in comparison with other state-of-the-art cepstral-based systems.

3.2 On sequence kernels

This section reviews some sequence kernels that have inspired research in text-independent speaker verification. Note that in this application, only kernels between "sets of vectors" have been used (they are called "sequence kernels" for simplicity). For context independence, it is better to use such a measure that is invariant w.r.t. the frame ordering within the sequences. Note that in practice, short-term dynamic information is taken into account in the derivatives incorporated in input vectors during front-end processing.

3.2.1 Sequence kernels based on generative models

As it is appealing to combine both advantages of generative and discriminative modeling, the design of sequence kernels based on generative models has been the center of interest of several studies in speaker verification [44, 50, 51, 52].

A well known kernel in this category is the Fisher kernel [53] which has been applied and evaluated in speaker identification in [45]. It can be seen as a special case of Mutual

Information Kernels [54], for which the underlying philosophy is to first train on unlabeled data a generative model that describes the *a priori* data distribution, then to use this model to construct a kernel suited to the data. Despite its theoretical power, the Fisher kernel has not been shown to significantly outperform GMMs. Moreover, it lacks efficiency for large databases because it involves the explicit computation of high-dimensional sequence maps with a size equal to the number of model parameters. Indeed, if G is the number of Gaussian components in the (diagonal covariance) GMM, the size of the Fisher score map is $D = G(1 + 2d)$, where d is the input space dimension. In practice $\{G\,,\,d\} \sim \{500\,,\,30\}$ leads to a feature space dimension $D \sim 30,500$.

An important family of sequence kernels based on generative models are defined as kernels between probability distributions estimated on the sequences. These include the probability product kernels [55] (amongst which the expected likelihood kernel [56] and the Battacharyya kernel [57]), and kernels which are *Exponential Embeddings* [54] of distances between distributions, such as the Kullback-Leibler (KL) divergence [58, 59]. The popular approach referred to as GMM supervector kernel [51] can be actually seen as a practical implementation of the KL-divergence kernel [47].

All the kernels mentioned above are based on input space probability distributions. An another approach [60] consists of fitting simple parametric distributions in the feature space, and implicitly compute kernels between the two empirical densities in the feature space. As we will see later, a general form of the kernel we defined, the FSMS kernel, belongs to this category of kernels.

3.2.2 Other sequence kernels

Besides sequence kernels based on generative models, there exist several sequence kernels that combine similarity (or distance) measures between inter-sequence pairs of frames.

For instance, the DTW kernels [61, 62] are based on dynamic time alignment of sequences ; these kernels, not invariant to frame permutations, are not appropriate to text-

independent speaker verification. Another "local" sequence kernel, which combines a selection of maximal (inter-sequence) vector kernel values [63], has been applied to speaker verification [64]. However, these kernels do not satisfy the Mercer condition [65], which threatens the convergence of the SVM training algorithm.

Another sequence kernel, that satisfies the Mercer condition and showed very good performances in speaker verification is the GLDS kernel [48]. A second reason for the success of the GLDS kernel in systems presented at NIST SRE is its efficiency, with a specially low complexity for scoring. The next sub-section reviews the GLDS kernel as it was the basis of our work.

3.2.3 Overview of the GLDS kernel

The GLDS kernel [48] involves a polynomial expansion $\boldsymbol{\Phi}_p$, with monomials (between each combination of input vector components) up to a given degree p. For example, if $p = 2$ and $\mathrm{x} = [x_1, x_2]^{\mathrm{T}}$ is a 2-dimensional input vector, then $\boldsymbol{\Phi}_p(\mathrm{X}) = [x_1,\ x_2,\ {x_1}^2,\ x_1 x_2,\ {x_2}^2]^{\mathrm{T}}$.

The GLDS kernel between two sequences of vectors $\mathrm{X} = \{\mathrm{x}_t\}_{t=1\ldots T_{\mathrm{X}}}$ and $\mathrm{Y} = \{\mathrm{y}_t\}_{t=1\ldots T_{\mathrm{Y}}}$ is given as a dot product between average whitened expansions :

$$\kappa_{\mathrm{GLDS}}(\mathrm{X}, \mathrm{Y}) = \left[\frac{1}{T_{\mathrm{X}}} \sum_{t=1}^{T_{\mathrm{X}}} \boldsymbol{\Phi}_p(\mathrm{x}_t)\right]^{\mathrm{T}} {\mathbf{S}_p}^{-1} \left[\frac{1}{T_{\mathrm{Y}}} \sum_{s=1}^{T_{\mathrm{Y}}} \boldsymbol{\Phi}_p(\mathrm{y}_s)\right] \tag{3.1}$$

where $\mathbf{S}_p$ is the second moment matrix of polynomial expansions $\boldsymbol{\Phi}_p$ estimated on some background population, or its diagonal approximation for more efficiency.

Proposed in this way, the GLDS kernel is difficult to tune, because the size of the explicit polynomial expansion $\boldsymbol{\Phi}_p$ becomes intractable for polynomial expansions with maximal degree p higher than 3. Indeed, if d is the dimension of the input space, then the dimension of the expansion size is $D = \frac{(d+p)!}{d!p!}$. In practice d is about 25, and D becomes too large when $p > 3$ (*eg* $D = 23,751$ when $d = 25$ and $p = 4$). That is why in practice GLDS SVM based systems use an expansion with monomials up to degree 3.

An interesting problem then is to find a tractable way to compute or approximate (3.1) for any p. A more general problem is how to use the kernel trick to provide a finite-

dimensional form of (3.1) for any expansion $\boldsymbol{\phi}$ including infinite ones. By this way, radial basis function (RBF) kernels, which have been proved very efficient in most kernel learning methods [66], could also be used. This is the purpose of the next section.

3.3 FSNS kernels

Conventions and notations

All the following discussion is valid for any kind of input data, which will be noted with normal letters (x, y, b). Variable length sets of data will be written in capital letters (X, Y, B). In our experiments, the data extracted on speech frames are acoustic vectors ($\mathrm{x} \in \Re^d$). For clarity, we refer to input data as "vectors", while "supervectors" refers to points in the (high-dimensional) feature space. Column supervectors ($\boldsymbol{\phi}$, $\boldsymbol{\psi}$) and matrices ($\mathbf{S}$, $\mathbf{K}$) are noted with bold letters. $\mathbf{I}_N$ denotes the identity matrix of size N, $\mathbf{A}^{\mathrm{T}}$ the transpose of $\mathbf{A}$, and $\overline{\boldsymbol{\phi}}$ the arithmetic mean of $\boldsymbol{\phi}$ on some set. $k(\mathrm{x}, \mathrm{y})$ denotes a vector kernel (sample similarity function), and $\kappa(\mathrm{X}, \mathrm{Y})$ a kernel between sets of vectors (ensemble similarity). Note that even if all the results are written with matrix inversion $\mathbf{A}^{-1}$, they still stand for not invertible matrices if the pseudo-inversion [67] $\mathbf{A}^{\dagger}$ is taken instead of $\mathbf{A}^{-1}$.

3.3.1 Definition

The following definition extend the GLDS kernel for any expansion $\boldsymbol{\phi}$. It leads to the formulation of a rich family of sequence kernels, which we refer to as *Feature Space Normalized Sequence kernels* (FSNS) kernels.

Definition :

Let

— $\mathrm{B} = \{\mathrm{b}_i,\ i = 1, ..., N\}$ *be a set of unlabeled training vectors (background corpus),*

— $\mathbf{\Phi}$ *a (feature space) map of size* $D \leq +\infty$ *defining a Mercer kernel* k*,*

$$k(\mathrm{x}, \mathrm{y}) = \mathbf{\Phi}(\mathrm{x})^{\mathrm{T}} \mathbf{\Phi}(\mathrm{y}) \tag{3.2}$$

— $\mathbf{S}_{\mathrm{B}}$ *the expected second moment matrix of the map estimated on the background population :*

$$\mathbf{S}_{\mathrm{B}} = \tfrac{1}{N} \sum_{i=1}^{N} \mathbf{\Phi}(\mathrm{b}_i) \mathbf{\Phi}(\mathrm{b}_i)^{\mathrm{T}} \tag{3.3}$$

— ε *a positive scalar.*

The FSN (Feature Space Normalized) kernel between two vectors x *and* y *is defined as the generalized linear kernel between supervectors :*

$$\widehat{k}(\mathrm{x}, \mathrm{y}) = \mathbf{\Phi}(\mathrm{x})^{\mathrm{T}} \big(\mathbf{S}_{\mathrm{B}} + \varepsilon \mathbf{I}_{D}\big)^{-1} \mathbf{\Phi}(\mathrm{y}) \tag{3.4}$$

Given two sequences X *and* Y *of lengths* T_{X} *and* T_{Y}*, the FSNS kernel is defined as the average FSN kernel on inter–sequence pairs of vectors :*

$$\begin{aligned} \widehat{\kappa}(\mathrm{X}, \mathrm{Y}) &= \tfrac{1}{T_{\mathrm{X}} T_{\mathrm{Y}}} \sum_{t=1}^{T_{\mathrm{X}}} \sum_{s=1}^{T_{\mathrm{Y}}} \widehat{k}(\mathrm{x}_t, \mathrm{y}_s) \\ &= \overline{\mathbf{\Phi}}(\mathrm{X})^{\mathrm{T}} \big(\mathbf{S}_{\mathrm{B}} + \varepsilon \mathbf{I}_{D}\big)^{-1} \overline{\mathbf{\Phi}}(\mathrm{Y}) \end{aligned} \tag{3.5}$$

where $\overline{\mathbf{\Phi}}(\mathrm{X}) = \frac{1}{T_{\mathrm{X}}} \sum_{t=1}^{T_{\mathrm{X}}} \mathbf{\Phi}(\mathrm{x}_t)$*.*

The so defined FSNS kernel satisfies Mercer's conditions because $\mathbf{S}_{\mathrm{B}}$ is symmetric and semi-positive definite, and also because the addition of the (symmetric) regularization term guarantees the positive definiteness. FSNS kernels can exploit many high-order statistics and have three important features : averaging, normalization and regularization.

The averaging allows to easily construct a Mercer kernel, even if it is not potentially the best way to combine information when consecutive observations are not independent. An advantage of averaging is that it makes our kernel invariant to vector permutations, which, in a text-independent speaker verification application, is welcome for context independence. Remind that short-time dynamics are taken into account in the front-end processing, by incorporating derivatives in input vectors.

The normalization is done in the feature space using the second order moment matrix.

The main advantage is to ensure invariance to linear transformation in the feature space, so as to make training algorithm more stable. In fact, kernels methods such as SVM are not invariant to feature rescaling. If a feature (encoded by one component of the kernel map) has more amplitude than the others, it will have more importance in the choice of the decision function, which can threaten robustness if this feature is not relevant for the classification problem. Therefore the normalization aims at giving the same a priori importance to each feature. The use of the covariance matrix is discussed in section 3.5, let's first focus on the use of the second moment matrix, as it is done with the GLDS kernel as well as with the Fisher kernel [53].

The regularization involving a small positive constant $\varepsilon \ll \mathrm{tr}(\mathbf{S}_{\mathrm{B}})$ is needed for statistical and numerical reasons in cases where the dimension of the feature space is of the same order or larger than the number of data points [68]. Indeed, numerically, the matrix $\mathbf{S}_{\mathrm{B}}$ is usually not invertible and thus adding a constant times the identity makes it well conditioned for inversion. Moreover, regularization makes the problem well behaved statistically ; in the context of second order moment estimation, shrinkage estimators such as the one obtained by adding a constant times the identity lead empirically and theoretically to estimators with lower expected mean-squared error [69]. Note that regularizing by a multiple of the identity matrix is common, while not being the one and only alternative [70].

The GLDS kernel can be written in the form (3.5) with a polynomial expansion $\boldsymbol{\Phi}_p$ and $\varepsilon = 0$. Note that multiplying any component of this expansion by a scalar does not have any influence on the kernel value, because the normalization by the second moment matrix implies scale-invariance. By this way, the GLDS kernel can be seen as the unregularized FSNS kernel corresponding to a polynomial vector kernel $k(\mathrm{x}, \mathrm{y}) = (c + \mathrm{x} \cdot \mathrm{y})^p$ (for any strictly positive value of the constant c).

3.3.2 Dual/Kernelized form of FSNS kernels

Now the question is how to compute FSNS kernels $\widehat{\kappa}$ when the underlying map $\boldsymbol{\phi}$ is infinite dimensional. For this purpose, we introduce the following notations :

— $\mathbf{K}$ is the Gram matrix of the kernel k on the background set B. It is of size $N \times N$ and contains the values $K_{i,j} = k(\mathrm{b}_i, \mathrm{b}_j)$.

— $\boldsymbol{\psi}_{\mathrm{B}}$ is the *empirical map* [68] defined as the N-dimensional supervector

$$\boldsymbol{\psi}_{\mathrm{B}}(\mathrm{x}) = \left[\begin{array}{ccc} k(\mathrm{b}_1, \mathrm{x}) & \cdots & k(\mathrm{b}_N, \mathrm{x}) \end{array} \right]^{\mathrm{T}} \tag{3.6}$$

— $\overline{\boldsymbol{\psi}}_{\mathrm{B}}$ is the average empirical map on a sequence :

$$\overline{\boldsymbol{\psi}}_{\mathrm{B}}(\mathrm{X}) = \left[\begin{array}{c} \frac{1}{T_{\mathrm{X}}} \sum_{t=1}^{T_{\mathrm{X}}} k(\mathrm{b}_1, \mathrm{x}_t) \\ \ldots \\ \frac{1}{T_{\mathrm{X}}} \sum_{t=1}^{T_{\mathrm{X}}} k(\mathrm{b}_N, \mathrm{x}_t) \end{array} \right] \tag{3.7}$$

3.3.2.1 Relationship with Gaussian processes

Using these notations, the Woodbury matrix inversion lemma [67] allows to rewrite the FSN kernel (3.4) as :

$$\widehat{k}(\mathrm{x}, \mathrm{y}) = \varepsilon^{-1} \left[k(\mathrm{x}, \mathrm{y}) - \tfrac{1}{N} \boldsymbol{\psi}_{\mathrm{B}}(\mathrm{x})^{\mathrm{T}} \left(\tfrac{1}{N} \mathbf{K} + \varepsilon \mathbf{I}_N \right)^{-1} \boldsymbol{\psi}_{\mathrm{B}}(\mathrm{y}) \right] \tag{3.8}$$

It turns out that this expression can be identified as the posterior covariance kernel for a Gaussian process regression model [71]. Modulo the multiplicative factor $N\varepsilon^{-1}$, it corresponds to a Gaussian process with kernel k, a zero-mean Gaussian prior with variance $1/N$ for regression weights, and a zero-mean Gaussian noise with variance ε. By this way, the FSN kernel can be seen as a posterior covariance kernel learnt on some unlabeled data. (3.8) cannot however use directly in our application because it would involve the computation of $k(\mathrm{x}, \mathrm{y})$ between all inter-sequence pairs of frames. This would be untractable in application such as NIST SRE. Still, this realtion with Gaussian processes is very intersting. Indeed, this shows that the FSN (and GLDS) kernel belong to the familly of posterior covariance kernels.

3.3.2.2 Factorized form

Now assume that some kernel machine is trained on a (labeled) subset of the background corpus. Whaba's representer theorem guarantees that the optimal scoring function can be computed at any point x using a set of N basis functions of the form $\mathrm{x} \mapsto k(\mathrm{b}_i, \mathrm{x})$. When it is the case, one can use without losing information the following expression for the vector kernel [68] :

$$k(\mathrm{x}, \mathrm{y}) = \boldsymbol{\psi}_{\mathrm{B}}(\mathrm{x})^{\mathrm{T}} \mathbf{K}^{-1} \boldsymbol{\psi}_{\mathrm{B}}(\mathrm{y}) \tag{3.9}$$

In our application, training data collected from a target speaker may not be included in the (previously observed) background set [1], so we have to consider the general case. Equality (3.9) still stands if supervector $\{\boldsymbol{\phi}(\mathrm{x})\}$ or $\{\boldsymbol{\phi}(\mathrm{y})\}$ lies in the span of background supervectors $\{\boldsymbol{\phi}(\mathrm{b}_i)\}$, we will refer to this by *the span condition*. When the dimension of the feature space is extremely high, this span condition may not be fulfilled. Then computing the right term of (3.9) amounts to project supervector $\boldsymbol{\phi}(\mathrm{x})$ on the subspace spanned by $\{\boldsymbol{\phi}(\mathrm{b}_i)\}$ *i.e* to compute the kernel $\boldsymbol{\phi}_{\mathrm{B}}(\mathrm{x})^{\mathrm{T}} \boldsymbol{\phi}(\mathrm{y}) = \boldsymbol{\phi}(\mathrm{x})^{\mathrm{T}} \boldsymbol{\phi}_{\mathrm{B}}(\mathrm{y}) = \boldsymbol{\phi}_{\mathrm{B}}(\mathrm{x})^{\mathrm{T}} \boldsymbol{\phi}_{\mathrm{B}}(\mathrm{y})$. In this worst case, the loss of information due to the projection $\boldsymbol{\phi}_{\mathrm{B}}$ is minor when B represents a large corpus. Besides, this loss can be minimized when B is chosen using a criterion similar to the one used in kernel PCA [72], which will be the concern of the next section.

Under the span condition, equation (3.9) can be used in combination with (3.8) to formulate FSNS kernels in a factorized form :

$$\widehat{\kappa}(\mathrm{X}, \mathrm{Y}) = \overline{\boldsymbol{\psi}}_{\mathrm{B}}(\mathrm{X})^{\mathrm{T}} \underbrace{\left(\tfrac{1}{N}\mathbf{K}^2 + \varepsilon \mathbf{K}\right)^{-1}}_{\mathbf{R}} \overline{\boldsymbol{\psi}}_{\mathrm{B}}(\mathrm{Y}) \tag{3.10}$$

If the span condition does not hold, a similar form can be obtained when the regularization term is omitted (actually, as we will see later, we choose $\varepsilon = 0$ in our experiments). This form is given in the following proposition.

Proposition 1 :

1. The background set is used to tune the kernel, whereas the training dataset is used to build the SVM classifier.

$$\textit{if } \varepsilon = 0, \quad \widehat{\kappa}(\mathrm{X}, \mathrm{Y}) = {}_N\overline{\boldsymbol{\psi}}_{\mathrm{B}}(\mathrm{X})^{\mathrm{T}}\ \mathbf{K}^{-2}\ \overline{\boldsymbol{\psi}}_{\mathrm{B}}(\mathrm{Y}) \tag{3.11}$$

Proof :

Consider the matrix $\boldsymbol{\Phi}$ whose columns are background supervectors $\boldsymbol{\phi}(\mathrm{b}_i)$, and the thin Singular Value Decomposition (thin SVD [67]) :

$$\boldsymbol{\Phi} = \mathbf{UDV}$$

If r denotes the rank of $\boldsymbol{\Phi}$, $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices of sizes $D \times r$ and $N \times r$, and $\mathbf{D}$ is a $r \times r$ invertible diagonal matrix. Such a decomposition permits to write the Gram matrix $\mathbf{K} = \mathbf{V}\mathbf{D}^2\mathbf{V}^{\mathrm{T}}$ and the second moment matrix $\mathbf{S}_{\mathrm{B}} = \frac{1}{N}\mathbf{U}\mathbf{D}^2\mathbf{U}^{\mathrm{T}}$. The inversion gives :

$$\mathbf{S}_{\mathrm{B}}{}^{-1} = {}_N\tfrac{1}{N}\mathbf{U}\mathbf{D}^{-2}\mathbf{U}^{\mathrm{T}} = {}_N\boldsymbol{\Phi}\mathbf{V}\mathbf{D}^{-4}\mathbf{V}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}} = {}_N\boldsymbol{\Phi}\mathbf{K}^{-2}\boldsymbol{\Phi}^{\mathrm{T}}$$

We finally obtain equation (3.11) by identifying the empirical map $\boldsymbol{\psi}_B(\mathrm{x}) = \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\phi}(\mathrm{x})$.

3.3.2.3 Computational complexity

Now we discuss the advantage and the limitation of the factorized form, in terms of computational complexity.

The whitening matrices $\mathbf{R}$, identified by middle terms of equation (3.10), can be computed off-line. Likewise, a square root $\mathbf{R}^{1/2}$ can be computed by Cholesky factorization and employed to normalize sequence maps $\overline{\boldsymbol{\psi}}_{\mathrm{B}}(\mathrm{X})$. By this way, the computation of FSNS kernel values between all pairs of a collection of sequences can be done in two steps : first compute and store the whitened maps $\mathbf{R}^{1/2}\overline{\boldsymbol{\psi}}_{\mathrm{B}}$ for all sequences, and then compute dot products between all pairs. This methodology is judged to be the most efficient for the training procedure in our application. Table 3.1 presents the corresponding complexity of computing FSNS kernels, according to the different aforementioned expressions. The kernelized forms (3.8) and (3.10) perform an implicit normalization in the feature space, without having to compute any map $\boldsymbol{\phi}$ that may be infinite dimensional. Generally spea-

TABLE 3.1 – Complexity of computing FSNS kernels

	form (3.5)	form (3.8)	form (3.10)
Kernel comput. on 2 sequences	$O\big(D(Td_{+}D)\big)$	$O\big(N(Td_{+}N)$ $_{+}T^2d\big)$	$O\big(N(Td_{+}N)\big)$
Kernel comput. on n_s sequences	$O\big(n_sD(Td_{+}D)$ $_{+}{n_s}^2D\big)$	$O\big(n_sN(Td_{+}N)$ $_{+}{n_s}^2(N_{+}T^2d)\big)$	$O\big(n_sN(Td_{+}N)$ $_{+}{n_s}^2N\big)$

d : Input Space dimension (size of x)
D : Feature Space dimension (size of $\boldsymbol{\Phi}(\mathrm{x})$)
N : Number of background vectors (size of $\boldsymbol{\Psi}_{\mathrm{B}}(\mathrm{x})$)
T : Sequences length

king, they are worthy when $N < D$. In contrast with the exact form (3.8), the factorized form (3.10) permits to save $O(T^2d)$ operations for each occurring pair of sequences, and above all to reduce considerably the computational complexity of the testing procedure by means of the following model compacting. If $\{\mathrm{X}_i\}$ denotes training sequences and $\{\alpha_i\}$ denotes Lagrangian coefficients learnt from an SVM training, the score function can be computed as a generalized linear model :

$$\begin{aligned} f(\mathrm{X}) &= \textstyle\sum_i \alpha_i \widehat{\kappa}(\mathrm{X}_i, \mathrm{X}) + \alpha_0 \\ &= \boldsymbol{\omega}^{\mathrm{T}}\, \overline{\boldsymbol{\Psi}}_{\mathrm{B}}(\mathrm{X}) + \alpha_0 \end{aligned} \tag{3.12}$$

where the supervector $\boldsymbol{\omega} = \sum_i \alpha_i \mathbf{R}\overline{\boldsymbol{\Psi}}_{\mathrm{B}}(\mathrm{X}_i)$ encompasses discriminative model's information.

Finally, the form (3.10) could be satisfactory in small size problems. However in large scale problems, such as speaker verification, it would be burdensome to compute because the number of background frames available is very high. In the case of monitoring conversations, the size N of background data available can be enormous. We thus need to find a way to make efficient the computation of (3.10) in large scale problems. In the next section, we use a low-rank matrix decomposition to provide an approximate but tractable form for (3.10).

3.4 Incomplete Cholesky Decomposition of the Gram Matrix

Kernel methods have the advantage of allowing to work implicitly on numerous non-linear complex features and leads to algorithms that manipulate Gram matrices of the form $\mathbf{K}_{i,j} = k\big(\mathrm{x}_i, \mathrm{x}_j\big)_{(\mathrm{x}_i,\mathrm{x}_j)\in \mathrm{B}^2}$. However, when the number of data points B is large, then these methods, if applied naively, may become computationally too intensive and much research in machine learning has been done to lower this complexity. One of the tools that has proved simple and efficient is the use of low-rank approximations of the Gram matrix [73, 74]. In our context, the goal is to pick up a subset $\mathrm{C} \subset \mathrm{B}$ that would lead to a good approximation of the Gram matrix $\mathbf{K}_{i,j} = k\big(\mathrm{x}_i, \mathrm{x}_j\big)_{(\mathrm{x}_i,\mathrm{x}_j)\in \mathrm{B}^2}$ using only a sub-matrix of the full matrix, so as to rewrite kernel formulas with lower complexity.

An appealing technique is the Incomplete Cholesky Decomposition (ICD). This greedy algorithm, used in [74] to reduce SVM complexity and also described in [75], has a relatively low complexity $O(m^2N)$, if m is the desired size of the representative set C. Besides, it does not require to keep in memory or compute the entire Gram matrix $\mathbf{K}$ at any time.

Given a Gram matrix $\mathbf{K}$ of size $N\times N$ (the actual rank of $\mathbf{K}$ may be smaller than N), the ICD of $\mathbf{K}$ is a $N\times m$ matrix $\mathbf{G}$, such that $\mathbf{K}$ can be approximated by $\mathbf{L} = \mathbf{G}\mathbf{G}^{\mathrm{T}}$. The approximate square root $\mathbf{G}$, with rank $m < N$, is spanned by the columns of $\mathbf{K}$ indexed by a list denoted by $\mathrm{I} = \{I_1, \ldots, I_m\} \subset \{1, \ldots, N\}$. This list corresponds to a codebook $\mathrm{C} = \{\mathrm{b}_{I_1}, \ldots, \mathrm{b}_{I_m}\}$, and leads to a low-rank approximation of the form [76] :

$$\mathbf{L} = \mathbf{K}(:, \mathrm{I})\mathbf{K}(\mathrm{I}, \mathrm{I})^{-1}\mathbf{K}(:, \mathrm{I}) \tag{3.13}$$

where $\mathbf{K}(:, \mathrm{I})$ denotes the sub-matrix of the columns of Gram matrix $\mathbf{K}$ indexed by the elements of I, and $\mathbf{K}(\mathrm{I}, \mathrm{I})$ denotes the columns and rows indexed by I. This last squared matrix is nothing but the Gram matrix on the codebook C.

Equation (3.13) is similar to the empirical map identity (3.9), and replacing $\mathbf{K}$ by $\mathbf{L}$ means that all background supervectors $\boldsymbol{\phi}(\mathrm{b}_i)$ are approximated by their projections on the codebook supervectors $\{\boldsymbol{\phi}(\mathrm{b}_{I_1}), \ldots, \boldsymbol{\phi}(\mathrm{b}_{I_m})\}$. The ICD picks out the codebook

elements so as to minimize the trace [2] of the residual $\mathrm{tr}(\mathbf{K}-\mathbf{L})$ [74]. It is straightforward to show that this criterion is the mean squared error induced by the aforementioned projection, estimated on the background supervectors.

The following proposition shows how to approximate our sequence kernel with a tractable form involving the fixed-size codebook C instead of all background data.

Proposition 2 :

The ICD approximation of the kernel $\widehat{\kappa}$ is given by

$$\widehat{\kappa}_{\mathrm{ICD}}(\mathrm{X},\mathrm{Y}) = \overline{\boldsymbol{\psi}}_{\mathrm{C}}(\mathrm{X})^{\mathrm{T}}\ \mathbf{R}\ \overline{\boldsymbol{\psi}}_{\mathrm{C}}(\mathrm{Y}) \tag{3.14}$$

where

$$\mathbf{R} = \left(\tfrac{1}{N}\mathbf{K}(:,\mathrm{I})^{\mathrm{T}}\mathbf{K}(:,\mathrm{I}) + \varepsilon\mathbf{K}(\mathrm{I},\mathrm{I})\right)^{-1} \tag{3.15}$$

and where we define similarly to (3.7) the sequence empirical map on the codebook

$$\overline{\boldsymbol{\psi}}_{\mathrm{C}}(\mathrm{X}) = \begin{bmatrix} \frac{1}{T_{\mathrm{X}}}\sum_{t=1}^{T_{\mathrm{X}}} k(\mathrm{b}_{I_1},\mathrm{x}_t) \\ \dots \\ \frac{1}{T_{\mathrm{X}}}\sum_{t=1}^{T_{\mathrm{X}}} k(\mathrm{b}_{I_m},\mathrm{x}_t) \end{bmatrix}$$

Proof :

Projecting all background supervectors on the codebook in the feature space comes down to replace the Gram matrix by $\mathbf{L}$ (3.13) and the empirical map (3.6) by :

$$\boldsymbol{\psi}_{\mathrm{B}}(\mathrm{x}) \approx \mathbf{K}(:,\mathrm{I})\mathbf{K}(\mathrm{I},\mathrm{I})^{-1}\boldsymbol{\psi}_{\mathrm{C}}(\mathrm{x})$$

Taking into account these approximations and the fact that the Gram matrix $\mathbf{K}(\mathrm{I},\mathrm{I})$ provided by the ICD is invertible, the Woodbury matrix inversion lemma gives after some lines of calculus :

$$\begin{aligned}\widehat{k}(\mathrm{x},\mathrm{y}) \approx \varepsilon^{-1}\big[k(\mathrm{x},\mathrm{y}) - \boldsymbol{\psi}_{\mathrm{C}}(\mathrm{x})^{\mathrm{T}}\ \mathbf{K}(\mathrm{I},\mathrm{I})^{-1}\ \boldsymbol{\psi}_{\mathrm{C}}(\mathrm{y})\big] \\ +\boldsymbol{\psi}_{\mathrm{C}}(\mathrm{x})^{\mathrm{T}}\left(\tfrac{1}{N}\mathbf{K}(:,\mathrm{I})^{\mathrm{T}}\mathbf{K}(:,\mathrm{I}) + \varepsilon\mathbf{K}(\mathrm{I},\mathrm{I})\right)^{-1}\boldsymbol{\psi}_{\mathrm{C}}(\mathrm{y})\end{aligned}$$

The first term vanishes if we project all supervectors $\boldsymbol{\phi}(\mathrm{x})$, $\boldsymbol{\phi}(\mathrm{y})$ on the codebook supervectors. We then extend the result to average sequence kernels by linearity to show the

2. The trace is a norm for positive semidefinite matrices and it can be shown that the residual is positive semidefinite.

proposition (3.14).

The computational complexity of the approximation $\widehat{\kappa}_{ICD}(\mathrm{X},\mathrm{Y})$ is now $O(m^2)$ instead of $O(N^2)$. In practice the value of m can be chosen to be much lower than N which leads in turn to an efficient kernel computation. We thus achieved (with (3.14)) an efficient algorithm to compute the FSNS kernels for any choice of the feature-space map $\boldsymbol{\phi}$ associated with a vector kernel k.

Note that computing a low-rank approximation via the ICD has itself a regularizing effect, as it implicitly performs dimensionality reduction (similar to [77], where a finite dimensional projection in the feature space is applied to regularize). Moreover, by construction of the ICD, the matrix $\mathbf{K}(:,\mathrm{I})^{\mathrm{T}}\mathbf{K}(:,\mathrm{I})$ is always numerically invertible. This allows to choose a simpler expression with $\varepsilon = 0$ in (3.15). Experiments confirm that taking a non zero ε regularization does not improve performance.

The idea behind complexity reduction is to consider a low-rank approximation of the Gram matrix, which amounts to a projection in a sub-space of the Feature Space. Such methods have been applied to save computations in kernel methods, with the goal to minimize information loss [73, 78, 74]. Even if the Gram matrix is full-rank, its eigenspectrum can decreases exponentially in many cases where the kernel is adapted to the data distribution [79]. Small eigenvalues correspond to axes of the feature space where the background data have small variance. These axes are discarded in low-rank approximation method likewise in kernel PCA [72]. Even if kernel PCA is optimal in the sense of information loss, it is not suitable to reduce the complexity in our case. The main reason is that principal eigenvectors are expressed in term of *all* background vectors because of the preimage problem [80]. The form (3.14) for kernel PCA with m principal eigenvectors $\{\boldsymbol{\alpha_1},...,\boldsymbol{\alpha_m}\}$ would involve a map with the form :

$$\left[\ \boldsymbol{\psi}_{\mathrm{B}}(\mathrm{x})^{\mathrm{T}}\boldsymbol{\alpha}_1\ \cdots\ \boldsymbol{\psi}_{\mathrm{B}}(\mathrm{x})^{\mathrm{T}}\boldsymbol{\alpha}_m\ \right]^{\mathrm{T}} \text{ instead of } \boldsymbol{\psi}_{\mathrm{C}}(\mathrm{x})\,.$$

Even if the dimension m of this map can be much lower than N, the kernel computation complexity remains $O(Nm)$. A natural solution to save computations using the kernel trick is to select a codebook in the input space like it is done in proposition 2. Another

method than the ICD to perform this goal is sparse kernel PCA [81] in which weight vectors α_i are sparse. But this method is cumbersome to run in our case, because of its complexity $O(N3)$ and its memory requirement $O(N2)$.

3.5 Mahalanobis kernels in the feature space

3.5.1 Definition

A natural variant of the FSNS kernels $\widehat{\kappa}$ is to replace in (3.5) the second moment matrix $\mathbf{S}_{\mathrm{B}}$ with the empirical covariance matrix :

$$\mathbf{\Sigma}_{\mathrm{B}} = \frac{1}{N}\sum_i \boldsymbol{\phi}(\mathrm{b}_i)\boldsymbol{\phi}(\mathrm{b}_i)^{\mathrm{T}} - \frac{1}{N^2}\Big(\sum_i \boldsymbol{\phi}\big(\mathrm{b}_i\big)\Big)\Big(\sum_i \boldsymbol{\phi}\big(\mathrm{b}_i\big)\Big)^{\mathrm{T}} \tag{3.16}$$

This would provide a more intuitive interpretation to the resulting sequence kernel that we note

$$\widetilde{\kappa}(\mathrm{X},\mathrm{Y}) = \overline{\boldsymbol{\phi}}(\mathrm{X})^{\mathrm{T}}\big[\underbrace{\mathbf{\Sigma}_{\mathrm{B}} + \varepsilon\mathbf{I}_D}_{\mathbf{\Sigma}_{\mathrm{reg}}}\big]^{-1}\overline{\boldsymbol{\phi}}(\mathrm{Y}) \tag{3.17}$$

Indeed the distance induced by the kernel $\widetilde{\kappa}$ is the Root-Mean-Square of the Mahalanobis distances (in the feature space) between all inter-sequence pairs of frames :

$$\begin{aligned}\tilde{d}(\mathrm{X},\mathrm{Y}) &= \sqrt{\tilde{\kappa}(\mathrm{X},\mathrm{X})^2 - 2\tilde{\kappa}(\mathrm{X},\mathrm{Y}) + \tilde{\kappa}(\mathrm{Y},\mathrm{Y})^2} \\ &= \sqrt{\big(\overline{\boldsymbol{\phi}}(\mathrm{X})-\overline{\boldsymbol{\phi}}(\mathrm{Y})\big)^{\mathrm{T}}\mathbf{\Sigma}_{\mathrm{reg}}^{-1}\big(\overline{\boldsymbol{\phi}}(\mathrm{X})-\overline{\boldsymbol{\phi}}(\mathrm{Y})\big)}\end{aligned}$$

We thus call the family of kernels $\widetilde{\kappa}$ *Feature Space Mahalanobis Sequence* (FSMS) kernels. Fig.3.1 illustrates that our kernel is a dot product between whitened centroids of sets of vectors. Note that with a kernel machine invariant to translation in the feature space, such as SVM, using a FSMS kernel is equivalent to using a FSNS kernel with centered map $\boldsymbol{\phi}(\mathrm{x}) - \frac{1}{N}\sum_i \boldsymbol{\phi}(\mathrm{b}_i)$.

Now assume that sequences' supervectors $\{\boldsymbol{\phi}(\mathrm{x}_t)\}_{t=1\ldots T_{\mathrm{X}}}$ and $\{\boldsymbol{\phi}(\mathrm{y}_t)\}_{t=1\ldots T_{\mathrm{X}}}$ are i.i.d samples of two respective random vectors, and that the latter have Gaussian densities with equal covariance (such an assumption is made for example in LDA [82]). Then, as the empirical means are respectively $\overline{\boldsymbol{\phi}}(\mathrm{X})$ and $\overline{\boldsymbol{\phi}}(\mathrm{Y})$, and as the empirical regularized

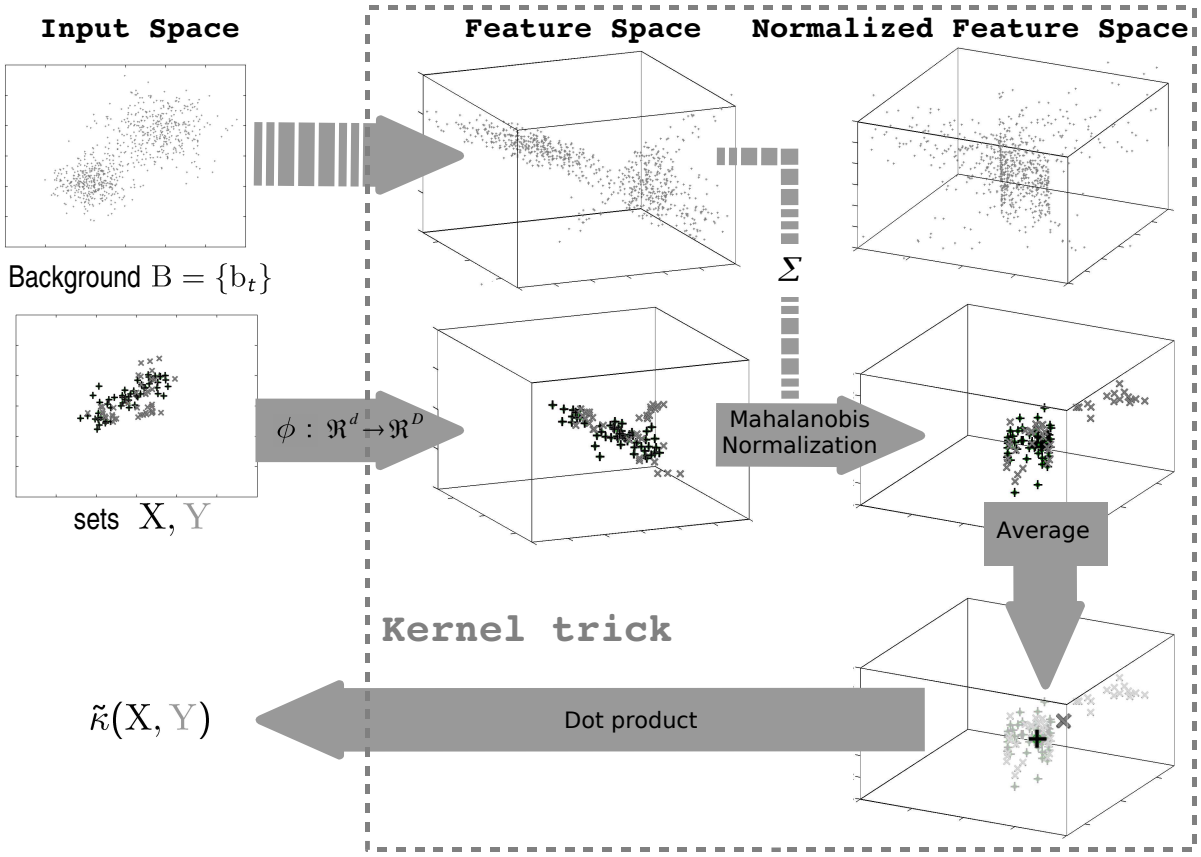


FIGURE 3.1 – Simplified view of what implicitly performs the FSMS kernel.

covariance on the background population is $\mathbf{\Sigma}_{\text{reg}} = \mathbf{\Sigma}_{\text{B}} + \varepsilon \mathbf{I}_D$, then the kernel (3.17) induces a Mahalanobis distance between Gaussian empirical means, which corresponds to the symmetric KL-divergence between estimated densities [60] (that is, in this special case, proportional to the Bhattacharyya distance [83]). By this way, the FSMS kernels can be seen as a kernels between probability densities in the feature space.

3.5.2 Dual/Kernelized form of the FSMS kernels

The following proposition shows that the results of the two previous sections can be readily extended to the FSMS kernels $\widetilde{\kappa}$.

Proposition 3 :

— *Using the span condition suggested by (3.9), the factorized kernelized form of FSMS kernels is :*

$$\widetilde{\kappa}(\mathrm{X},\mathrm{Y}) = \overline{\boldsymbol{\psi}}_{\text{B}}(\mathrm{X})^{\text{T}} \left[\tfrac{1}{N}\mathbf{K}\mathbf{\Pi}\mathbf{K} + \varepsilon \mathbf{I}_N\right]^{-1} \overline{\boldsymbol{\psi}}_{\text{B}}(\mathrm{Y}) \tag{3.18}$$

where $\mathbf{\Pi} = \mathbf{I}_N - \frac{1}{N}\mathbf{J}_N$ *is the centering matrix (*$\mathbf{J}_N$ *denotes the* $N \times N$ *matrix of all ones).*

— *Without any span condition, FSMS kernels without regularization can be written :*

$$\text{if } \varepsilon = 0, \quad \widetilde{\kappa}(\mathrm{X},\mathrm{Y}) = \overline{\boldsymbol{\psi}}_{\text{B}}(\mathrm{X})^{\text{T}} \left[\tfrac{1}{N}\mathbf{K}\mathbf{\Pi}\mathbf{K}\right]^{-1} \overline{\boldsymbol{\psi}}_{\text{B}}(\mathrm{Y}) \tag{3.19}$$

— *The ICD approximation of* $\widetilde{\kappa}$ *is given by*

$$\widetilde{\kappa}_{\text{ICD}}(\mathrm{X},\mathrm{Y}) = \overline{\boldsymbol{\psi}}_{\text{C}}(\mathrm{X})^{\text{T}}\, \mathbf{R}\, \overline{\boldsymbol{\psi}}_{\text{C}}(\mathrm{Y}) \tag{3.20}$$

where

$$\mathbf{R} = \left(\tfrac{1}{N}\mathbf{K}(:,\mathrm{I})^{\text{T}}\mathbf{\Pi}\mathbf{K}(:,\mathrm{I}) + \varepsilon \mathbf{K}(\mathrm{I},\mathrm{I})\right)^{-1} \tag{3.21}$$

Proof :

If $\mathbf{\Phi}$ denotes the $D \times N$ matrix containing all background supervectors $\boldsymbol{\phi}(\mathrm{b}_i)$, the expected covariance matrix can be written $\mathbf{\Sigma}_{\text{B}} = \frac{1}{N}\mathbf{\Phi}\mathbf{\Pi}\mathbf{\Phi}^{\text{T}}$. The Woodbury matrix inversion lemma

then gives :

$$\widetilde{k}(\mathrm{x},\mathrm{y}) = \varepsilon^{-1}\Big[k(\mathrm{x},\mathrm{y}) - \tfrac{1}{N}\boldsymbol{\psi}_{\mathrm{B}}(\mathrm{x})^{\mathrm{T}}\boldsymbol{\Pi}\big(\tfrac{1}{N}\boldsymbol{\Pi}\mathbf{K}\boldsymbol{\Pi} + \varepsilon\mathbf{I}_N\big)^{-1}\boldsymbol{\Pi}\boldsymbol{\psi}_{\mathrm{B}}(\mathrm{y})\Big]$$

Results (3.18) and (3.20) can be shown after some lines of calculus using the empirical map identity (3.9) as it was done in the two previous sections. Identity (3.19) is obtained using the same proof as for proposition 1 with a thin SVD of $\boldsymbol{\Phi}\boldsymbol{\Pi}$ (instead of $\boldsymbol{\Phi}$ itself).

3.6 Experiments

3.6.1 Database and front-end processing

All sequences used for development and evaluation come from the NIST SRE database in "core test" condition [84], limited to the female population. They include about two minutes of telephone speech pronounced by a same speaker. The development protocol was defined by the Biosecure project [85] and involves distinct sets of speakers by means of audio sequences from NIST 2003 and 2004 SREs. The *background* database includes 283 sequences, that correspond to about 9 hours of speech. Besides, 113 additional sequences which involve other speakers are available : they can be used to compute statistics for score normalization, or to increase impostor accesses for discriminative training. The validation corpus consists in 7062 trials that involve 181 target speaker and 368 test sequences. After having tuned the system to perform as well as possible on the validation set, NIST SRE 2005 is used to measuring the actual Detection Cost Function (DCF). This criterion to minimize is a weighted sum of False Rejection and False Alarm rates : $\mathrm{DCF} = 0.1\mathrm{FR}\% + 0.99\mathrm{FA}\%$. We insist on the fact that development, validation and evaluation involve non-overlapping sets of speakers.

As we concern ourselves with modeling strategy, we employed a classical front-end processing for speaker verification. To extract acoustic vectors from a speech sequence, 12 MFCC and their first order time derivatives are extracted on 16ms window, at a 10ms frame rate. The derivative of the energy logarithm is also added. Then, a speech activity detector discards silence frames, using an unsupervised bi-Gaussian modeling on

the energy level [86]. Finally, the 25-dimensional input vectors are normalized by *feature warping* [87] over 3 seconds windows.

3.6.2 System implementation

The first step is to run the ICD on the Gram matrix of background population. In our case, it would be computationally expensive to run this iterative algorithm on a huge amount N of data, as we need to memorize $O(Nm)$ real values. Our experiments showed that it is not necessary to consider all background data available. We have roughly the same performance when considering $20,000$ background vectors or $200,000$ background vectors. We thus run the ICD on the Gram matrix of $20,000$ background vectors picked up randomly in the background corpus.

In our SVM speaker verification scheme, we have to train several target speaker models using a common set of background sequences considered as impostors, whose sequence maps are computed off-line and kept in memory. To give an idea, about 15% of training sequences are retained as "support sequences" after SVM training with the (baseline) GLDS kernel as well as with (fine tuned) FSNS kernels.

3.6.3 Development of SVM system with FSNS/FSMS kernels

In this section, we discuss how to tune the parameters of the FSNS kernel, of the form :

$$\widehat{\kappa}(\mathrm{X},\mathrm{Y}) = \overline{\boldsymbol{\psi}_{\mathrm{C}}}(\mathrm{X})^{\mathrm{T}} \left[\tfrac{1}{N}\mathbf{K}(:,\mathrm{I})^{\mathrm{T}}\mathbf{P}\mathbf{K}(:,\mathrm{I}) + \varepsilon\mathbf{K}(\mathrm{I},\mathrm{I})\right]^{-1} \overline{\boldsymbol{\psi}_{\mathrm{C}}}(\mathrm{Y})$$

where ε is a positive scalar and where

$$\mathbf{P} = \begin{cases} \mathbf{I}_N & \text{for FSNS kernels (without centering)} \\ \boldsymbol{\Pi} & \text{for FSMS kernels} \end{cases}$$

The different parameters that characterize FSNS kernels are summed up in the following list :

— Application or not of centering in the Feature Space ;

— Value of regularization parameter ε ;

— Method of selection and size of the codebook C ;

— General form and parameters values of the vector kernel k.

Other parameters involved by the use of SVM include the soft-margin cost parameter (for model training) and the score threshold for decision taking. They are both chosen so as to minimize the DCF on the validation set.

3.6.3.1 Kernel normalization and regularization strategies

Experiments showed that taking a non-zero regularization parameter ε do not improve robustness. This suggests that applying ICD to compute FSNS kernels amounts to apply some regularization in the Feature Space. Besides, applying a projection matrix $\mathbf{\Pi}$ as was suggested in (3.21) only slightly changes performance. Table 3.2 show corresponding results for a Gaussian kernel k with $m = 5,000$ codebook vectors picked up with an ICD. In the following sections, all the experiments are thus presented with $\varepsilon = 0$ and $\mathbf{P} = \mathbf{I}_N$.

TABLE 3.2 – Performance on the validation set according to normalization strategy

Centering	Reg.	EER(%)	minDCF($\times 10^{-3}$)
none	$\varepsilon = 0$	10.95	49.9
$\mathbf{P} = \mathbf{\Pi}$	$\varepsilon = 0$	**10.62**	**49.3**
$\mathbf{P} = \mathbf{\Pi}$	$\varepsilon = 10^{-6}$	14.11	53.1
$\mathbf{P} = \mathbf{\Pi}$	$\varepsilon = 10^{-4}$	15.83	63.5

3.6.3.2 Choice of the codebook

As regards the choice of the codebook, we experimented some unsupervised Vector Quantization (VQ) algorithms [88] that are independent of the choice of the kernel (contrary to the ICD), as it was done in [89]. Results show that using the ICD leads to a better robustness. The gain in performance is relatively small when the vector kernel is suited to the data, as shown in Table 3.3. Therefore, a good strategy to develop our new system is to choose a first codebook with a classic VQ, then to tune the vector kernel,

and finally to compute a new codebook using ICD based on the vector kernel that leaded to the best results.

TABLE 3.3 – Performance on the validation set according to the choice of codebook (size 4096)

Kernel	Codebook extraction	EER(%)	minDCF($\times 10^{-3}$)
Gaussian	ICD	**11.61**	**50.3**
Gaussian	VQ	12.42	51.9
Polynomial	ICD	**12.13**	**51.4**
Polynomial	VQ	12.92	53.8

The number of input vectors in the codebook controls the approximation level but has an influence on the computational complexity of the kernel computations. Through this parameter, a compromise has to be done between robustness and efficiency. Table 3.4 shows that increasing the codebook size generally improves performance, up to a certain point where results remain roughly the same. We thus choose $m = 5,000$ in all the following experiments.

TABLE 3.4 – Performance on the validation set according to codebook size (Gaussian kernel)

Codebook size	EER(%)	minDCF($\times 10^{-3}$)
600	14.38	61.2
1,250	13.59	57.4
2,500	12.52	53.5
5,000	10.95	49.9
8,000	**10.17**	**49.8**

3.6.3.3 Choice of the vector kernel k

Considering a polynomial kernel $k(\mathrm{x}, \mathrm{y}) = (c + \mathrm{x}^{\mathrm{T}}\mathrm{y})^p$, first experiments showed that performance is better when taking a non-zero c. This means that it is better to take into account all monomials with a degree equal or lower than p, as it is done with the GLDS kernel (when $c = 0$ only monomials with degree p are taken into account). We also checked

that a FSNS polynomial kernel with $c > 0$ and degree $p = 3$ gives the same results as the GLDS kernel with a full whitening matrix $\mathbf{S}_p$. Improvement can be made by increasing the degree to a certain extent. Table 3.5 show the main results concerning the tuning of polynomial kernels.

TABLE 3.5 – Performance on the validation set for polynomial kernels

Sequence kernel	EER(%)	minDCF(x10^{-3})
GLDS (full $\mathbf{S}_p$)	12.21	52.6
FSNS, $k(\mathrm{x},\mathrm{y}) = (1 + \mathrm{x}^{\mathrm{T}}\mathrm{y})^3$	12.14	52.6
FSNS, $k(\mathrm{x},\mathrm{y}) = (1 + \mathrm{x}^{\mathrm{T}}\mathrm{y})^4$	11.82	51.7
FSNS, $k(\mathrm{x},\mathrm{y}) = (1 + \mathrm{x}^{\mathrm{T}}\mathrm{y})^5$	**11.61**	**50.8**
FSNS, $k(\mathrm{x},\mathrm{y}) = (1 + \mathrm{x}^{\mathrm{T}}\mathrm{y})^7$	12.00	52.8
FSNS, $k(\mathrm{x},\mathrm{y}) = (\mathrm{x}^{\mathrm{T}}\mathrm{y})^5$	14.66	59.8

The best results were obtained using a Gaussian RBF kernel $k(\mathrm{x},\mathrm{y}) = e^{-\gamma\|\mathrm{x}-\mathrm{y}\|^2}$. To tune the width parameter γ, [68] recommends to choose it in the order of $\gamma_\mathrm{o} = 1/2d\overline{\sigma}^2$, where $\overline{\sigma}^2$ is the mean of the variance of each component of input vectors. With the front-end processing used in our experiments, it corresponds to $\gamma_\mathrm{o} \approx 2{\times}10^{-2}$. Our experiments confirm this recommendation, as shown in table 3.6. Indeed, if γ is too high, the vector

TABLE 3.6 – Performance on the validation set of FSNS kernels with Gaussian RBF kernels $k(\mathrm{x},\mathrm{y}) = e^{-\gamma\|\mathrm{x}-\mathrm{y}\|^2}$

Value of γ	EER(%)	minDCF(x10^{-3})
$\gamma_\mathrm{o}/4$	15.7	60.2
$\gamma_\mathrm{o} = 2{\times}10^{-2}$	**10.95**	**49.9**
$4\gamma_\mathrm{o}$	11.90	50.5

kernel will fit too much to data : $\mathbf{K}$ will be close to identity matrix (maximal rank), and our sequence expansion defined in (3.14) will amount to counting how many frames from a sequence lie in a narrow neighborhood of each respective codebook vector. On the contrary, if γ is too low, the eigenvalues of $\mathbf{K}$ will decrease rapidly (low numerical rank) and we will only consider only a small number of features. Experiments show that if γ

lies in a reasonably range around γ_o, then performance is roughly the same (resp. are degraded when γ reaches some obsolete values).

3.6.3.4 Score normalization

A common method used in speaker verification to reduce the false alarm rate is the T-Norm score normalization [90]. This method amounts to training some impostor models (113 sequences in our protocol) and normalizing the score of a given test sequence with the empirical statistics (mean μ / standard deviation σ) of the scores on these impostor models : score $\mapsto \frac{\text{score}-\mu}{\sigma}$.

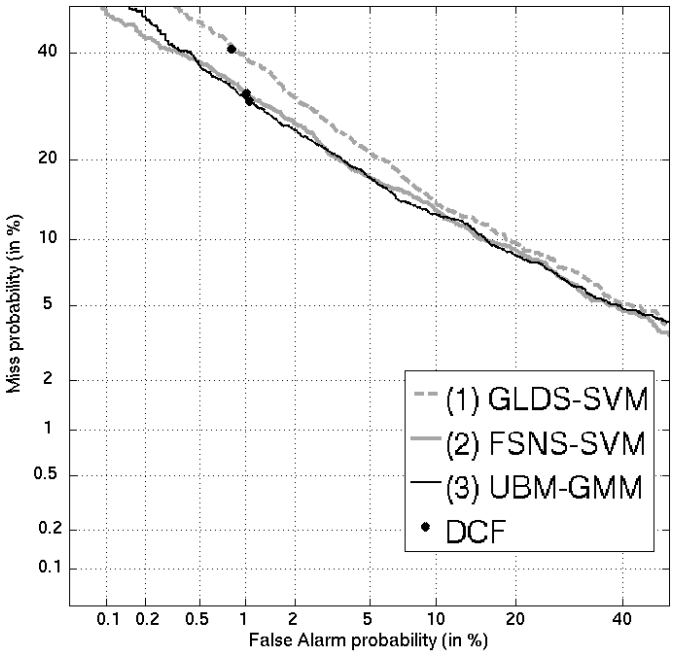
Even if T-Norm leads to a high decrease of the DCF with standard GMM systems [91], it only improves a bit with our SVM approach. An explanation is suggested by [48], who interpret the SVM score (3.12) as a kernel score with an appropriate cohort normalization. In our experiments, the gain in performance when applying T-Norm to the scores is comparable to the gain obtained when adding impostor sequences (normally used through models for T-Norm) to the set of training utterances (impostor accesses), as shows table 3.7. We chose this last strategy for all SVM systems since it leads to a much higher test efficiency.

TABLE 3.7 – Performance on the validation set of FSNS kernels according to the use of impostor utterances

No. of impostor training utterances	T-Norm	EER(%)	minDCF($\times 10^{-3}$)
283	_	10.95	49.9
283	113 *imp.models*	11.35	**47.0**
283+113	_	**10.55**	**47.3**

3.6.4 Evaluation

The best performance the novel SVM system was obtained using a Gaussian kernel with $\gamma = 2{\times}10^{-2}$. In this section, we compare this system with a SVM system based on the GLDS kernel in its current implementation [48], and with a state-of-the-art UBM-GMM system [41] based on Alize speaker verification software [92]. Note that the front-end processing of this last generative system is a bit different from the one we use for SVM systems, since the optimal settings for the two types of systems are different. The cepstral features extracted from the speech signal are also MFCC but are normalized differently : instead of the feature warping, we use a mean/variance normalization on the sequence (so that each vector component to has a zero mean unit variance on the sequence). Concerning technical details for the Alize GMM system, tuned for NIST evaluation, we used 2048 components in the GMMs with variance flooring during training, as well as 10-best scoring with T-Norm.



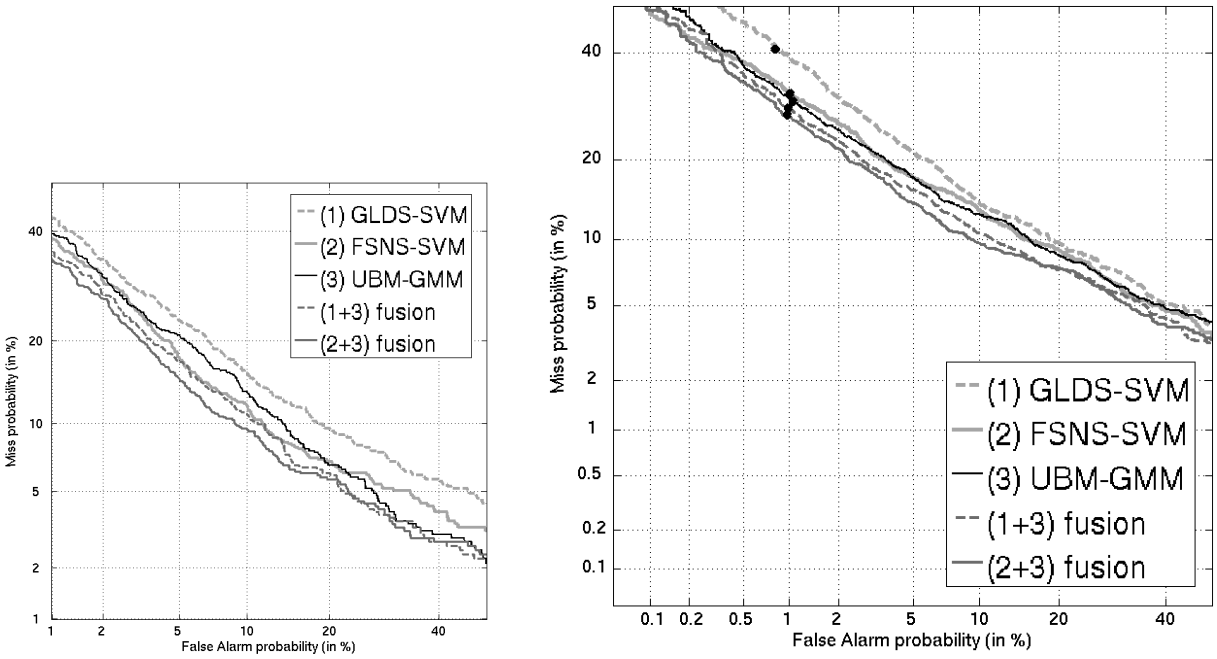
	EER(%)	DCF ($\times 10^{-3}$) min	DCF ($\times 10^{-3}$) actual
(1) GLDS-SVM	12.54	48.5	48.8
(2) FSNS-SVM	**11.91**	41.4	41.6
(3) UBM-GMM	12.06	40.5	**40.6**

FIGURE 3.2 – Performance of SVM systems and a UBM-GMM system on the evaluation set.

Individual results exhibited on Fig.3.2 show that the new approach significantly outperforms the baseline GLDS approach. Besides, SVM systems are competitive with respect

to conventional GMMs while they do not exploit probabilistic models.

Fig.3.3 shows the gain in performance achieved when fusing the GMM system and each previously mentioned SVM system. The fusion we used is a linear combination of output scores and the weight parameters are set so as to minimize the DCF on the validation set. Comparable improvement with such a simple fusion has been previously observed in speaker verification [48, 93]. It shows that a discriminative classifier can bring complementary information to a generative classifier. Fusing both SVM systems do not allow to reach better performance. Note finally that although the FSNS approach significantly



(a) DET curves – validation set

(b) DET curves – evaluation set

	Validation set		Evaluation set		
	EER(%)	min DCF ($\times 10^{-3}$)	EER(%)	DCF ($\times 10^{-3}$) min	DCF ($\times 10^{-3}$) actual
(1) GLDS-SVM	12.80	51.9	12.54	48.5	48.8
(2) FSNS-SVM	**10.55**	**47.5**	**11.91**	41.4	41.6
(3) UBM-GMM	11.48	49.1	12.06	40.5	**40.6**
(1+3) fusion	10.32	45.0	10.54	38.3	38.8
(2+3) fusion	**9.50**	**43.5**	**9.71**	36.7	**37.0**

FIGURE 3.3 – Effect of linear fusion between UBM-GMM and SVM systems.

outperforms the GLDS system individually, the performance improvement when fusing with the GMM system is less glaring.

Chapitre 4

Nonlinear speech processing

Sommaire

4.1 The production mechanism of the speech signal 93

4.2 Non-linear character of the speech signal 95

4.3 Speech as a realization of a non-linear dynamical system 97

4.4 The Microcanonical Multiscale Formalism 100

4.4.1 Singularity exponents 101

4.4.2 The Most Singular Manifold 103

4.4.3 Estimation of singularity exponents 104

4.5 Application to glottal closure instants detection 106

4.5.1 The significant excitation of the vocal tract 106

4.5.2 Review of existing methods 107

4.5.3 The relationship between MSM and GCIs 111

4.5.4 MSM-based GCI detection 113

4.5.5 Experimental results 116

4.6 Application to sparse linear prediction 125

4.6.1 Sparse linear prediction 126

4.6.2 Approximation of the l_0-norm 127

4.6.3 The weighted l_2-norm solution 130

4.6.4 Experimental results 133

This chapter first briefly reviews the evidences regarding the insufficiency of the existing linear techniques in capturing some important dynamics of the speech signal and also the motivations behind searching for alternative non-linear speech analysis methods. In section 4.1, the machinery involved in speech production is briefly recalled. This is then used in section 4.2 to explain the shortcomings of classical linear approaches. Section 4.3 presents our motivations for employing a particular formalism and the broad context of non-linear speech processing where our approach stands. Section 4.4 introduces the he Microcanonical Multiscale Formalism. The last 2 sections present some applications.

4.1 The production mechanism of the speech signal

Speech signal is produced by elaborate intervening of three main groups of organs [94] :

— the lungs, which are the source of energy for the speech signal.

— the larynx, or "voice box", which houses the vocal folds. The slit-like orifice between the two folds is called the glottis. The muscles in the larynx control the width of the glottis and also the amount of tension in the folds as two important parameters in controlling the pitch and the volume of the sounds [95].

— the vocal tract, including the pharynx plus the oral cavity which is coupled to the nasal cavity. The oral cavity may assume many different shapes, with non-uniform cross-sections, depending on the shape of jaws and the configuration of the articulators (the tongue, teeth and lips) during the speech.

Lungs act as the power supply of the vocal system and expel a burst of DC air, which later experiences a perturbation somewhere in the larynx or the vocal tract. The location where this perturbation happens varies depending the type of sound that is being produced [94] :

— for *voiced* sounds, this perturbation occurs in the larynx when a partial closure of

the glottis causes a self-sustained oscillation of the vocal folds. Hence the DC flow of air is converted to quasi periodic glottal pulses.

— for *fricative unvoiced* sounds, passage of the airflow through a quasi-narrow constriction in the oral cavity (generally shaped by the tongue), results in a turbulent flow of air.

— for *plosive unvoiced* sounds, the complete obstruction of the oral cavity (by the lips, tongue or teeth) is followed by a sudden release and hence results in an impulsive release of the air compressed behind the obstruction.

So that, through any one of the above disturbance mechanisms (or a combination of them) the DC airflow is converted to a flow of air with time varying velocity waveform, which later attains its specific phonetic character by traveling in the rapidly time-varying acoustic medium inside the vocal tract. Indeed, the vocal tract may attain a variety of shapes during continuous speech (depending on configuration of the articulators and the shape of its cavities).

This production mechanism is the motivation for the famous source-filter model : a raw excitation source is generated by any one of the above disturbance mechanisms and then, the resulting waveform is colored depending on the shape of the vocal tract system that it experiences [94]. Classically, the effect of the vocal tract system on the excitation waveform is approximated as a linear filter.The excitation source on the other hand, is idealized as periodic puffs for voiced sounds, white noise for fricatives and isolated impulses for plosive sounds. To account for time-varying character of excitation source and the vocal tract, the analysis is performed in windows of 10-30 ms length, during which the corresponding characteristics are assumed invariant.

This linear approximation is the basis for most of the current state-of-the-art speech analysis techniques. An example is the widespread Linear Prediction Analysis (LPA) technique, which is widely used in almost every field of Speech Processing. LPA is based upon consideration of the vocal tract system as an all-pole linear filter. This in turn suggests an Auto Regressive (AR) predictive formulation for speech samples. The LPA thus serves for

estimation of parameters of this AR model (i.e. the all-pole filter) representing the spectral shape of vocal tract filter and consequently the estimation of the excitation source would be possible by inverse filtering. The independent functioning of the excitation source and the vocal tract in linear source-filter model is an underlying assumption of many other conventional techniques, even when they involve non-linear mathematical manipulations (as it is the case for homomorphic [94] speech processing).

4.2 Non-linear character of the speech signal

There are many experimental evidences as well as theoretical reasonings regarding the existence of non-linear effects in the production mechanism of speech, which are generally ignored in the main-stream linear approaches. These nonlinear aspects are evidenced in the production process of all different types of sounds produced by human vocal system. They are even acknowledged in classical speech processing literature, just before settling down to their linear approximations.

A typical example of such non-linear phenomena is the existence of turbulent sound source in production process of unvoiced sounds [96, 97]. During the production of these sounds, vocal folds are more tensed and are closer to each other, thus allowing for turbulence to be generated ([94], pp. 64). This turbulence that is produced in the vocal folds is called aspiration and is the source of excitation for whispered and breathy voices. Although in linear techniques the turbulent sound source is accounted for by considering the excitation source as white noise, they still fail to fully characterize the sound as their underlying assumption is to consider the flow as a laminar one [98]. Also for the plosives, whose excitation source is idealized as an impulsive source in linear framework, there exist a time spread and turbulent component in practice ([94], pp. 88). The general existence of small or large degrees of turbulence in speech signal is discussed in [99], using the Navier-Stokes equation of fluid motion.

In case of voiced sounds, [100] reports some evidences regarding their characterization by highly complex airflows like jets and vortices in the vocal tract system. Moreover,

it is known that during voicing the vibrations of vocal folds are not exactly periodic as it is idealized in the linear model [101] and their shape changes depending on the amplitude [97]. This prohibits the exact estimation of glottal waveform using the linear inverse filtering approach (while this waveform has an essential role in voicing quality for synthesis applications [102]). These non-linear effects in vocal fold oscillations during voicing are attributed to non-linear feedback via Bernoulli forces [97].

Another major assumption made in linear framework is independent functioning of the vocal tract and the excitation source [102, 98, 97, 101]. When glottis is open there exist a coupling between the two which results in significant changes in formants compared to the closed phase [103]. If there were no interactions, the flow would be proportional to the glottal opening area. But the existence of source-filter interactions, causes the final glottal flow during voicing to be skewed towards right : it slowly increases to a maximum and suddenly ends [104].

The appropriateness of linear systems assumptions [105] are tested in [106, 107] using surrogate data techniques. It is empirically shown and argued (using biomechanical information on speech production) that the mathematical assumptions of LTI system theory cannot represent all the dynamics of the speech. It is then shown that a stochastic non-linear non-Gaussian model for speech signals offer a simplified but more general mathematical framework for speech (in the sense that more phenomena can be explained with fewer assumptions).

We can thus fairly conclude that the existence of non-linear phenomena is theoretically and experimentally established and so we can accept the fact that the production mechanism of the speech signal is not entirely linear. This has motivated a trend toward exploration of the possibilities to catch this non-linear character aiming both at improvement of conventional linear techniques and also at improvement of scientific understanding of this complex phenomena.

4.3 Speech as a realization of a non-linear dynamical system

There exists a broad spectrum of non-linear methods for speech analysis. Several reviews are provided on this subject in [98, 97, 106, 101] and also some collections of non-linear algorithms are presented in [108, 109, 110, 111]. In this work we were interested in an important class among these methods which considers speech as a realization of a non-linear dynamical system and attempts to use the available tools and methods in the study of such systems to catch non-linear features of this signal. Such attempts are motivated by the theoretic and experimental evidences regarding the association of the non-linear aspects in speech production to the turbulent nature of airflow in the vocal tract [96, 99, 100, 112], which in turn justifies the use of methods from the study of chaotic, turbulent systems for speech analysis [98].

In [99], motivated by the considerations on the dynamics of the airflow, it is conjectured that short-term speech sounds contain various degrees of turbulence at time scales below phoneme time scale. The Multi Fractal Dimension (MFD) is thus conceptually equated to the degree of turbulence in a speech sound. Consequently, short-time MFD is shown to be a useful feature for speech-sound classification, segmentation and recognition. Moreover, it is shown that unvoiced fricatives, affricates, stops and some voiced fricatives have a high fractal dimension at all scales, which is consistent with the presence of turbulence phenomena. The MFD value for some voiced fricatives and also vowels is medium-to-high at different scales, which corresponds to turbulence.

It is assumed in [113] that speech production can be regarded as a low-dimensional nonlinear dynamical system $Z(n+1) = H(Z(n))$ and speech signal $s(n)$ as a 1-D time-series resulted from the application of a vector function to the original D_e dimensional dynamic variable $Z(n)$. According to embedding theorem [114], under general assumptions, the lag vector $\mathbf{X(n)} = [s(n), s(n - T_d, \cdots, s(n - (D_e - 1)T_d)]$ has many common aspects with the original phase space of the original (but unknown) phase space $Z(n)$. Hence, several measurements such as correlation dimension and general dimension are tried on the attractor constructed by $\mathbf{X(n)}$ as examples of invariant quantities. These raw

measurements are then used as descriptive feature sets whose values (or their statistical trends) are shown to be [on average] dependent on general characteristics of sounds such as voicing, the manner and place of articulation of broad phoneme classes.

Another characteristic of a dynamical system that might be conserved by the embedding procedure are Lyapanov Exponents (LE). These quantities are often considered as quantitative fingerprints of chaos as they are related to the concept of predictability in a dynamical system. LEs are computed in [115] for isolated phonemes and they are shown to be useful features for phoneme classification.

LEs and correlation dimension are also shown to be valuable tools for voice disorder detection applications. It is discussed in [112] that the classical linear methods can not characterize the whole typography of disordered voice sounds and cannot be used to quantify two main symptoms of voice disorders which do not satisfy near periodicity requirement. Hence, the non-linear dynamical systems theory has been considered in the literature as a candidate for a unified mathematical framework modeling the dynamics seen in all types of disordered vowels. Authors in [112] provide a review on successful application of quantities like LEs and correlation dimension for classification of disordered sounds and also point out some practical limitation in computation of these quantities (for instance, correlation dimension is a quantity which is sensitive to the variance of the signal [116]). They mention that the deterministic non-linear methods are not appropriate for characterization of signals of random and very noisy nature and hence, they introduce an extended non-linear framework covering both deterministic and stochastic components of the speech signal. Recurrence and scaling analysis methods are thus employed using two measures : recurrence period density entropy [114] (measuring the extent of aperiodicity) and detrended fluctuation analysis [117] (which characterizes the self-similarity of the graph of the outcome of a stochastic process, and is used for characterizing the increased breath noise in disordered voices). It is shown that such combination can distinguish healthy from disordered voices with high accuracy [118] with the added benefit of reliability due to reduction of adjustable algorithmic parameters.

In fact, the above experiments altogether reveal the potential of methods related to

the characterization of turbulence and the study of chaotic dynamical system in speech analysis. Also, apart from their applicative relevance, they may bring up some interesting indications about the dynamics of the speech signal, if the limitations in their precise estimation can be neglected. The resulted positive value of LEs in [115] for most of the phonemes, which corresponds to chaotic behavior, at least implies the importance of predictability issues in speech analysis. Moreover, the above than 1 value of MFD in most of the cases in [99], at least confirms the existence of meaningful scale-dependent quantities. However, there exist some practical issues which put some limitations on these methods.

For instance, the determination of the appropriate dimension for the phase space and the appropriate time delay for the construction of the phase-space in embedding procedure, which are important factors in proper estimation of dynamically invariant parameters. In case of LEs, their computation is non-trivial for experimental data and particularly for the speech signal, in which the stationary assumption necessitates the procedures to be performed on phone-level which corresponds to very short data lengths. This is the reason why the authors in [119], have used sustained vowels for the computation of local LE (rather than naturally uttered phonemes). In fact, the dynamical models they have used require longer data samples compared to the actual length of ordinary vowels. This problem is addressed in [115], where several dynamical models in phase-space are extensively compared regarding their fidelity in estimating the LEs of a known dynamical system while a very short amount of data is available and finally TSK model [120] is chosen for the computation of LEs. Also, recent works insist on the limitations brought by classical LEs w.r.t. predictability : a Lyapunov exponent is a *global* quantity measuring an average divergence rate. In the general case, there are some fluctuations in finite time, playing an important role in predictability, which lead to the consideration of large deviations [121]. In case of deterministic chaos, Finite Time LE (FTLE) [1]exhibit multi-scale behavior and is related to large deviations in finite time intervals [121]. But FTLE and classical LE only

1. The FTLE is a scalar value that quantifies the amount of deviation between two particles flowing in a fluid, over a given time interval [122].

coincide in the limit ($t \to \infty$) and their precise numerical computation can be difficult.

From these considerations it appears that the computation of key quantities related to non-linearity in speech is worth contemplating. However, the use of classical embedding techniques are computationally challenging in the case of fully developed turbulence ; and they usually provide global descriptions, which is suitable for classification applications, but not for geometric local analysis.

As a consequence, the emergence of new computational approaches for accessing quantitatively and robustly to local scaling exponents around each point using specific measures of predictability opens vast areas of research for understanding the geometric multi-scale implications of a complex signal such as speech, that is to say, the geometrical interplay between statistical information content and the multi-scale organisations, predicted by them. The work we have developed and which is presented in the following was new in the sense that it focused on the implications contained on accessing localized scaling exponents in the speech signal which can be related to a geometric concept of predictability in the framework of reconstructible systems. To achieve this agenda, I chose the framework of the microcanonical multiscale formalism presented in the next section.

4.4 The Microcanonical Multiscale Formalism

Our work on nonlinear speech analysis was based on a novel formalism (back then) called the Microcanonical Multiscale Formalism (MMF). This formalism has its roots in the study of disordered systems in statistical physics and is related to a precise [quantitative] study of the notion of transition inside a complex system or signal. Statistical physics shows that complexity in a system is intrinsically related to the existence of a hierarchy of multi-scale structures inside the system. A typical example of such multi-scale organization is related to the cascade of energy in the case of fully developed turbulence. The fingerprints of these multi-scale structures are observed in a wide range of natural signals acquired from different complex systems.

In this context, the MMF is an extension of its standard canonical counter-

part [123, 124] which provides *global* views upon such complex structures. The particularity of MMF is that it is based on *geometrical* and *local* parameters rather than relying on *global* quantities. Therefore, MMF makes it possible to locally study the dynamics of complex signals from a multi-scale perspective. Meanwhile, rigid mathematical links are made between this geometric analysis of complexity in the MMF and the global statistical view in the canonical framework. So that, MMF provides tools and methods for both geometric and global description of non-linear phenomena in complex signals characterizing their intermittent signature. In other words, it allows the study of local geometrico-statistical properties of complex signals from a multi-scale perspective.

In practice, the MMF has been shown to be a valuable approach to model and analyze this multi-scale hierarchy in empirical complex and turbulent systems having corresponding statistical properties at different scales and it has shown outstanding results in a wide range of applications from diverse scientific disciplines [125, 126, 127, 128, 129, 130, 131]. An originality of our work was to identify the potential of this framework in speech analysis.

4.4.1 Singularity exponents

The microcanonical framework (the MMF) provides computationally efficient tools for *geometrical* characterization of this inter-scale relationship. MMF provides access to local scaling parameters which provide valuable information about the local dynamics of a complex signal and can be used for precise detection of critical events inside the signal. As such, the micro-canonical formalism not only recognizes the global existence of complex multi-scale structures, but also it shows *locally* where the complexity appears and how it organizes itself.

MMF does not rely on statistical values for ensemble averages, but rather look at what is going on around any given point. It is based on the computation of a scaling exponent $h(t)$ at every point in a signal domain and out of any stationarity assumption. These exponents are formally defined by the evaluation of the limiting behavior of a multi-scale

functional $\Gamma_r(s(t))$ at each point t over a set of fine scales r :

$$\Gamma_r(s(t)) = \alpha(t)\, r^{h(t)} + \mathrm{o}\left(r^{h(t)}\right) \qquad r \to 0 \tag{4.1}$$

where $\Gamma_r(s(t))$ can be any multi-scale functional complying with this power-law and the multiplicative factor $\alpha(t)$ generally depends on the chosen Γ_r, but for signals conforming to multi-scale hierarchy, the exponent $h(t)$ is independent of it. The term $\mathrm{o}\left(r^{h(t)}\right)$ means that for very small scales the additive terms are negligible compared to the factor and thus $h(t)$ dominantly quantifies the multi-scale behavior of the signal at the time instant t. Indeed, close to a critical point, the details on the microscopic dynamics of the system disappear and the macroscopic characteristics are purely determined by the value of this exponent, called the Singularity Exponent (SE) [132]. A central concern in MMF is the proper choice of the multi-scale functional $\Gamma_r(s(t))$ so as to precisely estimate these exponents. We will address this subject in following sections, but for now let us assume the availability of precise estimates of $h(t)$ and develop the link between this geometric representation and the global one in the canonical formalism.

When correctly defined and estimated, the values of singularity exponents $h(t)$ define a hierarchy of sets having a multi-scale structure closely related to the cascading properties of some random variables associated to the macroscopic description of the system under study, similar to the one observed in the canonical framework. Formally, this hierarchy can be represented by the definition of singularity components F_h as the level-sets of the SEs :

$$F_h = \{t \mid h(t) = h\} \tag{4.2}$$

These level sets, each highlight a set of irregularly spaced points having the same SE values. Consequently, they can be used to decompose the signal into a hierarchy of subsets, the "multi-scale hierarchy". Particularly, they can be used to detect the most informative subset of points called the Most Singular Manifold (MSM) and also, they can be used

to provide a global statistical view of the complex system by the use of the so-called singularity spectrum.

4.4.2 The Most Singular Manifold

In the MMF, a particular set of interest is the level set comprising the points having the smallest SE values and provides indications in the acquired signal about the most critical transitions of the associated dynamics [133]. These are the points where sharp and sudden local variations take place and hence, they have the lowest predictability : the degree in which they can be predicted from their neighboring samples is minimal. MSM is formed as the collection of points having the smallest values of SE. In other words, the smaller the h(t) is for a given point, the higher the predictability is in the neighborhood of this point. It has been established that the critical transitions of the system occurs at these points. This property has been successfully used in several applications [128, 129, 130]. The formal definition of MSM reads :

$$\mathcal{F}_\infty = \{t \mid h(t) = h_\infty\}, \qquad h_\infty = min(h(t)) \tag{4.3}$$

In practice, once the signal is discretized, h_∞ should be defined within a certain quantization range and hence MSM is formed as a set of points where $h(t)$ is below a certain threshold.

The significance of the MSM is particularly demonstrated in the framework of reconstructible systems : it has been shown that, for many natural signals, the whole signal can be reconstructed using only the information carried by the MSM [128, 133]. For example, a reconstruction operator is defined for natural images in [128] which allows very accurate reconstruction of the whole image when applied to its gradient information over the MSM. The reconstruction quality can be further improved, using the Γ_r measure defined in [134] which makes a local evaluation of the reconstruction operator to geometrically quantify the unpredictability of each point.

Although simple, the notion of MSM played an important role in most of the applications we have developed in this work. By those applications, we showed how the MSM corresponds to the subset of physically important points in the speech signal.

4.4.3 Estimation of singularity exponents

As mentioned earlier, the MMF provides numerically stable methods for estimation of the SEs at each point in the signal domain. It consists of methods which are appropriate for empirical data as they filter all the common artifacts that could arise due to discretization, aliasing, noise, lack of stationarity, correlations, instabilities, and other problems related to the nature of real signals or to the numerical analysis of them.

Several approaches are possible in evaluating the power-law scaling of Eq. (4.1), which are theoretically expected to provide the same estimates. However, they each have their own merits and drawbacks in the way they cope with different real-world situations. These approaches may differ either in the employed multi-scale functional $\Gamma_r(\cdot)$, or in the way the multi-scale behavior is being assessed.

4.4.3.1 The choice of $\Gamma_r(\cdot)$

One important factor in precise computation of SEs is the choice of the scale-dependent functional operating on the signal. In a purely turbulent signal, with no more regular dynamics superimposed, different multi-scale functionals should lead to the same values of SEs [133]. But for practical physical processes, different dynamics might be added to the purely turbulent ones and hence, different strategies might be adopted in the choice of this functional. Depending on the application, many functionals have been used : linear increments, gradient-modulus measure, wavelet transform and other. In our work, we ended up defining and using the two-sides variation measure. This functional is based on two-sided measurement of multi-scale variations of the signal $s(t)$:

$$\mathcal{D}_\tau s(t) \;=\; |2s(t) - s(t-\tau) - s(t+\tau)| \tag{4.4}$$

then, the variations are summed up to form the final multi-scale functional :

$$\Gamma_r(s(t)) = \int_0^r |\mathcal{D}_\tau s(t)| d\tau \tag{4.5}$$

4.4.3.2 Estimation of $h(t)$

Once the proper multi-scale functional is chosen, the singularity exponent $h(t)$ can be estimated by the evaluation of Eq. (4.1) over a set of reasonably fine scales. Several approaches for such multi-scale evaluation have been proposed in the literature, such as punctual estimation, log-log regression and inter-scale modulation. We proposed an other approach for multi-scale evaluation of $\Gamma_r(\cdot)$ based on the concept of microcanonical cascades introduced in [135], which associates the power-law scaling of Eq. (4.1) to the existence of an underlying microcanonical cascade process. The latter refers to a cascading process which rather than being valid for distributions, is valid for any given point in the signal. This is being made possible by assuming the existence of inter-scale correlations in form of Eq. (4.1). In such cascade processes, energy or information is transferred between scale levels of the signal. This way, the MSM actually corresponds to the set of points where information concentrates as it transfers across scales and, in that sense, it is a *least predictable/reconstructible manifold*. The cascade variable of this process must follow an infinitely divisible distribution ; a property which permits a simple estimation of the desired scaling exponents, as the sum of a set of *transitional* exponents [135] :

$$h(t) = \frac{1}{k} \sum_{i=1}^{k} h_{r_i}(t) \tag{4.6}$$

where $h_{r_i}(t)$ are the *transitional* exponents, which can be computed by direct evaluation of Eq. (4.1) at each scale, using anyone of the multi-scale functionals in section 4.4.3.1 as :

$$h_{r_i}(t) = \frac{log(\Gamma_{r_i}(s(t)))}{log(r_i f_s)} \tag{4.7}$$

where f_s is the sampling frequency of the signal.

4.5 Application to glottal closure instants detection

This section studies the relationship between the Most Singular Manifold and a particular mechanism in speech production. MSM is indeed shown to be related to the instants of significant excitations of the vocal tract system. This led us to develop an algorithm for automatic detection of these physically interesting instants which are called the Glottal Closure Instants (GCI). We showed that this algorithm has competitive performance to the state-of-the-art methods and effectively outperforms existing methods in the presence of noise. Indeed, as it is based on both time domain and inter-scale smoothing, it provides higher robustness against several types of noises. In the mean time, the high geometrical resolution of singularity exponents prevents the accuracy from being compromised. Moreover, this algorithm extracts GCIs directly from the speech signal and does not rely on any model of the speech signal (such as the autoregressive model in linear predictive analysis). We will see in section 4.6 how this property makes the algorithm suitable for the problem of sparse linear prediction.

This section is organised as follows : section 4.5.1 presents some background information about the production mechanism of the speech signal and the importance of GCIs. Section 4.5.2 briefly reviews some existing methods for GCI detection. Section 4.5.3 presents observations regarding the correspondence of the MSM to the GCIs which result in the development of a noise robust algorithm for GCI detection in section 4.5.4. The experimental results are presented in section 4.5.5.

4.5.1 The significant excitation of the vocal tract

In the classical model of speech production, voiced sound is presented as the output of vocal tract system excited by a quasi-periodic source located in the glottis. According to the aerodynamic theory of voicing, during the production of a voiced sound, a stream

of breath is flowing through the glottis and a push-pull effect is created on the vocal fold tissues that maintains self-sustained oscillation [136]. The push occurs during glottal opening, when the glottis is convergent, whereas the pull occurs during glottal closing, when the glottis is divergent. During glottal closure, the air flow is cut off until breath pressure pushes the folds apart and the flow starts up again, causing the cycles to repeat [136].

In this way, the steady (DC) airflow from the lower respiratory system is converted into a periodic train of flow pulses [137]. The excitation source is hence represented as glottal pulses. However, to a first approximation, the significant excitations of the vocal tract systems (the epochs) can be considered to occur at discrete instants of time (within these pulses) [138]. There can be more than one epoch during a pitch period, but the major excitation usually coincides with the Glottal Closure Instants (the GCIs) [139]. Indeed, when the glottal closing caused the vocal folds to become sufficiently close, the Bernoulli force results in an abrupt closure, which in turn causes an abrupt excitation of vocal tract system [140].

The precise detection of GCIs has found many applications in speech technology. For instance, as the glottal flow is zero immediately after GCIs, the speech signal in this interval represents the force-free response of the vocal tract system [139] and hence, more accurate estimates of vocal tract system can be realized by the analysis of speech signal over this interval (due to the decoupling of the source contribution) [141, 142]. Also, GCIs can be used as pitch marks for pitch synchronous speech processing algorithms, for speech conversion (of pitch and duration) [143], prosody modification [144] and synthesis [145, 146]. GCIs has been also used for speech enhancement in reverberant environments [147], casual-anticasual deconvolution of speech signals [148, 149] and glottal flow estimation [150].

4.5.2 Review of existing methods

The glottal closing (and opening) can be detected accurately and reliably from the contemporaneous Electro-Glotto-Graph (EGG). By a set of skin electrodes on both sides

of the larynx, this device provides a non-invasive measurement of the electrical impedance change caused by vocal fold vibration and hence, it can be used for monitoring the vibratory motion of vocal folds. It is known that the GCIs correspond to a rapid decrease in EGG. Consequently, they can detected as the minimum of the differentiated EGG (dEGG) in each pitch period [151].

However, as contemporaneous EGG is not always available, there has been great interest in GCI detection from the speech signal itself and EGG is usually used just as a reference for performance evaluation. Among these algorithms, many of them are based on detection of large values in the residual signal of Linear Prediction (LP) analysis, which is expected to indicate the GCI locations [152]. However, there are practical issues in epoch extraction from LP residuals as they are vulnerable to noise (and reverberation [147, 153]) and they contain peaks of random polarities [138]. Epoch extraction from LP residuals is extensively studied in [139] and an epoch filtering method is proposed to alleviate the problems in LP based epoch extraction. Still, many of the recent methods use LP residuals for GCI detection as they provide accurate estimates on clean speech [149].

In [138], the impulse like nature of significant excitations is exploited to detect GCIs by confining the analysis around a single frequency. It is argued that an impulsive excitation in the input of vocal tract system, causes discontinuities in the whole frequency range of the output signal. However, the time-varying response of the vocal tract system, makes it difficult to observe these discontinuities directly from the signal. Instead, the effect of these impulses is examined on the output of a narrow band filter centered around a certain frequency. The output of this filter is expected to contain a single central frequency component, while the discontinuities due to impulsive excitations are manifested as deviations from the center frequency. Since the discontinuities due to impulsive excitation are reflected over the whole frequency spectrum (including zero frequency), the authors have opted for zero frequency filtering so as to benefit from the fact that the time-varying characteristics of the vocal tract system is not present at this frequency (as it has resonances at much higher frequencies).

In [154], the properties of the phase spectrum of the speech signal is used for GCI

detection. The term phase spectrum refers to the unwrapped phase function (the group delay) of the short time Fourier transform of the signal. It is known that for a minimum phase signal the average slope of the phase spectrum is zero, while for a shifted minimum phase signal, this average will attain a value proportional to the shift. As the impulse response of a minimum phase system (system whose poles and zeros are located within the unit circle) is a minimum phase signal, the average slope of its phase spectrum would depend on the location of excitation impulse. Moreover, note that the phase spectrum of the LP residuals has a similar property. Indeed, as LP residuals are computed by passing the signal through a minimum phase inverse system, the phase slope characteristics of the excitation will not be altered. So finally, in [154], The negative derivative of the phase spectrum of the LP residuals is used to estimate this average and its positive zero crossings are considered as the locations of major excitations, i.e. the GCIs. The advantage of working on LPs is that it minimizes the effects of the position of the analysis window with respect to the impulse response of the vocal tract system, as LP residuals are a first approximation to the excitation signal. The method has performed well, but only for clean speech.

For the the introduction of DYPSA algorithm in [153], it is mentioned that the choice of analysis window size significantly affects the occurrence of zero crossing in the phase slope function. Ideally, the window should span exactly one impulsive event. When the window is larger, it covers more than one impulsive excitation and hence the zero crossings occur in the mid-way between the two events. The smaller windows causes a raise in false alarm, as spurious zero-crossings would occur in the windows which do not contain any impulsive event. Consequently, the authors use a moderately small window to minimize the risk of missing GCIs. All of these zero-crossings are then taken as candidates (along with additional candidates taken from a procedure called phase slope projection), and then a refined subset is taken as true GCIs by minimizing a cost function using N-best Dynamic Programming. The cost function includes terms considering the spectral quasi stationarity and also the periodic behavior of vocal folds. About the latter, based on the assumption of smooth variations of pitch over short segments, major pitch deviations are

penalized heavily (although the method does not require a supplemental pitch estimator). So in effect, for each pitch period, only one of the candidates is picked, which is the one providing the maximum consistency in terms of pitch-period variation. The same idea of dynamic programming is employed in YAGA to refine candidates which are taken by detection of discontinuities in an estimate of voice source signal [140].

Time-scale representation of the voiced speech is employed in [155, 156] for GCI detection. Lines of Maximum Amplitudes (LoMA) across scales in the wavelet transform domain are shown [157] to be organized in *tree* patterns, with exactly one such tree for each pitch period, while the GCI is located at the top of the strongest branch of it. In [155], an algorithm is described for automatic detection of GCIs using LoMA. First a coarse estimation of the fundamental frequency (F_0) is made so as to define the largest scale containing the F_0. Then, following a dyadic wavelet decomposition, LoMA are built according to a local dynamic programming technique. The pitch information is then used to select the scale containing the first harmonic, and thus selecting one [optimal] LoMA per pitch period. The time position along this LoMA is taken as the GCI. Finally, two heuristics are applied to reduce the errors corresponding to detection of more than one GCI per pitch period. The method is shown to compare favorably with EGG data and DYPSA algorithm as a reference.

As mentioned in [158, 149], a class of method use smoothing or measures of energy for GCI detection. The smoothing attenuates the effect of noise, reverberation and vocal tract resonances while preserving the periodicity of the speech signal. The smoothening can be performed on time, or on multiple scale [140, 159]. The drawback is that the accuracy might be compromised. That is why in SEDREAMS method [160], LP residuals are used in conjunction with a smoothing function so as to benefit from accuracy of the residuals in localizing GCIs. The method consists of two steps : first a mean based signal is computed which has the property of oscillating at the local pitch period. Hence, it can be used to locate short regions in each period as expected intervals of GCI presence. The GCIs are then extracted by detecting discontinuities in LP residuals within the determined interval of presence. In this way, mean based signal provides robustness against noise as it limits

the search space within each pitch period and LP residual provides accuracy. This can be considered as a more reliable way of imposing smooth pitch variation constraint to GCI detection, compared to the dynamic programming techniques.

In our work, we proposed to use the MSM as a simple and computationally efficient criterion for localizing the GCIs and to exploit some properties of the singularity exponents to handle noisy speech in a simple but yet a robust manner.

4.5.3 The relationship between MSM and GCIs

It is shown in [138] that significant impulsive excitations are reflected over the whole speech spectral band. Consequently, excitation impulses would produce *strong* local singularities at different scales of the waveform. This legitimates the use of the multi-scale power-law of Eq. (4.1) to identify and quantify these singularities : it is natural to expect the co-existence of negative transitional SE (Eq. (4.7)) at different scales around these singularities. The summation of these transitional singularities (Eq. (4.6)) would thus result in lower negative values and hence, those points would belong to the MSM (recall that the MSM is defined as the subset of points having the lowest values of SEs).

The relevance of SE values to the instances of significant excitation is illustrated in Fig. 4.1. The top panel shows a part of a voiced sound, while the bottom panel shows the corresponding SE values. The reference GCIs are also shown. It can be seen that $h(t)$ shows a sudden negative peak around GCIs.

Fig.4.2 shows another example which confirms the intuition about the correspondence of the MSM with GCIs. The top panel shows another segment of a voiced sound along with its corresponding pitch marks taken around the GCI points [161]. The bottom panel shows the MSM points of this segment with their value of SE. The MSM is formed as the 5% of samples having lowest value of SE. It can be seen that MSM points are indeed located around the reference GCI. Note also that, around every single GCI, the MSM point with the lowest SE value is the closest one to the GCI mark. This example shows that MSM can indeed identify the locations where significant impulsive excitations occur.

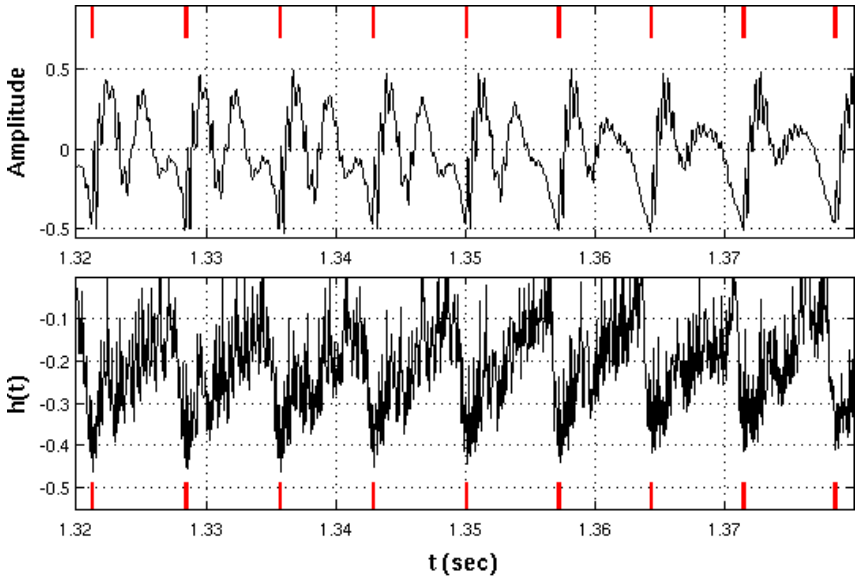


FIGURE 4.1 – **top :** A voiced segment of the speech signal "arctic_a0001" from CMU ARCTIC database [161] and **bottom :** the singularity exponents h(t). The reference pitch marks are represented by vertical red lines.

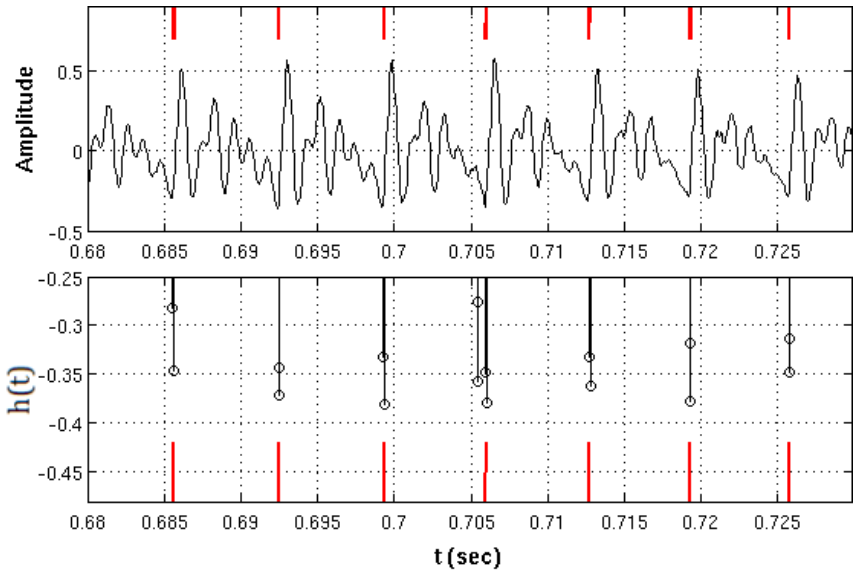


FIGURE 4.2 – **top :** A voiced segment of the speech signal "arctic_a0001" of the male speaker BLD from the "CMU ARCTIC" database [161]. **bottom :** MSM samples and their corresponding SE values. The reference pitch marks are represented by vertical red lines.

We also made a more objective study, by comparing the MSM against the reference GCIs provided by contemporaneous EGG in the KED database which is free available on Festvox website [162]. This comparison confirmed this correspondence. We observed that 95% of the points in the MSM (containing 5% of the points having the lowest values of singularity exponents) coincide with the reference GCI points. However, as the 5% density of MSM is not guaranteed to be equal to density of GCIs, to develop an automatic GCI detection algorithm more care must be taken to cope with false alarms and missing GCIs.

4.5.4 MSM-based GCI detection

The preliminary observations presented in section 4.5.3 showed that MSM effectively coincides with the instants of significant excitations. However, there are some practical considerations in development of an automatic algorithm which must be taken into account. Indeed, practical formation of the MSM requires the specification of a threshold to be applied to singularity exponent values $h(t)$. However, a global specification of the threshold would be impractical : it may happen that a GCI point does attain a lower h(t) value compared to its surrounding points in one pitch period, but in a larger neighborhood, it may have higher value even compared to non-GCI points. This may specially occur for the starting and ending parts of a voiced phoneme, where the energy of the signal is lower compared to the central parts of phonemes. This case is demonstrated in Fig. 4.3 (around the ime instant of 0.77 sec, there are two GCI points which do have the smallest $h(t)$ in smaller neighborhoods, but in a larger window their $h(t)$ is even larger than non-GCI points). Also, the presence of noise may cause the location of the points having the lowest value of singularity exponent to be slightly changed from desired GCI locations.

To overcome these issues, note first that GCIs can be identified using two properties of singularity exponents :

1. In each pitch period, $h(t)$ has the lowest value at the GCI. This is always true for clean speech and hence, the location of *local* minimum in each period, can be taken

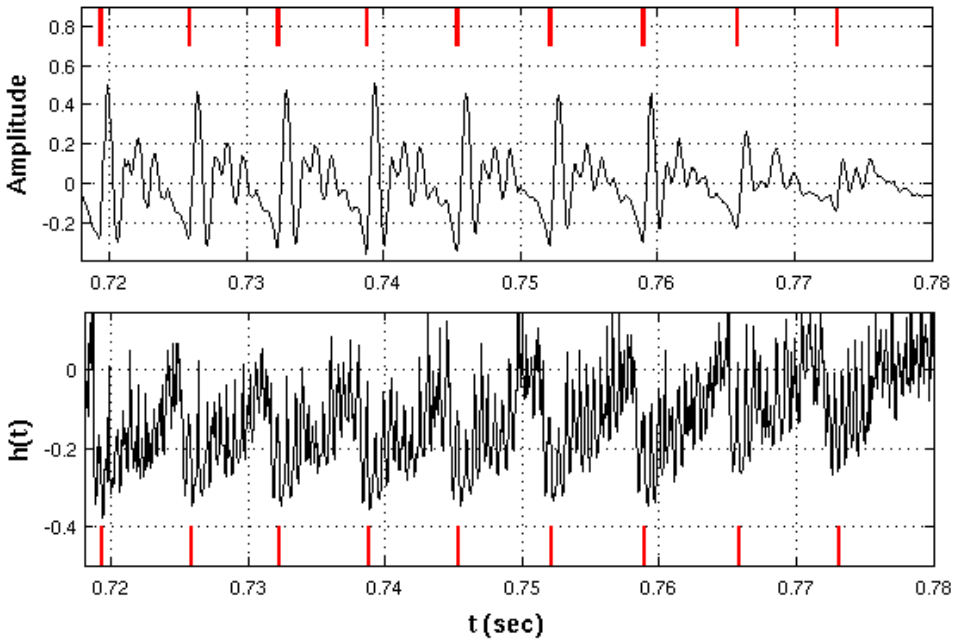


FIGURE 4.3 – **top :** A voiced segment of the speech signal "arctic_a0001" of the male speaker BLD from "CMU ARCTIC" database [161]. **bottom :** MSM samples and their corresponding SE values. The reference pitch marks are represented by vertical red lines.

as the GCI.

2. There is a sharp and clear *level change* before the GCI. Moreover, GCI is the closest point to the instant of level-change. Note that, level change is a relative concept and in our case is independent of the local energy.

Our experiments showed that the first property indeed provided very precise GCI detection in high-energy segments of clean speech, within a single pitch period. However as mentioned before, in a larger window, the GCIs at low-energy parts of speech may attain relatively higher values compared to the non-GCIs belonging to the high-energy parts. Also, the presence of noise may cause a GCI point to attain slightly higher values compared to its immediate neighbors. The second property on the other hand, is a relative quantity itself and local energy have no effect on the level-change. Hence, this criterion seems more suitable for GCI detection in segments with lower energy. Also, the presence of noise would not drastically affect this property. Consequently, we defined a new functional to make explicit and easy use of this level-change. We defined the level-change functional as :

$$\mathcal{L}_c(t) = \sum_{t-T_L}^{t-\delta t} h(t) - \sum_{t}^{t+T_L} h(t) \tag{4.8}$$

where T_L is a parameter controlling the length of averaging window. Fig. 4.4, illustrates the resulting functional for a segment of voiced speech along with the reference GCIs. It can be seen that indeed, the peaks of $\mathcal{L}_c(t)$ corresponds to the GCIs. Moreover, it oscillates with the pitch period. In that sense, this is similar to the mean-based signal used in [160], with the advantage that the GCI is located on its peak in each period.

Of course, as this functional is obtained through a smoothing procedure, its precision in detection of GCIs may not be competitive with $h(t)$. We thus used this $\mathcal{L}_c(t)$ only to limit the search space and we used $h(t)$ itself for the final GCI detection. Indeed, the level-change of $h(t)$ occurs once per pitch period. So, in each period, $\mathcal{L}_c(t)$ experiences a peak at GCI which is proceeded by a positive-going zero-crossing and is followed by

a negative-going zero-cross (the reason for these zero-cross can be easily observed in the definition of $\mathcal{L}_c(t)$ in Eq. (4.8) which includes the difference of averages of $h(t)$ on two windows of length T_L, on both sides of any given time instant). As these zero-crossings can be easily detected without any ambiguity, we used them as guiding lines in our algorithm. The final implementation is provided in Algorithm 2.

Algorithm 1 GCI detection algorithm

1: Calculate $h(t)$ and $\mathcal{L}_c(t)$.
2: In $\mathcal{L}_c(t)$, for any positive-going zero-cross t_p, find the next negative-going zero-cross t_n.
3: $t_{peak} \leftarrow \underset{t}{\text{argmax}}\ \mathcal{L}_c(t), \quad t \in (t_p, t_n)$.
4: MSM formation : take t_1, t_2, t_3 having the lowest values of $h(t)$ in $t \in (t_p, t_n)$.
5: $t_{msm} \leftarrow \underset{t_i}{\text{argmin}}\ |t_i - t_{peak}|$
6: $t_{gci} \leftarrow 0.5 \times (t_{peak} + t_{msm})$

Note that in step 4 of the algorithm, we took 3 points with the lowest value of singularity exponent so as to cope with noisy scenarios where $h(t)$ at GCI may be slightly higher than one or two of immediate neighbors. That is why the criterion of closeness to the peak of $\mathcal{L}_c(t)$ is used in step 5, to make the final decision. Indeed, $\mathcal{L}_c(t)$ is not simply used for constraining the detection to one detection per period, but rather, as its peak is expected to be located on GCI it is also contributing in increase of accuracy.

4.5.5 Experimental results

We tested our algorithm on CMU ARCTIC databases, which consists of 3 sets of 1150 phonetically balanced sentences, each uttered by a single speaker : BDL (US male), JMK (US male) and SLT (US female) [162]. We also tested on the KED Timit database which contains 453 utterances spoken by a US male speaker. All these freely available [162] datasets contain contemporaneous EGG recordings. The reference GCIs are thus taken as the negative peaks of differentiated EGG (manual synchronization of EGG signal and speech recordings are made to compensate for larynx-to-microphone delay). The only parameter to be selected for our algorithm is T_L. The only constraint considerd in selection

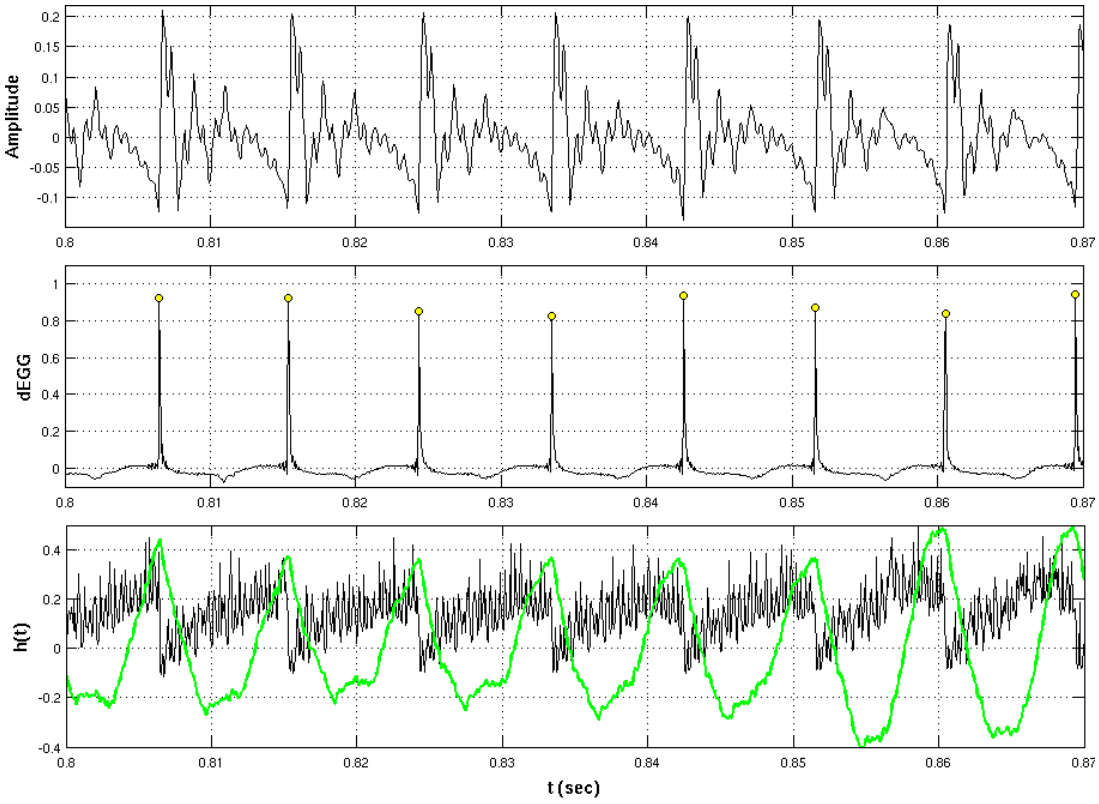


FIGURE 4.4 – **top :** A voiced segment of the speech signal taken from KED database. **middle :** the differenced EGG signal which serves for extraction of reference GCI points. The peaks are marked with yellow circles as the reference GCIs. **bottom :** singularity exponents are shown by black color and the level-change functional $\mathcal{L}_c(t)$ is shown by green color.

of T_L is to keep it smaller than half of the local pitch period for different speakers. On the other hand, the longer it is, the higher the robustness would be against white noise. In the experiments we used $T_L = 1.5ms$.

An extensive comparison is made in [158] between stat-of-the-art GCI detection method which shows that for clean speech, SEDREAMS [160] and YAGA [140] have the best of performances. In the noisy case however, SEDREAMS significantly over-performs all the other methods. Consequently, we compared our method with SEDREAMS [158] using the implementation that is made available online by its author [163]. Note that for each speaker, the EGG signal is synchronized to the speech recordings such that SEDREAMS performance is maximized.

4.5.5.1 Performance measures

We used the set of performance measures defined in [153] to evaluate the performance of our method. If each reference GCI is denoted by t_{gci_k}, the corresponding larynx cycle can be defined as the range of samples $t \in (\frac{t_{gci_k}+t_{gci_{k-1}}}{2}, \frac{t_{gci_k}+t_{gci_{k+1}}}{2})$. Consequently, two sets of performance measures are defined using the graphical representation in Fig. 4.5. The first set consists in three measures of the *reliability* of the algorithms :

— Hit Rate (HR) : the percentage of larynx cycles for which exactly one GCI is detected.

— Miss Rate (MR) : the percentage of larynx cycles for which no GCI is detected.

— False Alarm Rate (FAR) : the percentage of larynx cycles for which more than one GCI is detected.

And the second set defines two measures of the *accuracy* of the algorithms :

— Accuracy to ±0.25 ms (A25m) : the percentage of larynx cycles for which exactly one GCI is detected and the identification error ζ is less than±0.25 ms.

— Identification Accuracy (IDA) : the standard deviation of identification error ζ (the timing error between the reference GCIs and the detected GCIs in larynx cycles for which exactly one GCI has been detected).

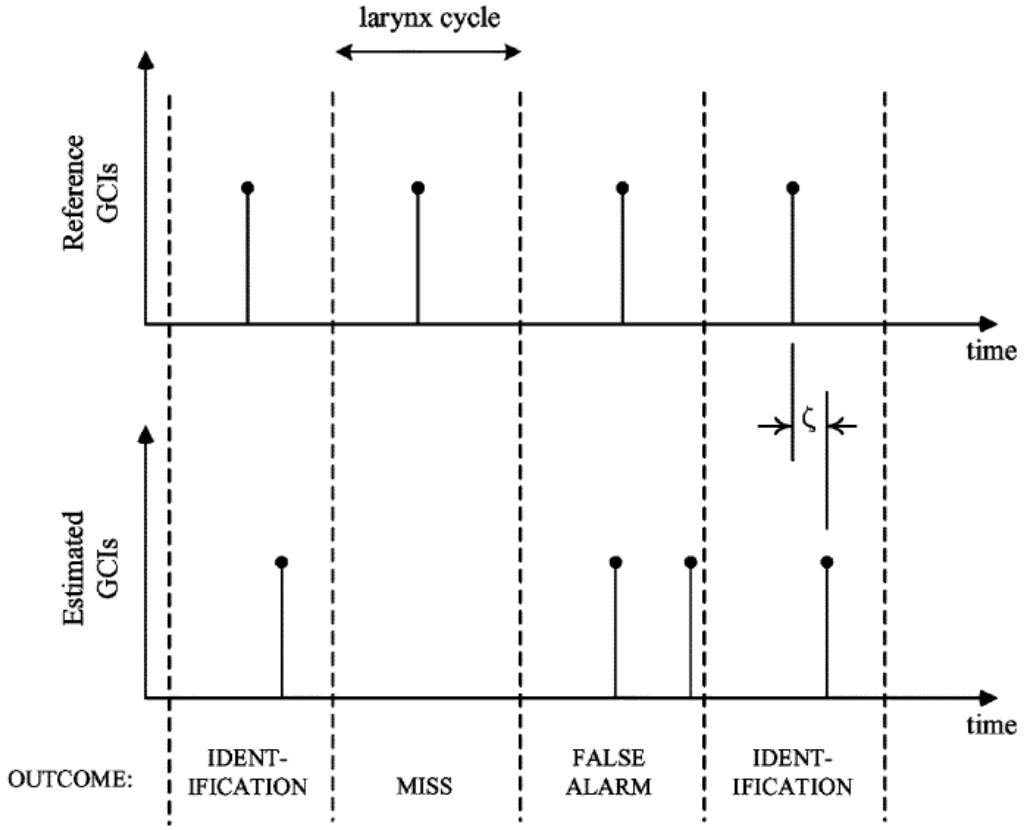


FIGURE 4.5 – Characterization of GCI Estimates showing 4 larynx cycles with examples of each possible outcome from GCI estimation. Identification accuracy is measured by ζ (the graphical representation is taken from [153])

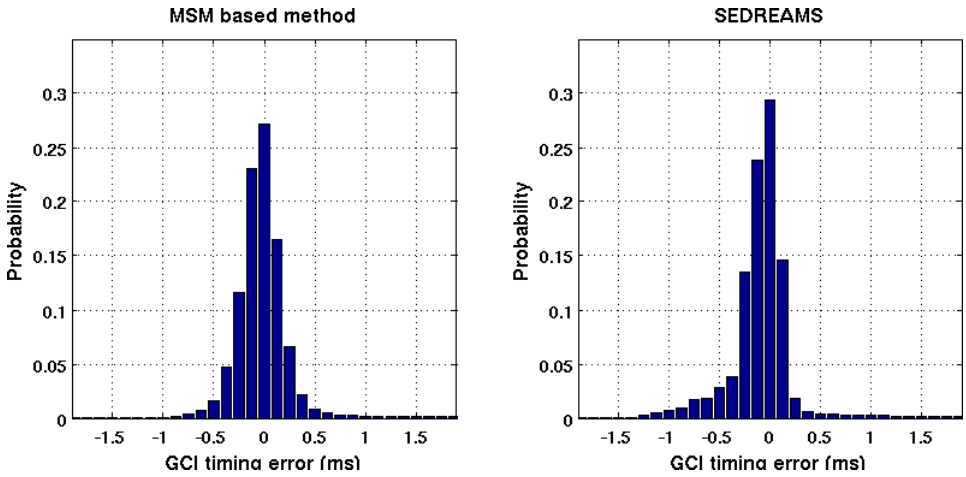


FIGURE 4.6 – Histogram of GCI detection timing error ζ. A reference GCI is considered to be correctly detected when exactly one detection has happened for the corresponding larynx cycle.

In our experiments, the reference GCIs were extracted from the contemporaneous EGG recordings provided in KED database. We used the significant peaks of dEGG signal as the reference GCIs. For this, manual synchronization is done to compensate for the delay between the EGG recording and the speech signals.

4.5.5.2 Clean speech

Table 4.1 compares the performance of different GCI detection method for clean speech signals 4.5.5.1. Overall, it can be seen that SEDREAMS is slightly more reliable, but the accuracy of the two methods are the same. Fig. 4.6 shows histograms of GCI detection timing error for the two algorithms (over the whole four datasets). It can be seen that the distribution of timing error is identification of GCIs are almost the same.

4.5.5.3 Noisy speech

To assess the performance of our algorithm in more realistic scenarios, we evaluated its robustness against 14 different types of noises taken from the NOISEX-92 database [164]. We compared the results of our MSM-based algorithm with that of the SEDREAMS method [158], which was shown in [158, 149] to be the most robust method compared to

TABLE 4.1 – The comparative table of GCI detection performances for clean speech signals.

BDL dataset :					
	HR(%)	MR (%)	FAR (%)	IDA (ms)	A25m (%)
MSM-based	94.7	2.7	2.5	0.54	79.5
SEDREAMS	97.4	0.85	1.7	0.38	85.43
JMK dataset :					
	HR(%)	MR (%)	FAR (%)	IDA (ms)	A25m (%)
MSM-based	94.9	1.38	3.6	0.55	85.5
SEDREAMS	97.8	0.52	1.6	0.53	78.9
SLT dataset :					
	HR(%)	MR (%)	FAR (%)	IDA (ms)	A25m (%)
MSM-based	94.1	4.42	1.4	0.39	80.91
SEDREAMS	98.3	0.02	1.6	0.31	80.25
KED dataset :					
	HR(%)	MR (%)	FAR (%)	IDA (ms)	A25m (%)
MSM-based	97.4	1.07	1.5	0.39	96.24
SEDREAMS	98.8	0.05	1.14	0.34	94.33
Overall results for four speakers :					
	HR(%)	MR (%)	FAR (%)	IDA (ms)	A25m (%)
MSM-based	95.5	2.3	2.2	0.48	82.3
SEDREAMS	98.0	0.4	1.6	0.39	82.5

the other state-of-the-art methods.

Fig 4.7 shows the results in presence of different types of noises. To make the comparison easier, only two performance measures are shown : the Hit Rate (HR) as a measure of reliability and the Accuracy to ± 0.25 ms as a measure of accuracy. It can be seen that in terms of reliability (Hit Rate), SEDREAMS overperforms in cases of white noise, Babble noise and destroyer engine noise. However, the MSM based method is more reliable in presence of car interrior noise, factory floor noise, Leopard military car noise and tank noise. For the remaining 7 types of noises, the reliability of the two methods are quite close, while SEDREAMS shows slightly better results specially for higher SNRs. However, in terms of accuracy, the MSM based methods is showing significantly higher performance for all the 14 types of noises.

SEDREAMS reliability can be explained by the adaptive control of the window length with a rough estimation of pitch period. This permits the algorithms to smoothen the signal as much as possible. That is why SEDREAMS shows much more reliable results in presence of an uncorrelated noise like white noise. The more accurate result of our MSM-based algorithm compared to SEDREAMS might be explained by the difference between the level-change function in our method and the mean-based signal used in SEDREAMS (both of them are used to constrain the number of detections in each pitch period to one). Apparently both of these functionals serve a similar goal to increase the *reliability* of the algorithms. However, the level-change functional $\mathcal{L}_c(t)$ has two distinctive features that contribute not only to improve the *reliability* of our algorithm but also serves to improve its *accuracy* : first, its peak is located on the GCI and hence, it is a smooth (noise-robust) pointer to the GCI (while the mean-based function has no indication about location of GCI). The second difference is that $\mathcal{L}_c(t)$ is a relative quantity which results in its independence from long-range correlations or DC shifts due to changes in energy, or presence of noises like car-noise. It must be always noted however, that the high geometrical resolution of our algorithms is mainly attributed to the geometric resolution of singularity exponents.

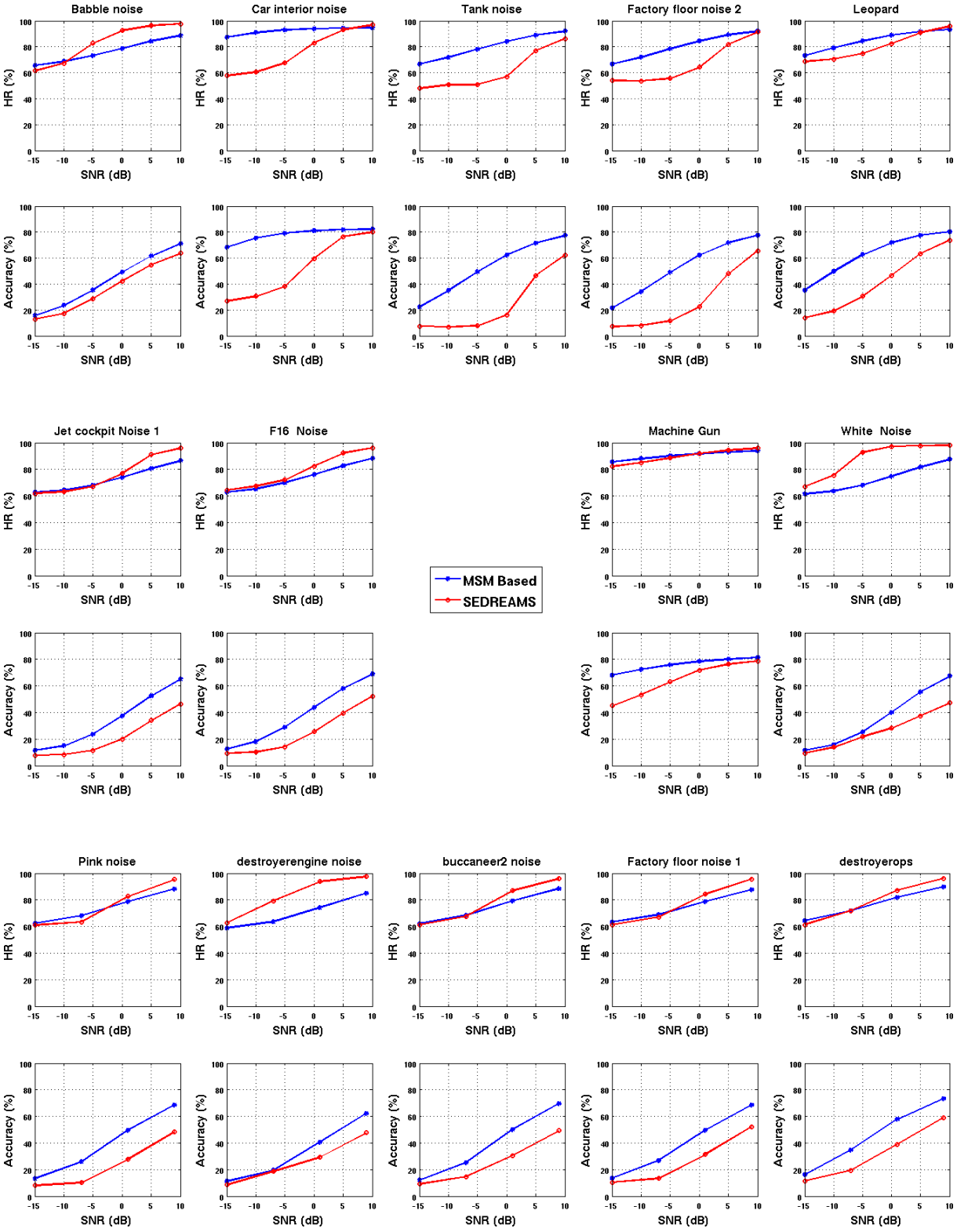


FIGURE 4.7 – Performance comparison in presence of 14 different types of noises taken from NOISEX database [164].

TABLE 4.2 – Comparison between the Relative Computation Time (RCT).

Method	RCT (%)
MSM-based	2.2
SEDREAMS [158]	25.1
fastSEDREAMS [158]	43.8

4.5.5.4 Computational complexity

We compared the computational complexity of our algorithm with that of fastSEDREAMS [158], which is shown to be the most efficient algorithm compared to other state-of-the-art algorithms [158, 149]. As the computational complexity of a GCI detection algorithm is highly data-dependent, it is not easy to provide order of computation details [149]. Instead, we used an empirical metric called Relative Computation Time (RCT) and is defined as [158] :

$$RCT(\%) = 100.\frac{CPU\ time\ (s)}{Sound\ duration\ (s)} \tag{4.9}$$

The RCTs of the MATLAB implementation of the two methods are compared in Table 4.2, where the processing times are averaged over the whole database (Note that RCT is a relative quantity that is dependent on the specific processor that is used for the experiment. It can be seen that our MSM based method is almost 30 times faster than SEDREAMS. Also, compared to the fast implementation of SEDREAMS [149], the MSM based method is 17 times faster. We used the fast implementation that is provided by the author of SEDREAMS [163]. It must be noted that the results of sections 4.5.5.2 and 4.5.5.4 are reported using the original implementation and not the fast one. However, the fast implementation is not always as reliable as the original one, specially in the noisy scenarios.

4.6 Application to sparse linear prediction

This section presents a case study on the application of our GCI detection algorithm in a main-stream speech technology. We address the problem of sparse Linear Prediction (LP) modeling, that is, the estimation of vocal tract model such that the corresponding LP residuals are as sparse as possible : for voiced sounds, one desires the residual to be zero all the time, except for few impulses at GCIs. Classical LP analysis is based on minimization of l_2-norm of residuals, which fails in providing the desired level of sparsity. The standard solutions for this problem are complex as they generally consist in the minimization of the l_1-norm of the linear prediction error through complex convex programming optimization techniques. To bypass this complexity, we introduced a simple closed-form solution based on minimization of weighted l_2-norm of residuals. The weighting function plays the most important role in our solution in maintaining the sparsity of the resulting residuals. To do so, we used our MSM-based GCI detector to extract from the speech signal itself, the points having the potential of attaining largest norms of residuals and then we constructed the weighting function such that the prediction error is relaxed on these points. Consequently, the weighted l_2-norm objective function can be efficiently minimized by the solution of normal equations of the liner least squares problem. The choice of our MSM-based GCI detector is particularly justified, considering the fact that most of successful GCI detection methods use LP residuals for their detection and hence, they cannot be used for constraining the LP problem. This relatively simple algorithm provides better sparseness properties and does not suffer from instabilities. An experiment was carried out to show how such sparse solution leads to more realistic estimates of the vocal tract by decoupling of the contribution of the excitation source from the vocal tract filter one. Moreover, to show a potential application of such sparse representation, we used the resulting linear prediction coefficients inside a multi-pulse synthesizer and showed that the corresponding multi-pulse estimate of the excitation source results in slightly better synthesis quality when compared to the classical technique which uses the traditional non-sparse minimum variance synthesizer.

This section is organized as follows : Ssection 4.6.1 introduces the general problem of sparse linear prediction. Section 4.6.2 recalls the mathematical formulation of the problem and reviews some existing methods. Our novel efficient solution is described in section 4.6.3. In section 4.6.4, the experimental results are presented.

4.6.1 Sparse linear prediction

Linear Prediction analysis is a ubiquitous analysis technique in current speech technology. The basis of LP analysis is the source-filter production model of speech. For voiced sounds in particular, the filter is assumed to be an all-pole linear filter and the source is considered to be a semi-periodic impulse train which is zero most of the times, i.e., the source is a sparse time series. LP analysis results in the estimation of the all-pole filter parameters representing the spectral shape of the vocal tract. The accuracy of this estimation can be evaluated by observing the extent in which the residuals (the prediction error) of the corresponding prediction filter resembles the hypothesized source of excitation [94] (a perfect impulse train in case of voiced speech). However, it is shown in [94] that even when the vocal tract filter follows an actual all-pole model, this criterion of goodness is not fulfilled by the classical minimum variance predictor.

It is argued in [165] that the reason behind the failure of the classical method in providing such sparse representation is that it relies on the minimization of l_2-norm of prediction error. It is known that the l_2-norm criterion is highly sensitive to the outliers [166], i.e., the points having considerably larger norms of error. Hence, l_2-norm error minimization favors solutions with many small non-zero entries rather than the sparse solutions having the fewest possible non-zero entries [166]. Hence, l_2-norm is not an appropriate objective function for the problems where sparseness constraints are incorporated. Indeed, the ideal solution for sparse residual recovery is to directly minimize the cardinality of this vector, i.e. the l_0-norm of prediction error which yields a combinatorial optimization problem. Instead, to alleviate the exaggerative effect of l_2-norm criterion at points with large norms of error, it is usual to consider the minimization of l_1-norm as it puts less emphasis on

outliers. l_1-norm can be regarded as a convex relaxation of the l_0-norm and its minimization problem can be re-casted into a linear program and solved by convex programming techniques [167].

The l_1-norm minimization of residuals is already proven to be beneficial for speech processing [168, 169, 170]. In [168], the stability issue of l_1-norm linear programming is addressed and a method is introduced to achieve intrinsically stable solution as well as keeping the computational cost down. The approach is based the Burg Method for autoregressive parameters estimation using the least absolute forward-backward error.

In [169], the authors have compared the Burg method with their l_1-norm minimization method using the modern interior points method and showed that the sparseness is not preserved with the Burg method. Later, they have proposed a re-weighted l_1-norm minimization approach in [170], to enhance the sparsity of the residuals and to overcome the mismatch between l_0-norm minimization and l_1-norm minimization while keeping the problem solvable with convex programming tools. Initially the l_1-norm minimization problem is solved using the interior points method and then the resulted residuals are used iteratively, to re-weight the l_1-norm objective function such that less weight is given to the points having larger residual norms. The optimization problem is thus iteratively approaching the solution for the ideal l_0-norm objective function.

4.6.2 Approximation of the l_0-norm

The consideration of the vocal tract filter in the source-filter production model as an all-pole filter results in the well-known autoregressive model for the speech signal $x(n)$:

$$x(n) = \sum_{k=1}^{K} a_k x(n-k) + r(n) \tag{4.10}$$

where a_k are the prediction coefficients, K is the order of prediction filter and $r(n)$ is the prediction error or the residual. In the ideal case, when the $\{a_k\}$ coefficients are perfectly estimated and the production mechanism verifies the all-pole assumption, the residual

should resemble the hypothesized excitation source. In case of voiced speech, it should be a perfect semi-periodic impulse train which is zero most of the times, i.e., it is a sparse time series. The linear prediction analysis problem of a frame of length N can be written in the general matrix form as the l_p-norm minimization of the residual vector $\mathbf{r}$:

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmin}} \parallel \mathbf{r} \parallel_p^p, \quad \mathbf{r} = \mathbf{x} - \mathbf{X}\mathbf{a} \tag{4.11}$$

where $\mathbf{a}$ is a vector representing the set $\{a_k\}$ and

$$\mathbf{x} = \begin{bmatrix} x(N_1) \\ \vdots \\ x(N_2) \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x(N_1-1) & \cdots & x(N_1-K) \\ \vdots & & \vdots \\ x(N_2-1) & \cdots & x(N_2-K) \end{bmatrix}, \mathbf{r} = \begin{bmatrix} r(N_1) \\ \vdots \\ r(N_2) \end{bmatrix}$$

and $N_1 = 1$ and $N_2 = N + K$ (For $n < 1$ and $n > N$ we put $x(n) = 0$). The l_p-norm is defined as $\parallel \mathbf{x} \parallel_p = (\sum_{k=1}^{N} x(k)^p)^{\frac{1}{p}}$. Depending on the choice of p in Eq. (4.11), the estimated linear prediction coefficients and the resulting residuals would possess different properties.

The ideal solution to the LP analysis problem of Eq. (4.11), so as to retrieve the sparse excitation source of voiced sounds, is to directly minimize the number of non-zero elements of the residual vector, i.e. its cardinality or the so-called l_0-norm [171]. As this problem is an N-P hard optimization problem [170], its relaxed but more tractable versions ($p = 1, 2$) are the most widely used.

Setting $p = 2$ results in the classical minimum variance LP analysis problem. Although the latter suggests the highest computational efficiency, it is known that this solution cannot provide the desired level of sparsity, even when the vocal tract filter is truly an all-pole filter [94]. It is known that l_2-norm has an exaggerative effect on the points having larger values of prediction error (the so-called outliers). Consequently, the minimizer puts much effort on forcing down the value of these outliers, with the cost of more non-zero elements. Hence, the resulting residuals are not as sparse as desired.

It is known that this exaggerative effect on the outliers is reduced with the use of l_1-norm and hence, its minimization could be a meliorative strategy w.r.t the minimum variance solution, in that the error on the outliers are less penalized [171]. The solution to the l_1-norm minimization is not as easy as the the classical minimum variance LP analysis problem but it can be solved by recasting the minimization problem into a linear program [172] and then using convex optimization tools [167]. However, it is argued in [168] that linear programming l_1-norm minimization, suffers from stability and computational issues and instead, an efficient algorithm is introduced, based on a lattice filter structure in which the reflection coefficients are obtained using a Burg method with l_1 criterion and the robustness of the method is shown to be interesting for voiced sound analysis. However, it is shown in [169] that the l_1-norm Burg algorithms behaves somewhere in between the l_2-norm and the l_1-norm minimization. Instead, the authors have shown that enhanced sparsity level can be achieved using modern interior points method [167] of solving the linear program. They have shown interesting results of such analysis and have argued that the added computational burden is negligible considering the consequent simplifications (granted by such a sparse representation) in applications such as open and closed loop pitch analysis and algebraic excitation search.

An iteratively re-weighted l_1-norm minimization approach is consequently proposed by the same authors in [170] to enhance the sparsity of residuals, while keeping the problem solvable by convex techniques. The algorithm starts by plain l_1-norm minimization and then, iteratively, the resulting residuals are used to re-weight the l_1-norm cost function such that the points having larger residuals (outliers) are less penalized and the points having smaller residuals are penalized heavier. Hence, the optimizer encourages small values to become smaller while augmenting the amplitude of outliers [173].

The enhanced sparsity properties of the re-weighted l_1-norm solution compared to the l_1-norm minimization, and also the better performance of the l_1-norm criterion compared to l_2-norm criterion, can be explained numerically with the help of the graphical representation in Fig. 4.8. There, the numerical effect of different residual values on l_p-norm cost functions is graphically depicted. It can be seen that the penalty on outliers is increasing

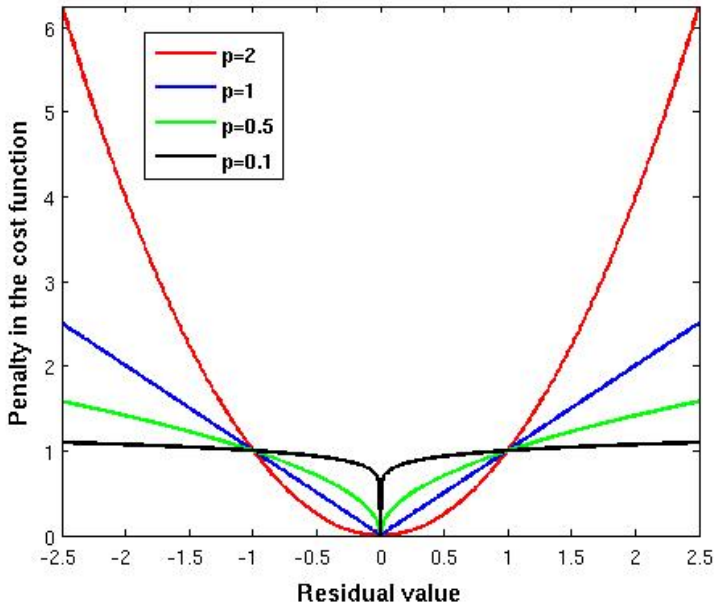


FIGURE 4.8 – Comparison between l_p-norm cost functions for $p \leq 2$. The "democratic" l_0-norm cost function is approached as $p \to 0$. The term "democratic" refers to the fact that l_0-norm weights all the nonzero coefficients equally [171].

with p. Indeed, as $p \to 0$ the penalty of the corresponding cost function on non-zero values approaches l_0-norm cost function (where any non-zero value is equally penalized and there is no penalization of larger values). This will force the minimization to include as many zeros as possible as their weight is zero. In case of the re-weighted l_1-norm solution [170], any residual is weighted by its inverse at each iteration and hence, the equal penalization of any non-zero value (as in l_0-norm criterion) is achieved. In other words, if a point has a very large (resp. very small) residual, it will be less (resp. much more) penalized in the next iteration and so, the sparsity is enhanced iteratively.

4.6.3 The weighted l_2-norm solution

We proposed an alternative and efficient optimization strategy which approximates the desired sparsity of the residuals. Our approach is based on the minimization of a weighted version of the l_2-norm criterion. The weighting function plays the key role in maintaining the sparsity of the residuals. Other than pure numerical motivations on de-emphasizing

the exaggerative effect of l_2-norm on outliers (as discussed in the previous section), the design of this function is physically motivated. Indeed, we extract from signal the points which are susceptible of attaining larger values of residuals, the GCIs, then we construct the weighting function such that the error at those points is less penalized.

4.6.3.1 Optimization algorithm

We opted for l_2-norm cost function to preserve computational efficiency, then we coped with its exaggerative effect on outliers by careful down-weighting of the cost function at those points. Formally, we defined following optimization problem for the recovery of sparse residuals :

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmin}} \quad \sum_{k=1}^{N} w(k)(r(k)^2) \tag{4.12}$$

where $w(\cdot)$ is the weighting function. Once $w(\cdot)$ is properly defined, the solution to Eq. (4.12) is straight-forward. Indeed, setting the derivative of the cost function to zero results in a set of normal equations which can be solved as in the classical l_2-norm approach :

$$\hat{\mathbf{a}} = \mathbf{R}^{-1}\mathbf{p} \tag{4.13}$$

while in our case, $\mathbf{R} = (\mathbf{W} \odot \mathbf{X})\mathbf{X}^T$, $\mathbf{p} = \mathbf{w} \odot (\mathbf{X}^T\mathbf{x})$, $\odot$ denotes the element-wise product of the two matrices and :

$$\mathbf{w} = \begin{bmatrix} w(N_1) \\ \vdots \\ w(N_2) \end{bmatrix}, \mathbf{W} = \begin{bmatrix} w(N_1) & \cdots & w(N_1) \\ \vdots & & \vdots \\ w(N_2) & \cdots & w(N_2) \end{bmatrix}$$

It is interesting to mention that our experiments show that as long as the smoothness of the $w(\cdot)$ is maintained the stability of the solution is preserved. Indeed, the special

form of the input vector $\mathbf{X}$ in Eq. (4.11), is the one used in autocorrelation formulation of LP analysis using l_2-norm minimization. It is proven that autocorrelation formulation always results in a minimum-phase estimate of the all-pole filter, even if the real vocal tract filter is not minimum phase [94]. As our formulation is similar to the autocorrelation formulation, we can fairly expect the same behavior (though we don't have a theoretical proof). This is indeed beneficial, as having a non-minimum phase spectral estimate results in saturations during synthesis applications. Our experiments show that such saturation indeed never happens. This is an interesting advantage of our method compared to l_1-norm minimization methods which do not guaranty a minimum phase solution, unless if additional constraints are imposed to the problem [169].

4.6.3.2 The weighting function

The weighting function is expected to provide the lowest weights at the GCI points and to give equal weights (of one) to the remaining points. To put a smoothly decaying down-weighting around GCI points and to have a controllable region of tolerance around them, a natural choice is to use a Gaussian function. We thus defined the weighting function as :

$$w(n) = 1 - \sum_{k=1}^{N_{gci}} g(n - T_k) \tag{4.14}$$

where $T_k, k = 1 \cdots N_{gci}$ denotes the detected GCI points and $g(\cdot)$ is a Gaussian function ($g(x) = \kappa e^{(\frac{x}{\sigma})^2}$). The parameter σ allows the control of the width of the region of tolerance and κ allows the control of the amount of down-weighting on GCI locations. Fig. 4.9 shows a frame of voiced sound along with the GCI points and the weighting function of Eq. (4.14). It can be seen that this weighting function puts the lowest weights around the GCI locations (i.e. the expected outliers) and equally weights the remaining points. Numerically speaking, the minimizer is free to pick the largest residual values for the outliers and it concentrates on minimizing the error on the remaining points. This can also be explained with regard to physical production mechanism of the speech signal :

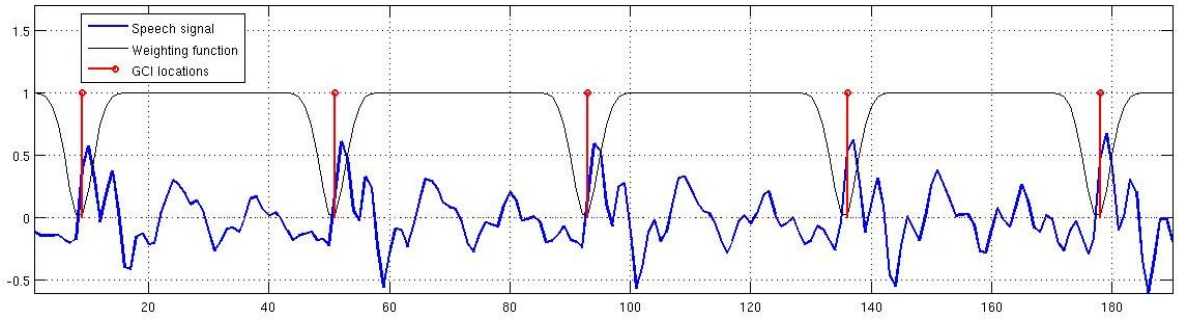


FIGURE 4.9 – A frame of a voiced sound along with the detected GCI locations and the constructed weighting function (with $\sigma = 50$ and $\kappa = 1$).

as the coupling of excitation source and vocal tract filter is maximized on GCIs, such weighting function assists the minimizer to exclude the points on which the coupling is maximized and concentrate its effort on speech samples where the source contribution is minimized. Such decoupling is investigated in the context of Glottal volume velocity estimation by closed phase inverse filtering techniques [140]. There, the whole time interval on which the glottis is expected to be open is localized and discarded from the analysis frame. Consequently, these methods require the availability of both GCI and Glottal Opening Instants (GOI). However, the determination of GOIs is much more difficult than GCI detection [140]. Moreover, as the analysis window is strictly limited to the closed phase [150], another practical issue may arise : this time-frame might be too short (for high-pitched voices for instance) such that the analysis becomes ineffective [140].

4.6.4 Experimental results

We carried out experiments to show : i) the ability of our approach in retrieving sparse residuals for stationary voiced signals ; ii) how it can provide a better estimation of the all-pole vocal-tract filter parameters ; iii) how our sparse modeling can enhance the performance of a multi-pulse excitation estimation. All the results were obtained using the parameter values $w(\cdot)$: $\kappa = 0.9$ and $\sigma = 50$. The choice of these values was obtained using a small development set (few voiced frames) taken from the TIMIT database [174].

TABLE 4.3 – Quantitative comparison of the level of sparsity of different LP analysis strategies.

Method	kurtosis on the whole sentence	kurtosis on voiced parts
l_2-norm	51.7	39.7
l_1-norm	81.9	65.9
weighted-l_2-norm	85.4	69.1

4.6.4.1 Voiced sound analysis

We compared the performance of our weighted-l_2-norm solution with that of the classic l_2-norm minimization and also the l_1-norm minimization via convex programming. For minimization of the l_1-norm, we used the publicly available l_1-magic toolbox [172] which uses the primal-dual interior points optimization [167]. Fig. 4.10 shows the residuals obtained for all these different optimization strategies. It is clear that the weighted-l_2 and also l_1-norm criteria achieve higher level of sparsity compared to the classic l_2-norm criterion. Moreover, a closer look reveals that our weighted-l_2-norm solution shows better sparsity properties compared to the l_1-norm minimization : in the former, each positive peak of residuals is followed by a single negative peak (of almost the same amplitude) while for the latter, any positive peak is surrounded by two negative peaks of smaller (but yet significant) values.

This comparison can be further formalized by using a quantitative measure of sparsity. There exists plenty of such measures on which a review is provided in [175]. We used the kurtosis as it satisfies three of the most important properties that are intuitively expected from a measure of sparsity : scale invariance, rising tide and Robin Hood [175]. Kurtosis is a measure of peakedness of a distribution, higher values of kurtosis implies higher level of sparsity. Table 4.3 shows the kurtosis of the residuals obtained from the three optimization strategies, averaged over 20 randomly selected sentences of both male and female speakers taken from TIMIT database. From this table, it is clear that our method achieves the highest level of sparsity as it obtains the highest values of kurtosis.

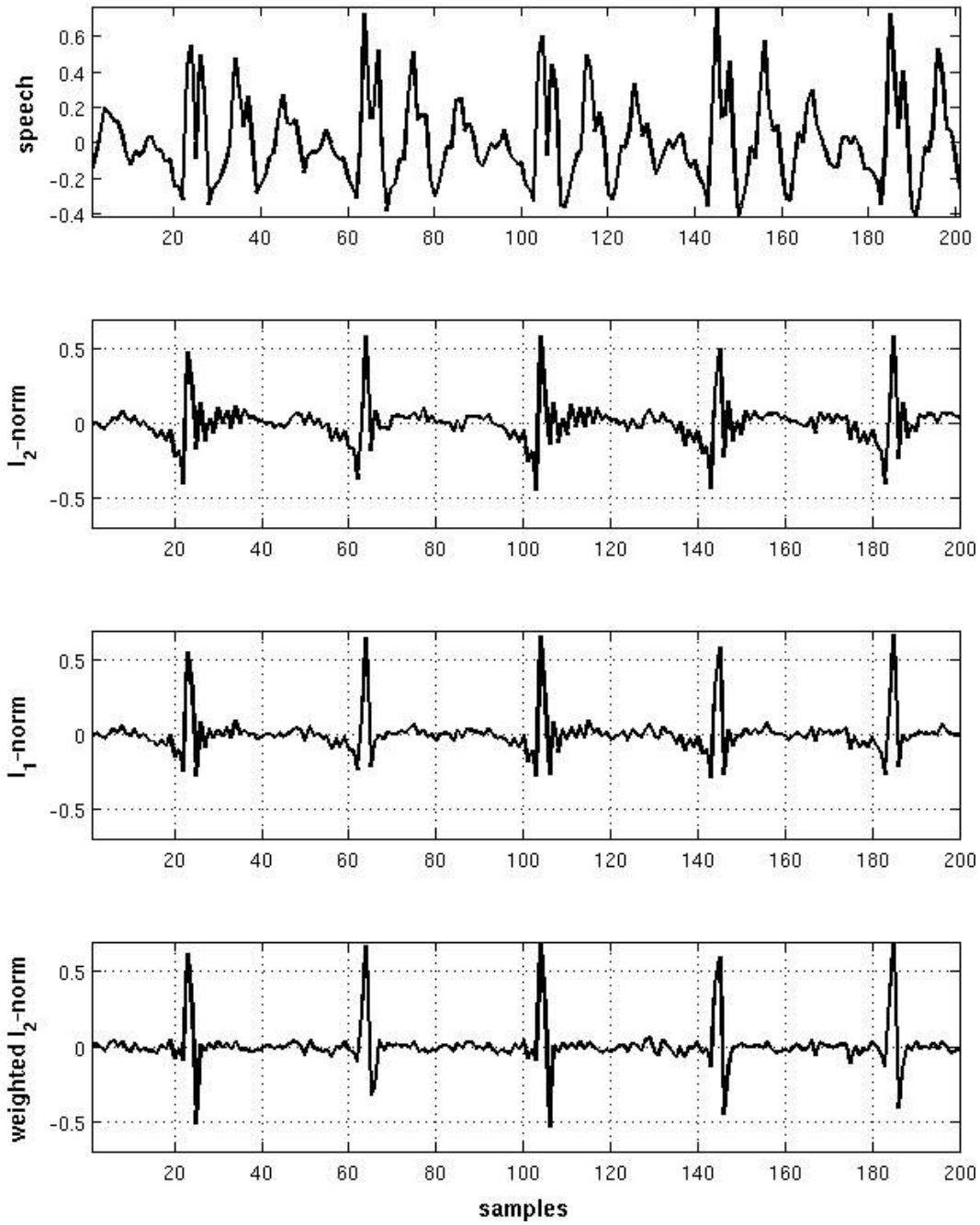


FIGURE 4.10 – The residuals of the LP analysis obtained from different optimization strategies. The prediction order is K = 13 and the frame length is N = 160.

4.6.4.2 Estimation of the all-pole vocal-tract filter

We also investigated the ability of our method in the estimation of the all-pole filter parameters. To do so, we generated a synthetic speech signal by exciting a known all-pole system with a periodic sequence of impulses at known locations. We then estimated these parameters from the synthetic signal by LP analysis using our method and the classical l_2-norm method. Fig. 4.11 shows the frequency response of the resulting estimates along with the frequency-domain representation of the synthetic excitation source. It can be seen that for the l_2-norm minimizer, there is a clear shift in the peaks of the estimated filter towards the harmonics of the excitation source. Specifically, the first spectral peak is shifted toward the forth harmonic of the excitation source. Indeed, the effort of l_2-norm minimizer in reducing large errors (the outliers due to the excitation source) has caused the estimated filter to be influenced by the excitation source. However, our weighted-l_2-norm minimization makes a very well estimation of the original all-pole filter and there is no shift in the spectral peaks. Our method effectively decouples the contributions of the excitation source and the all-pole filter (as the source contribution is de-emphasized by the weighting function).

4.6.4.3 Multi-pulse excitation estimation

The sparseness of the excitation source is a fundamental assumption in the framework of Linear Predictive Coding (LPC) where the speech is synthesized by feeding the estimated all-pole filter by an estimate of the excitation source. The coding gain is achieved by considering a sparse representation of the excitation source. In the popular Multi-Pulse Excitation (MPE) method [176, 177], the synthesis filter is estimated through the classic l_2-norm minimization and then a sparse multi-pulse excitation sequence is extracted through an iterative Analysis-by-Synthesis procedure. However, as discussed in previous sections this synthesizer is not intrinsically a sparse one. Hence, it would be logical to expect that the employment of an intrinsically sparse synthesis filter, such as the one we developed, could enhance the quality of the synthesized speech using the correspon-

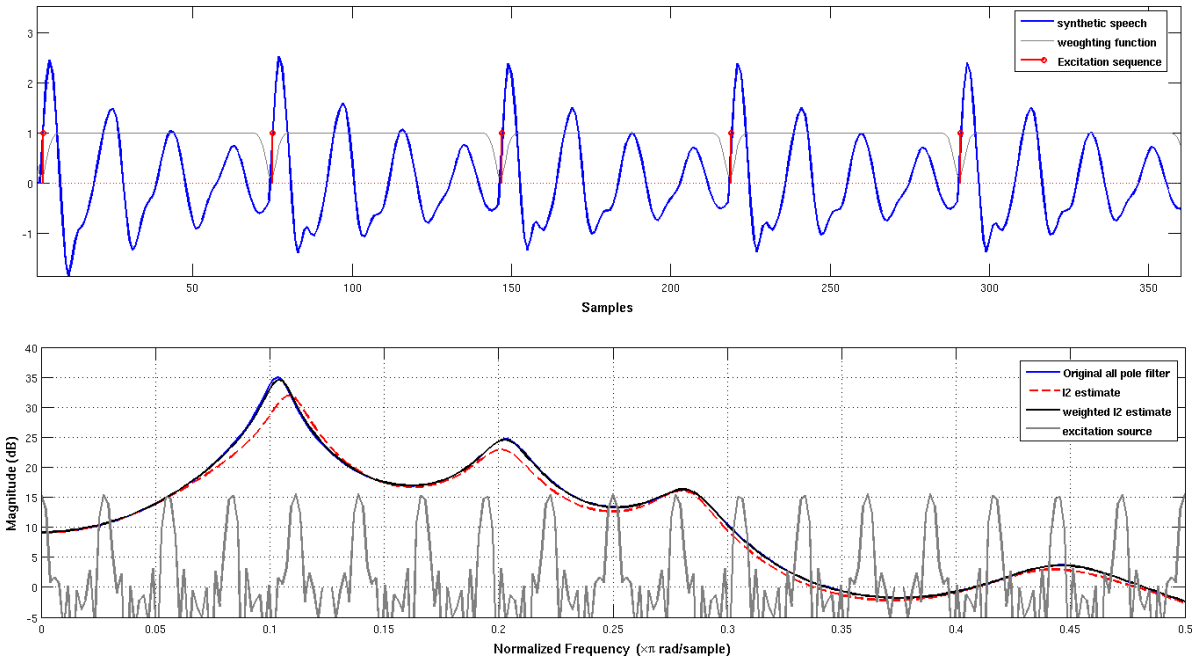


FIGURE 4.11 – **top :** Synthetic speech signal, **bottom :** frequency response of the filters obtained with l_2-norm and weighted-l_2-norm minimization (prediction order $K = 13$). Note that only the first half of the frequency axis is shown so as to enhance the presentation.

ding multi-pulse estimate. We compared the performance of the classical MPE synthesizer which uses minimum variance LPC synthesizer with the one whose synthesizer is obtained trough our weighted l_2-norm minimization procedure. We follow exactly the same procedure for estimation of multipulse coders for both synthesizers, as in the classical MPE implementation in [177] (iterative minimization of perceptually weighted error of reconstruction).

We tried to follow the same experimental protocol as in [178]. That is, we evaluate our method using about 1 hour of clean speech signal randomly chosen from the TIMIT database (re-sampled to 8 kHz) uttered by speakers of different genders, accents and ages which provides enough diversity in the characteristics of the analyzed signals. 13 prediction coefficients are computed for frames of 20ms (N=160) and the search for the multi-pulse sequence (10 pulses per frame) is performed as explained in [177]. We evaluated the quality of reconstructed speech in terms of SNR and the PESQ measure [179] which provides a

TABLE 4.4 – The quality of Multi-pulse excitation coding using two different synthesizer filters. The sparse excitation source is constructed by taking 10 pulses per 20ms.

Method	PESQ	SNR
MPE + l_2-norm	3.3	9.5 dB
MPE + weighted-l_2-norm	3.4	10.2 dB

score of perceptual quality in the range of 1 (the worst quality) to 5 (the best quality). The results are shown in Table 4.4, which shows that our method achieved slightly higher coding quality than the classical MPE synthesizer.

From the computational efficiency perspective, the superior performance of our weighted-l_2-norm solution in retrieving sparse residuals in section 4.6.4.2 and the slight improvement of the coding quality in section 4.6.4.3 were achieved with roughly the same computational complexity as the classical l_2-norm minimization (if the computational burden of the GCI detector is neglegted). This is a great advantage compared to the computationally demanding l_1-norm minimization via convex programming (as in [169] or in [170] where multiple re-weighted l_1-norm problems are solved) which also suffers from instability issues. Moreover, another important feature of our solution is that, during the coding experiment we observed that by using the Gaussian shape for the weighting function, the solution is always stable and it does not encounter the instability issues as l_1-norm minimization.

Chapitre 5

Dysarthric speech processing

Sommaire

5.1 Introduction 140

5.2 Parkinsonism 141

5.2.1 Parkinson's disease 141

5.2.2 Multiple system atrophy 141

5.2.3 Progressive supranuclear palsy 142

5.3 The challenge of differential diagnosis 143

5.4 Dysarthria 143

5.4.1 Hypokinetic dysarthria 145

5.4.2 Spastic dysarthria 147

5.4.3 Ataxic dysarthria 149

5.5 Dysarthria-based differential diagnosis 151

5.6 The Voice4PD-MSA project 154

5.6.1 The consortium 155

5.6.2 Beyond Voice4PD-MSA 157

5.6.3 The challenge of data collection 159

5.7 Differential diagnosis between MSA and PSP 159

5.7.1 Dataset 160

5.7.2 Acoustic features 161
5.7.3 Methodology and experiments 161
5.7.4 Discussion and conclusion 166
5.8 Distortion of voiced obstruents for differential diagnosis between PD and MSA-P 167
5.8.1 Introduction 167
5.8.2 Database 169
5.8.3 Method and results 169
5.8.4 Discussion and conclusion 174

5.1 Introduction

Speech is a unique, complex, dynamic motor activity through which we express thoughts and emotions. Speech production requires the integrity and integration of numerous neurocognitive, neuromotor, neuromuscular and musculoskeletal activities [180]. Deficit in any of this activity may be the cause of speech disorder. Aging is one of the natuaral cause of speech disorder due to different types of physiological degeneration. It is important to note, the elderly population is growing fast all over the world and, as a consequence, the number of elderly subjects with speech/language disorders has also increased rapidly [181]. Presence of voice and speech disorders in adult age is commonly manifested as a consequence of degenerative, traumatic, vascular, infectious, demyelinating or other diseases. Besides these pathophysiology, speech disorder for elderly subjects can be herbinger of different parts of nervous system disorder. Therefore, recognizing and understanding predictable patterns of speech disturbances and their underlying neurophysiological bases are important to understand nervous system organization for speech motor control, differential diagnosis, localization, prevalence and management.

My research has focused on the speech analysis and characterization of a particular group of neurodegenerative diseases, the Parkinsonian disorders or Parkinsonism.

5.2 Parkinsonism

Parkinsonism refers to the group of neurological diseases which includes Parkinson's Disease (PD) and Atypical Parkinsonian Disorders (APD). According to etiology, PD and APD can be divided as sporadic and familial. APD includes rare diseases like Progressive Supranuclear Palsy (PSP), Multiple System Atrophy (MSA), Huntington's disease, Corticobasal Degeneration (CBD), Neurodegeneration with brain iron accumulation (NBIA), Dementia with Lewy Bodies (DLB) and few more. My research has been focused on the analysis of dysarthric speech caused by PD, MSA and PSP.

5.2.1 Parkinson's disease

Parkinson's disease (PD) is the most common neurodegenerative disease after Alzheimer's disease. The prevalence is 1.5% of the population over 65 years and around 170,000 French are affected [182]. Given the general aging of the population, the prevalence is likely to increase over the next decades. PD is characterized by the death of dopaminergic neurons within the substantia nigra pars compacta (SNpc) due to intraneuronal aggregation of α-synuclein in the form of Lewy bodies and Lewy neurites in the majority of cases. The clinical diagnosis requires the presence of bradykinesia, i.e. slowness of movement, together with one additional motor manifestation among rigidity, resting tremor and postural instability [183]. The clinical diagnosis is confirmed by a sustained response to dopamine replacement therapy. Clinical criteria have a sensitivity of 89% and a positive predictive value of 82% for the diagnosis of PD, while the definitive diagnosis is based on post-mortem confirmation of alpha-synuclein containing Lewy bodies. The direct cost of care increases with disease progression [184].

5.2.2 Multiple system atrophy

Multiple system atrophy (MSA) is a relentlessly progressing rare neurodegenerative disease of unknown etiology. It is characterized by a variable combination of parkinso-

nism, cerebellar impairment, autonomic failure and pyramidal tract signs. Together with PD, it belongs to the group of the synucleinopathies that are characterized by progressive cell loss in the brain due to abnormal aggregation of alpha-synuclein in neurons and glia. MSA usually begins in the sixth decade [185, 186] and has a very poor prognosis, median survival ranges between 5.8 to 9.5 years [187, 188, 189]. Revised consensus criteria allow the clinical diagnosis of MSA with two degrees of certainty, "possible" and "probable", while the diagnosis of "definite" MSA requires post-mortem confirmation of alpha-synuclein containing glial cytoplasmic inclusions [190]. Another clinical diagnosis is confirmed by poor response to dopaminergic therapy. Revised consensus criteria include the results of brain imaging as additional features for the diagnosis of "possible" MSA. However, the sensitivity of these criteria remains relatively low for the diagnosis of "possible" MSA and requires further improvement [191]. Depending on the predominant clinical phenotype, revised consensus diagnosis criteria distinguish between MSA-P where parkinsonism predominates and MSA-C where cerebellar symptoms are most prominent. MSA-P accounts for two thirds of cases in Western populations.

5.2.3 Progressive supranuclear palsy

Symptoms of PSP are most commonly seen in people in their early 60's but may begin in some people who are in their 40's [192]. It is also a progressing rare neurodegenerative disease of unknown etiology [193]. Early symptoms of this disease may be related to a person's increased difficulty with walking and balance, often resulting in frequent falls. It is common for a person in the early stages of PSP to develop other motorrelated symptoms like slowed or awkward movements while walking. Symptoms that help to differentiate PSP from other neurodegenerative diseases, like Parkinson's, are often related to a person's vision and eye movements [194]. People with PSP often experience blurred vision and an inability to control eye movements [195]. Some cannot look downward or cannot open their eyelids. Speech and swallowing complaints are seen early. Another characteristic feature is poor response to dopaminergic agents. The mental and physical slowness is also observed

for this disease but not in early stage. There is neither confirm diagnostic test nor specific treatment for this fast progressive disease.

5.3 The challenge of differential diagnosis

No reliable biomarker exists for the differential diagnosis between PD and APD or between APD subtypes. Magnetic resonance imaging (MRI) of the brain may help the clinician by revealing distinct abnormalities in APD patients. It has been recently reported distinct patterns of nigro-striatal involvement in PD and MSA by using multimodal MRI techniques. However, brain MRI can also be normal, especially for patients where the clinical differential diagnosis between PD and APD is difficult. Other imaging techniques such as [18F]-fluorodeoxyglucose positron emission tomography (FDG-PET) allow identifying distinct metabolic patterns in PD and MSA. However, this technique is very costly and not available in clinical routine. Beyond imaging, several studies have compared plasma and cerebrospinal fluid levels of alpha-synuclein, markers of axonal degeneration and catecholamines between PD and MSA-P patients. At this time, no major conclusions can be drawn from these and further efforts are urgently needed to improve diagnostic accuracy between PD and APD and between APD subtypes. An accurate early diagnosis is indeed not only important for assessing prognosis and follow appropriate treatment, but also to understand underlying pathophysiology and development of new therapies [192].

5.4 Dysarthria

Speech disorder is often one of the early manifestation of neurological disorders. Thus, voice/speech analysis can provide valuable information about the underlying disease. Motor speech disorder includes apraxia and dysarthria. The latter affects the streanth, speed, range, steadiness, tone or accuracy of movements required for the breathing, phonatory, resonatory, articulatory and prosodic aspect of speech production. Dysarthria is mostly caused by control or execution impairment of one or more sensorimotor. It can be charac-

terized by weakness, spasticity, incoordination, involuntary movements or variable muscle tone. Conversely, apraxia is another type of motor speech disorders resulting from impaired capacity to plan or program the sensorimotors responsible for directing synchronous movements that produce phonetically and prosodically normal speech. It is quite different from dysarthria and aphasia by it's localization and management. In the pioneer work [196], Darley&Aronson&Brown (DAB) classified dysarthria in 7 types depending on their localization and neuromotor symptoms, as reported in Figure 5.1. Since then, this classification has been the reference, known as the DAB dysarthria classification.

There is consensus that PD patients develop essentially hypokinetic dysarthria [197, 198] while MSA and PSP patients typically exhibit mixed dysarthria with various combination of hypokinetic, ataxic and spastic components [199, 198]. The next sections provide a brief description of these 3 dysarthria types as well as some of the most known findings in their perceptual and objective analysis and evaluation.

	Type	Localization	Neuromotor bases: general	Neuromotor bases: specific
Dysarthria	Flaccid	Lower motor neuron (final common pathway, motor Unit)	Execution	Weakness
	Spastic	Bilateral upper motor neuron (direct and indirect activation pathways)	Execution	Spasticity
	Ataxic	Cerebellum (cerebellar control circuit)	Control	Incoordination
	Hypokinetic	Basal ganglia control circuit (extrapyramidal)	Control	Rigidity, reduced range of movements, scaling problems
	Hyperkinetic	Basal ganglia control circuit (extrapyramidal)	Control	Involuntary movements
	Unilateral upper motor neuron	Unilateral upper motor neuron	Execution/Control	Upper motor neuron weakness, incoordination, or spasticity
	Mixed	More than one	Execution and/or control	More than one
Apraxia of speech		Left (dominant) hemishpere	Motor planning/programming	Planning/ programming errors

FIGURE 5.1 – Types of motor speech disorders

5.4.1 Hypokinetic dysarthria

Hypokinetic dysarthria is a perceptually distinct motor speech disorder (MSD) associated with basal ganglia control circuit pathology. It may manifest in any or all of respiratory, phonatory, resonatory and articulatory level of speech, but its characteristics are more evident in voice, articulation and prosody. This disorder reflects the effect of rigidity, reduced force movement and slow individual but sometimes fast repetitive movements on speech. Decreased range of movements is a significant contributor to this disorder. Parkinson disease (PD) is the typical example of this group, but there are other diseases also associated with hypokinetic dysarthria. Hypokinetic dysarthria affects aspects of speech motor control, such as the preparation, maintenance, and switching of motor programs.

5.4.1.1 Perceptual evaluation

Speech related abnormalities mostly reflected in conversation speech or reading, AMRs and vowel prolongation. Hypokinetic dysarthria can be evident at all level of speech system. Neuromuscular deficits for hypokinetic dysarthria was summarized by Darley, Aronson, and Brown (DAB) in Table 5.1. It shows that range and force are reduced. Muscle tone is excessive. However, reduced range of movement may be the most significant underlying neuromuscular deficit in hypokinesia. General profile of hypokinetic dysarthria was defined in [200]. He showed that 89% of patients had voice abnormalities characterized by hoarseness, roughness, tremulousness, and breathiness, and 45% had articulation problems. 20% had rate abnormalities such as syllable repetitions, shortened sylabbles, lengthened syllables, and excessive pauses. Conversely, DAB found prosodic insufficiency as primary abnormalities in hypokinetic dysarthria. The characteristics include monopitch, monoloudness, reduced stress, short phrases, variable rate, short rushes of speech, and imprecise consonants. In summary, we can organize speech abnormalities according to the speech level components for perceptual analysis. Reduced loudness and utterances length are example of phonatory-respiratory abnormalities. In articulatory system, repea-

ted phonemes, palilalia, rapid or blurred or galloping AMRs are most evident. In addition, reduced stress, monopitch, monoloudness, inappropriate silences, short rushes of speech, variable rate, increased rate in segments, and increased overall rate are the prosodic characteristics for hypokinesia. Most deviant speech dimension associated with hypokinetic dysarthria listed according to severity from most to least in the Table 5.2.

	Direction	Rhythm	Rate		Range		Force	Tone
	Individual movements	Repititive movements	Individual movements	Repititive movements	Individual movements	Repititive movements	Individual movements	Muscle tone
Hypokinetic	Normal	Regular	Slow	Fast	Reduced	Very reduced	Reduced	Reduced Excessive (balanced)
Spastic	Normal	Regular	Slow	Slow	Reduced (weak)	Reduced (biased)	Reduced	Excessive
Ataxic	Inaccurate	Irregular	Slow	Slow	Excessive to normal	Excessive to normal	Normal to excessive	Reduced

TABLE 5.1 – Neuromuscular deficits associated with ataxic dysarthria

5.4.1.2 Objective evaluation

Perceptual heterogeneity among patients in hypokinetic dysarthria is a major issue. To alleviate this limitation, acoustic and physiologic measures have contributed to a richer description and better understanding of the disorder. Four different levels of speech components and related speech characteristics are presented in the folowing.

Respiratory abnormalities occur frequently in hypokinetic dysarthria. Abnormalities in respiratory system reflected as reduced maximum phonation time, reduced airflow volume during vowel prolongation, fewer syllables per breath, shorter utterance length, increased breathing at the time of reading. Some of these characteristics are also related to laryngeal function abnormalities.

Laryngeal abnormalities are mostly reflected in phonatory speech characteristics. Acoustical method confirm hypotheses made by perceptual analyses and provide some additional insights of underlying voice abnormalities.

Resonance abnormalities are not perceptually prominent. Nonetheless, nasal airflow can be increased across the consecutive syllables. It probably due to the reduced degree

Speech dimension	Speech component
Monopitch*	Phonatory prosodic
Reduced stress*	Prosodic
Monoloudness*	Phonatory-respiratory-prosodic
Imprecise consonants	Articulatory
Inappropriate silences*	Prosodic
Short rushes of speech*	Articulatory-prosodic
Harsh voice quality	Phonatory
Breathy voice	Phonatory
Low pitch	Phonatory
Variable rate*	Articulatory-prosodic
Increased rate in segments*	Prosodic
Increase of rate overall*	Prosodic
Repeated phonemes*	Articulatory

TABLE 5.2 – Deviant speech dimensions in hypokinetic dysarthria according to severity (descending) ; "*" means most distinctive dimension.

and velocity of velar movement. Velopharyngeal dysfunction is possibly the secondary to slow movements, rigidity, reduced range of movements for resonance abnormalities.

Articulatory dynamics provide qualitative support for perception of imprecise articulation, rate abnormalities, and reduction in range of articulatory movements. Rigidity and reduced range of motion affect different aspect of speech production. Precision of articulatory movements are affected significantly. Articulatory undershoot is responsible for spirantization, is characterized acoustically by stop-gap by low intensity fricative. In addition, range of movement ; rate, strength and endurance ; tremor, steadiness, and control are affected in articulatory dysfunction.

5.4.2 Spastic dysarthria

Spastic dysarthria is a perceptually distinct MSD produced by bilateral damage to the direct and indirect activation pathway of Central Nervous System (CNS). It may be manifest in all of the respiratory, phonatory, resonatory and articulatory level of speech components. It is characterized by weakness and spasticity which slows movements and reduces its range and force. Spasticity, a hallmark of Upper Motor Neurone (UMN), seems to be the major contributor of this dysarthria. Spasticity is primarily a problem of neuromuscular execution, rather than planning, programming and control.

5.4.2.1 Perceptual evaluation

Spastic dysarthria is studied effectively on conversational and reading speech, speech AMRs, and vowel prolongation. Spastic dysarthria is usually associated with deficits at all the speech valves and for all components of speech system. We can see neuromuscular deficits of spastic dysarthria in Table 5.1. It shows that direction and rhythm of movements are unaffected. The major abnormalities are in slowness and reduced range of individual and repetitive movements, reduced force of movement, and excessive muscle tone. Biased muscle tone is most apparent in laryngeal valve, in which the bias is toward hyperadduction during phonation.

DAB found four clusters of deviant dimensions in patients with pseudobilbar palsy. First cluster is prosodic excess. It include excess and equal stress and slow rate. Increased syllable duration is also evident in spastic dysarthria. Second cluster is articulatory-resonatory incompetencies. It is represented by imprecise consonants, distorted vowels, and hypernasality. Third cluster is prosodic insufficiency. It covers monopitch, monoloudness, reduced stress, and short phrases. Fourth cluster is called phonatory stenosis. It is characterized by low pitch, harshness, strained-strangled voice, pitch breaks, short phrases and slow rate. Slow rate and slow AMR are pervasive and perceptually salient feature for spastic dysarthria particularly for connected speech tasks. So far, features of spastic dysarthria, distinguishable from other dysarthria are strained-harsh voice, monopitch and monoloudness, slow speech rate, and slow and regular speech AMRs. Most deviant speech dimension associated with spastic dysarthria listed according to severity from most to least in the Table 5.3.

5.4.2.2 Objective evaluation

Some of acoustic feature correlates of perceived slow speech rate and prosodic abnormalities commonly associated with spastic dysarthria. Effect of respiratory abnormalities is unclear for spastic dysarthria. However, reduced volume of inhalatory and exhalatory volumes, reduced vital capacity, and reduced maximum vowel prolongation are observed

Speech dimension	Speech component
Imprecise consonants	Articulatory
Monopitch	Phonatory prosodic
Reduced stress	Prosodic
Harsh voice quality	Phonatory
Monoloudness	Phonatory-respiratory
Low pitch*	Phonatory
Slow rate*	Articulatory-prosodic
Hypernasality	Resonatory
Strained-strangled voice quality*	Phonatory
Short Phrases	Respiratory-phonatory-resonatory or articulatory
Distorted vowels	Articulatory
Pitch breaks	Phonatory
Breathy Voice	Phonatory

TABLE 5.3 – Deviant speech dimensions in spastic dysarthria according to severity (descending)

in respiratory abnormalities. In phonatory level, increased jitter and shimmer, decreased HNR and intensity variability are observed. In resonatory part of speech production, palate may move sluggishly or not at all during vowel prolongation. In addition, Palatal immobility, slow movement, and incomplete Velopharyngeal closure may be evident in nasoendoscopy. Incomplete closure may result in hypernasality. Slowness, reduced range and precision in articulatory system lead to slow speech rate, increased word or syllable duration, reduced rate of amplitude variation and slow speech AMRs.

5.4.3 Ataxic dysarthria

Ataxic dysarthria is also a perceptually distinct MSD associated with cerebellar control circuit. It may be manifest in all for level of speech systems, but it is most evident in articulation and prosody. The disorder reflects the effects of incoordination and perhaps reduced muscle tone which results in slowness and inaccuracy in the force, range, timing, and direction of speech movements. Unlike flaccid and spastic dysarthria, which are predominantly problems of neuromuscular execution, ataxic dysarthria predominantly reflects problem of motor control. It is primarily a abnormality of timing and coordination.

The cerebellar control circuit consists of of the cerebellum and its connections. Vermis form the mid-portion of the anterior and posterior lobes of the cerebellum. Right and

left of the cerebellar hemisphere are connected to the opposite thalamus and cerebral hemisphere. Each cerebellar hemisphere helps control movements of ipsilateral side of the body.The lateral cerebellar hemisphere are particularly important to the coordination of skilled voluntary muscle activity.

5.4.3.1 Perceptual evaluation

Conversational speech, reading, and speech AMRs are the most useful task for observing the salient and distinguishing characteristics of ataxic dysarthria. Sentence with multiple syllables may provoke articulatory breakdowns and prosodic abnormalities. Irregular speech AMRs are the salient feature of ataxic dysarthria. It is predominantly an articulatory and prosodic disorder. Neuromuscular deficits in ataxic dysarthria are tabulated in Table 5.1.

DAB found three distinct cluster in ataxia dysarthria. First cluster is articulators inaccuracy, represented by imprecise consonants, Irregular articulatory breakdowns, and vowel distortion. These features reflect inaccurate direction of articulatory movements and dysrhythmia of repetitive movements. Second cluster is prosodic excess, composed of excess and equal stress, prolonged phonemes, prolonged interval and slow rate. Third cluster is phonatory-prosodic insufficiency. It is represented by harshness, monopitch, and monoloudness. Most deviant speech dimension associated with ataxic dysarthria listed according to severity from most to least in the Table 5.4.

5.4.3.2 Objective evaluation

Incoordination of respiratory and phonatory section of speech for isolated phonatory task lead to excessive loudness variation and fundamental frequency variation. Stability of long term and short term phonation are found to be abnormal for cerebellar disease. It has been speculated that asymmetrically distributed motor deficits at laryngeal level and altered sensory control of laryngeal and respiratory reflexes could account for impaired control of tension in intrinsic laryngeal muscle, leading to phonatory instability. Tremor

Speech dimension	Speech component
Excess and equal stress*	Prosodic
Irregular articulatory breakdown*	Articulatory
Distorted vowels*	Articulatory-prosodic
Reduced stress	Prosodic
Harsh voice quality	Phonatory
Prolonged phonemes	Articulatory-prosodic
Prolonged interval	Prosodic
Monopitch	Phonatory-prosodic
Monoloudness	Resonatory-phonatory
Slow rate	Articulatory
Excess loudness variability*	Respiratory-phonatory-resonatory or articulatory
Voice tremor	Phonatory

TABLE 5.4 – Deviant speech dimensions in ataxic dysarthria according to severity (descending)

is not always present in ataxic dysarthria, but a perceptible 3 Hz voice tremor has been observed by acoustic analysis.

In articulatory section, slow rate, abnormalities of rhythm on speech AMR task, and timing abnormalities in VOT are evident for ataxic dysarthria. VOT is a sensitive measure of laryngeal control or laryngeal-articulatory coordination. Some of prosodic abnormalities are also observed in acoustic studies like excessive interword pauses, increased pause length, and irregularity in segment duration. Rhythm metrics (acoustically measured indices of vocalic and consonant segment durations that capture speech rhythm as reflected in patterns of stressed and unstressed syllables in utterances) or envelope modulation spectra (quantification of temporal regularities in the amplitude envelope of the acoustic speech waveform) can distinguish ataxic dysarthria from other kinds of dysarthria.

5.5 Dysarthria-based differential diagnosis

During the last decades, there has been a increasing interest in PD speech and voice analysis. The large majority of research have however focused on discriminating between PD and healthy controls with the motivation to use speech assessment as a supporting method for early PD diagnosis. There exists a very large set of publications in this area, a good and recent review is given in [201]. While this can have an interest from a fun-

Speech dimension	Level of speech	Type of dysarthria	Explanation
MPT	Respiratory	Hypokinetic	MPT is more dependant on air volume at the time of exhalation. Many studies established that rigidity is an common clinical sign for hypokinetic dysarthria. Reduced movements of diaphragm results in reduced airflow volume. Therefore, it is hypokinetic by nature.
Harsh voice	Phonatory	Hypokinetic	Range and rate are usually reduced for hypokinetic dysarthria. Reduced range of neuromuscular activity is the possible reason to abnormal closure of vocal folds. It results in breathy, rough and raspy voice.
Rapid AMR	Articulatory	Hypokinetic	Speech acceleration (short rushes) is an salient feature of hypokinesia. It may be a cause of abnormality of inhibitory activities of basal ganglia control circuit.
Inappropriate silences	Articulatory	Hypokinetic	Rigidity and reduced range of movements are the major reason for inappropriate silences.
Gaping in-between voiced intervals (GIV)	Phonatory	Ataxic/hypokinetic	Phonatory aspects provide information about disabilities to control opening and closing of vocal folds.
Speech rate	Articulatory	Hypokinetic	It is also a parameter to test articulatory movements
Reduced loudness	Phonatory-respiratory	Hypokinetic	It is mostly evident in hypokinetic dysarthria. Reduced range of movement of diaphragm leads to airflow insufficiency. In addition, air leakage at glottis also reduce the loudness.
Monopitch	Phonatory	Hypokinetic	Reduced contrstivity is a common underlying feature for hypokinetic dysarthria. During sentence reading the F_0 variability reduced.
Pitch range	Phonatory	Hypokinetic	It is also a measures of neuromuscular range of vocal folds. We have already observed that reduced range of neuromuscular activities for hypokinetic dysarthria.
Imprecise vowels	Articulatory	Hypokinetic/spastic	Due to reduced range of neuromuscular movements vowel space is reduced.
Dysfluency	Articulatory	Hypokinetic	It is due to short
Impresise consonant articulation	Articulatory	Hypokinetic, ataxic, spastic	Precision and range of movements are responsible for this feature. Articulatory undershoot is the example for Imprecise consonant articulation. Due to abnormal articulatory closure, stop consonants are perceived as noisy fricative.
Articulation decay	Articulatory	Hypokinetic	It is also related to reduced range and rate, rigidity of articulatory movements.
Rate of speech timing		Hypokinetic	Don't know exactly
Duration of stop consonants (DUS)	Articulatory	Hypokinetic	Spirantization is related to the abnormal closure of velar. It introduce fricative type noise to stop consonants.
Relative loudness of respiration (RLR)	Respiratory-phonatory	Hypokinetic	Hypokinesia and decreased range of rib cage motion were measured using relative loudness of respiration.
Rate of speech respiration (RSR)	Respiratory	Hypokinetic	Rigidity and reduced range of movements lead to air capacity. Therefore, increase rate of inspiration is usually observed.
pause intervals per respiration (PIR)	Respiratory	Hypokinetic	Inappropriate silence is part of hypokinesia. Pauses are more due to initiation problem. It is probably due to the degradation of basal ganglia control circuit.
Latency of respiratory exchange (LRE)	Respiratory	Hypokinetic	Increased latency of exchange between expiration and inspiration associated with rigidity and bradykinesia of respiratory muscles.
Acceleration of speech timing (AST)	Articulation	Hypokinetic	Similar as rapid AMRs ; it is computed on monologue.
Entropy of speech timing (EST)	Articulatory-phonatory-respiratory	Probably mixed H,A,S	

TABLE 5.5 – Different dysarthrias

Speech dimension	Level of speech	Type of dysarthria	Explanation
Strained-strangled voice	Phonatory	Spastic	Biased to excessive adduction or resistance to abduction. It is the example of laryngeal spasticity.
Slow AMR	Articulatory	Spastic	In spastic dysarthria, rate of movement is slow. In addition, range and precision is reduced. Slow movement results in increased syllables duration. It leads to slow AMRs.
Slow rate	Articulatory	Spastic	Prolongation of phonemes, syllables and words leads to slow speech rate. Prolongation is mostly due to slow movements of articulators.
Duration of voiced intervals (DVI)	Phonatory	Ataxic	Phonemes prolongation is more evident in ataxic dysarthria. It is probably due to the abnormality of inhibitory activity of cerebellar control circuit.
Excess pitch fluctuations	Phonatory	Ataxic	In sustained vowel pronunciation, movements are regular but mostly lacks of timing and coordination lead to excess pitch fluctuation. Incoordination of respiratory and phonatory system lead to abnormal pitch fluctuations.
Vocal tremor	Phonatory	Ataxic	Cerebellar tremor of around 3Hz is generally observed in ataxic dysarthria.
Irregular AMRs	Articulatory	Ataxic	Irregular activity is linked to abnormalities of control and coordination. Therefore, it is an salient feature of ataxic dysarthria.
Excessive intensity variation	Respiratory-phonatory	Ataxic	It is also mostly attributed as abnormalities of control and coordination of laryngeal and resonatory system.
Rhythm instability (RI)	Articulatory	Ataxic	RI was calculated as the sum of absolute deviations of each observation in terms of gaps duration from the regression line, weighted to the total speech ; Measure irregularity of rhythm.
Duration of voiced intervals (DVI)	Articulatory	Atxic	Prolonged phoneme is an example of ataxic dysarthria, which might be due to cerebellar damage

TABLE 5.6 – Different dysarthrias

damental perspective, it has actually a limited impact from the clinical point of view. Indeed, early diagnosis of PD cannot be claimed (as often done) because APD dysarthria is not taken into account. Moreover, most of the time the clinical diagnosis even neglects the possibility of an APD. The resulting speech dataset can thus be noisy in the sense that patients considered as PD may actually be APD. Such studies may claim at best methods/features which correlate with a diagnosis of *Parkinsonism* (which groups PD and APD).

On the other hand, there exists only few studies on comparison/discrimination between PD and APD or between APD subgroups [202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221]. I thus launched the clinical research project, Voice4PD-MSA (voice4pd-msa.inria.fr), which focuses on dysarthria-based differential diagnosis between PD and MSA-P.

5.6 The Voice4PD-MSA project

As mentioned earlier, because of the similarity between PD and MSA-P symptoms in early disease stages, the differential diagnosis between these two diseases can be very challenging. Indeed, despite recent efforts, no validated objective marker is currently available to guide the clinician in this differential diagnosis. The need of such markers is thus very high in the neurology community, particularly given the severity of the MSA-P prognosis. The ultimate goal of this research is two-fold, that is, to develop a non-invasive objective vocal biomarker to assist first in the early differential diagnosis between PD and MSA-P, then in the early diagnosis of PD. This would be a world premiere as we would be the first scientists to achieve such an important contribution in the field of clinical diagnosis of neurodegenerative diseases. Moreover, our ambition is to develop a very low cost and portable digital tool. In case of success, this would allow a large scale and world-wide use ; every neurologist would have the possibility to use the digital marker in his clinic or private office.

The Voice4PD-MSA projet is a **pilot** study with the objective of developing a **proof-**

of-concept on the differential diagnosis between PD and MSA-P. That is, we search for distinctive speech features which can be useful in early differential diagnosis. To do so, we have to address and solve several scientific challenges :

— We have to identify and design discriminative speech features which lead to a high classification accuracy (higher than 80%).

— We have to develop a classifier which is robust to intra/inter speaker variations and also to recording conditions.

— The final digital software have to be easy to use by the clinicians and have to be flexible enough to incorporate new knowledge and/or new data.

5.6.1 The consortium

The high quality of the consortium is a major asset of the Voice4PD-MSA project. Each partner has indeed a long and solid experience in its area of expertise as well as strong scientific indicators and is internationally recognized. GeoStat has now an established position in the field of nonlinear analysis of complex signals/systems as well as a good expertise in machine learning. SAMoVA has a long history and a well-known expertise in (healthy and pathological) speech processing. IMT has strong expertise and statistical analysis and modeling. The neurology departments of the Bordeaux and Toulouse university hospitals are the two French reference centres for MSA and have a strong expertise and world-wide leadership in the field of PD and MSA. This makes our consortium highly complementary with regard to the project objectives. Moreover, the project methodology is structured in such a way to develop a synergy that exploits at best the skills of each partner. In addition, all the partners have a long and solid experience in running national and international projects :

— INRIA (Coordinator of the project). GeoStat (geostat.bordeaux.inria.fr/) is an INRIA project-team located at INRIA Bordeaux Sud-Ouest. The team makes fundamental and applied research on new nonlinear methods for the analysis of complex systems, turbulent and natural signals, using paradigms and tools coming from the

notions of scale invariance, predictability, universality classes and nonlinear physics. GeoStat's research thematics are centred on nonlinear signal processing using methods from complex systems, statistical physics and nonlinear physics, sparse and compact representations, signal reconstruction, predictability in complex systems, analysis, classification, detection in complex signals.

— IRIT is a French laboratory organized in seven scientific themes among which Information Analysis and Synthesis. The SAMoVA team (www.irit.fr/recherches/SAMOVA/) belongs to this theme. SAMoVA focuses its research activities mainly on audiovisual content analysis, modeling and automatic structuring : from signal (audio, speech, music, images) analysis for feature extraction and segmentation at different temporal scales to multimedia objects modeling (speakers, languages, human shapes and gestures) and content structuring. This is applied to different kinds of content for different kinds of applications such as : structured audiovisual content analysis, spoken and music content analysis, media and video surveillance, multimedia indexing and multimodal human robot interaction. SAMoVA is also working on automatic speech intelligibility and comprehension for older persons presbycusis metering, and on the quality of pronunciation of disordered speech based on goodness of pronunciation measures.

— IMT is a French laboratory organized in three scientific themes among which Statistics and Probability. The Statistics and Probability team covers all the areas of the random domain from theoretical aspects to practical ones in various contexts including epidemiology, biometrics and biostatistics. Many past and current collaborations with biology laboratories enable to develop skills in the identification of biomarkers from high-throughput technologies data (genomics, proteomics etc...). Methodologies used in this context rely on discriminant analysis from either multivariate exploratory analysis or Machine Learning techniques. The transfer of methodologies from the statistical domain to the application field is made easier through the involvement of research engineer in this project.

— The University Hospital Bordeaux hosts since 2007 the Bordeaux site of the French National Reference Centre for MSA, creation in the framework of the first French national plan for rare disorders 2005-2008. The French Reference Centre for MSA is unique in Europe bringing together expertise in multidisciplinary patient care (100 MSA patients are seen every year in Bordeaux) and research. Since 2012 the University Hospital Bordeaux also hosts an expert center for PD (creation in the framework of the national plan for PD 2011-2014 ; 1000 PD patients are seen every year). The investigators of the MSA and PD centres are the same and have high-level research expertise as indicated by their capacity to attract public and private funding for clinical research in MSA and PD, as well as their publication track record (imn-bordeaux.org/).

— The University Hospital Toulouse hosts since 2007 the Toulouse site of the French National Reference Centre for MSA,since 2012 the regional expert center of PD (www.chu-toulouse.fr/-centre-de-reference-de-l-atrophie-multisystematisee) and since 2004 the Clinical Investigation Center. The cohort of followed MSA and PD patients are similar compared to Bordeaux. The investigators of the MSA and PD centres in Toulouse also have high-level research expertise as indicated by their capacity to attract public and private funding for clinical research in MSA and PD, as well as their publication track record. The Clinical Investigation Center has an internationally recognized expertise and has been working in close collaboration with the investigators of the MSA and PD centres for many years.

5.6.2 Beyond Voice4PD-MSA

It is important to emphasize that we have mid and long-term ambitious objectives beyond Voice4PD-MSA. Indeed, based on the results obtained in this project, we plan to launch a large multi-center study involving de novo patients with parkinsonian syndromes (i.e. patients who see for the first time in their life a neurologist because of their parkinsonism). This trial may start as early as next year, if the results of the clinical

proof-of-concept study are conclusive or promising. General neurologists will participate in this study by recording patients with a parkinsonian syndrome in their office at first visit. Disease progression will then be followed and the patients' voice/speech will be recorded every 12 months for at least four years. The goals of this large multi-center study will be three-fold :

— Validation and refinement of the objective vocal biomarker that will be developed for the differential diagnosis between PD and MSA-P in the framework of the Voice4PD-MSA proof-of-concept study.

— Extension of the validity to the differential diagnosis of other neurodegenerative parkinsonian disorders which may resemble to PD in early stages, such as progressive supranuclear palsy and corticobasal syndrome. Indeed, the methodology and the final software developed in Voice4PD-MSA will be structured in a way that would allow easy adaptation to other APDs.

— Assessing the validity of the tool for the early diagnosis of PD. As mentioned earlier, most research in this area uses voice datasets of PD and HC, followed by the extraction of speech/voice features that achieve the best classification accuracy. This is a "normal" approach given the relative youth of such research and the lack of data. In our vision, while such research is interesting to improve fundamental knowledge about PD speech characteristics, focusing on it at this point of time is not a priority. The first reason is ethical because of the absence of a PD treatment. From an ethical point of view, an early diagnosis of PD without an accompanying treatment can be seriously questionable. Another reason is technical : differential diagnosis between PD and similar symptomatic diseases should be first addressed (as we do in Voice4PD-MSA). Otherwise such a diagnosis would actually collapse to early diagnosis of Parkinsonism, which is already achievable by current clinical tools. Our strategy is thus to first achieve the Voice4PD-MSA objectives and then address the PD early diagnosis task in a long-term strategy, given that PD treatments are expected to be available in 2 or 3 decades. A long-term process is also necessary to

follow disease declaration/progress and speech quality of patients (in order to pretend achieving an early diagnosis). From this perspective, Voice4PD-MSA can be seen as the necessary first step towards achieving these ultimate mid and long-term crucial and ambitious objectives.

5.6.3 The challenge of data collection

Since the official start of the project mid-2017, we have been facing numerous and unexpected difficult problems in data collection. The reasons behind these difficulties range from technical, medical and administrative problems to Covid-19 pandemic consequences. Despite tremendous efforts, it's only few months ago that we reached a sufficient number of inclusions which allow acceptable statistical analysis.

I then immediately started the investigation of distortion of consonants given that is known to be a frequent impairment in dysarthria. The next section describes this work.

5.7 Differential diagnosis between MSA and PSP

Because of the delay in collecting the Voice4PD-MSA data, and prior to the work presented in the previous section, we actually started investigating the problem of discrimination between MSA and PSP. The first reason was the availability of the richest existing dataset in this field, which has been kindly provided to me by my collaborators at the SAMI team of the Czech technical university in Prague (sami.fel.cvut.cz). The second reason is that this problem is fundamentally similar to differential diagnosis between PD and MSA (my main objective).

We investigated this problem by considering the constraint that only a small amount of training data can be available in this setting. To do so, we performed univariate statistical analysis followed by a supervised learning that forces the designed new features to be 1-dimensional. We carried out experiments using speech recordings of MSA and PSP Czech patients.

This research was a continuation of a pioneer work [209] which provided a quantitative and objective analysis of speech characteristics for the discrimination between PD and APD and between MSA and PSP. The basic conclusion was that PD speakers manifest pure hypokinetic dysarthria, ataxic components are more affected in MSA whilst PSP subjects demonstrate severe deficits in hypokinetic and spastic elements of dysarthria. Using an SVM with a Gaussian radial basis kernel and an exhaustive search, [209] reported a score of 75% in discrimination between PSP and MSA.

We first explored standard linear and generalized linear models to address the curse of dimensionality problem in this setting. This led us to build simple linear models which achieved an 80% accuracy in classification between MSA and PSP [222]. Then, in [214] , we focused on defining new speech features which can objectively measure particular dysarthria attributes and which are disease-specific, in the sense that such features would have a (statistical) behavior for PSP which is significantly different than for MSA. Moreover, we wanted such features to be clinically interpretable in order to improve the understanding of speech impairments in PSP and MSA. An important benefit of such a strategy is to potentially allow drawing hypothesis regarding the early stage of the diseases. I present in the following the development we made in order to achieve this objective.

5.7.1 Dataset

From 2011 to 2014, 12 consecutive patients with the clinical diagnosis of probable PSP (10 men, 2 women) and 13 patients with the diagnosis of probable MSA (6 men, 7 women) were recruited. In this series, 9 PSP patients were diagnosed with the Richardson's syndrome (PSP-RS), 2 with PSP-parkinsonism (PSP-P) and 1 with PSP-pure akinesia with gait freezing (PAGF), whereas 10 MSA patients were diagnosed as the parkinsonian type (MSA-P) and 3 as cerebellar type (MSA-C). The diagnosis of PSP was established by the NINDS-PSP clinical diagnosis criteria [223], MSA according to consensus diagnostic criteria for MSA [224]. Speech severity did not perceptually differ between PSP and MSA based on UPDRS speech item 18. A detailed description of the patients is given in [209].

Speech recordings were performed in a quiet room with a low ambient noise level using a head-mounted condenser microphone (Bayerdynamic Opus 55, Heilbronn, Germany) situated approximately 5 cm from the mouth of each subject. Speech signals were recored with 48 kHz sampling frequency and 16-bit resolution. Each participant was instructed to perform sustained phonation of the vowel /a/ per one breath as long and steadily as possible, fast /pa/-/ta/-/ka/ syllable repetition at least seven times per one breath, a reading passage and a monologue on a given topic for approximately 90 s. All participants performed the sustained phonation and syllable repetition tasks twice.

5.7.2 Acoustic features

In earlier studies [225, 226] several acoustic parameters were explored, the detailed description is given in [227]. In the same manner as in our study [222], in order to allow easy future comparisons or reproduction, we considered the same set of 13 features that can be computed with existing and established scripts. In [222], we adopted a "phonetic" point of view to group the feature because our main concern was the investigation the usability of linear models to address the curse of dimensionality. In [214], our concern was to find/define features which can be specific to each disease and which can lead to a clinical interpretation. We thus adopted a "symptomatic" point of view (as in [209]) to categorize the features into the well-known 3 dysarthria groups : Hypokinetic, Spastic and Ataxic. The set of the 13 features we considered is presented in Table 5.7. All the features were computed using Python and Praat scripts [228]. For phonation and syllable repetition, the final value of each feature was calculated as the mean of the values obtained from two vocal tasks.

5.7.3 Methodology and experiments

In all experiments, the MSA and PSP data is used to normalize the acoustic features to zero mean and unit variance.

Speech features	Vocal task	Description
Hypokinetic :		
1. Harsh voice	Sustained phonation /a/	Jitter : Frequency perturbation ; Shimmer : Amplitude perturbation ; Harmonics-to-noise ratio (HNR) : amount of noise in voiced speech
2. Rapid Alternating Motion Rate (AMR)	Syllable repetition	It is measured as pace acceleration ; It provides impression of rapid and blurred speech
3. Inappropriate silences		
i) Percent pause time (PPT)	Reading passages	PPT is measured as the percentage of pause time relative to total speech time
ii) Number of pauses (No. of pauses)	Reading passages	No. of pauses measured as the average number of pauses per second
iii) Intra-word pause ratio	Reading passages	The intra-word pause ratio measured as the ratio between the total pause time within polysyllabic words and the total pause time.
5. Monopitch	Monologue	Monotone voice, lacking normal pitch and inflection changes
Spastic :		
1. The degree of voicelessness (DUV)	Sustained phonation /a/	DUV represents the fraction of pitch frames marked as unvoiced.
2. Slow AMR	Syllable repetition	It is measured as the DDK rate of the first seven repetitions of the /pa/-/ta/-/ka/ syllables.
Ataxic :		
1. The excess pitch fluctuation (F0_SD)	Sustained phonation /a/	Pitch fluctuation measured as the standard deviation of voice pitch
2. Irregular AMR	Syllable repetition	It is measured as the standard deviation of distances between consecutive positions of syllables in the first seven repetitions of /pa/-/ta/-/ka/
3. Vocal tremor	Sustained phonation /a/	Vocal tremor is measured as the frequency tremor intensity index (FTRI) defined as the intensity/-magnitude of the strongest low-frequency modulation of F0 [229].

TABLE 5.7 – List and description of the 13 features, grouped by dysarthria type

5.7.3.1 Univariate statistical analysis

Given that we were faced with the curse of dimensionality problem, we needed to perform feature selection and/or reduction. The simplest way to do feature selection is to carry out a univariate statistical analysis. We used the latter as a first pass to discard the highly correlated (thus less discriminative) features. To do so, for each feature, we first computed the intra MSA (resp. PSP) class mean μ_{msa} (resp. μ_{psp}) and variance σ_{msa} (resp. σ_{psp}). We then used Mahalanobis distance, defined in Equation 5.1, as a measure of the the inter-class separation.

$$D = \sqrt{(\mu_{psp} - \mu_{msa})^2 . (\frac{\sigma_{psp} + \sigma_{msa}}{2})^{-1}} \tag{5.1}$$

The result of this univariate analysis is given in Figure 5.2. This analysis suggested that monopitch can be fairly discard from the hypokinetic set of features. Likewise, this analysis suggested that DUV and slow AMR can be disregarded. We kept them however in our second stage of analysis in order to check whether this "first indication" would stand in our second stage of analysis. As for the ataxic group, it seemed like all features are contributing towards discrimination. We thus ended up considering the following sets of features :

— H = {jitter, shimmer, HNR, intra-word pause, rapid AMR, no. of pauses, PPT}

— A = {F0_SD, irregular AMR, vocal tremor}

— S = {DUV, slow AMR}

5.7.3.2 Learning a new speech feature

We then performed a second stage of feature selection and dimension reduction. Nowadays, most of "hot" machine learning problems and methods deal with large data sets (big data problems). In many applications however, such as biomedical engineering, data are rare and/or are collected with a (very) low time resolution. The problem we addressed falls in this setting ; we were in a small dataset machine learning scenario. Indeed,

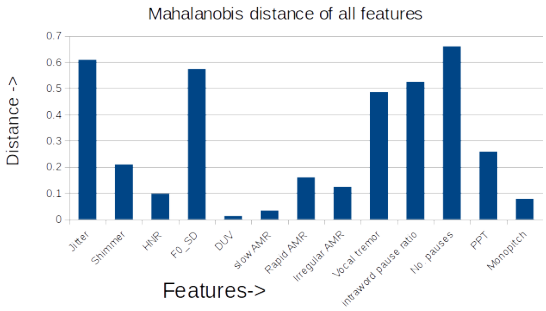


FIGURE 5.2 – Feature-wise distance between PSP and MSA

	H	A	S
Accuracy (%)	68	8	0

TABLE 5.8 – Classification accuracy for individual dysarthric groups

only 12 PSP and 13 MSA patients were available. Thus, typically only a 1-dimensional feature space may provide acceptable statistics. Univariate statistical analysis, as performed in the previous section, is one solution. In fact, a more involved univariate analysis has been carried out [209], using the same dataset. This led to the argument that ataxic components are more affected in MSA whilst PSP subjects demonstrate severe deficits in hypokinetic and spastic elements of dysarthria. However, univariate analysis did not allow us to achieve the goal we targeted, that is, finding a **scalar** speech feature which is disease specific and which is clinically interpretable.

In [222], we presented a methodology to deal with the machine learning part of our small dataset problem. We showed that classical linear and generalized linear models, such as Factorial Discriminant Analysis (FDA) (also known as descriptive LDA) [230], can provide a simple solution to this problem. We thus naturally investigated whether such models could lead us to our goal.

	H+S+A	H+S	S+A	H+A
Accuracy (%)	72	60	24	**84**

TABLE 5.9 – Classification accuracy for combined dysarthric groups

	Hypokinetic							Ataxic		
	Jitter	Shimmer	HNR	Intra-word pause	No. of pauses	PPT	Rapid AMR	F0_SD	Irregular AMR	Vocal
Weight	1.03	3.02	4.77	-0.10	-0.02	-1.11	0.80	2.14	-0.49	0

TABLE 5.10 – Feature weights obtained by FDA

For classification, as the amount of data is small, a Leave-One-Speaker-Out (LOSO) training approach was adopted in all the experiments. In order to have a clear understanding of the data behavior, we used a simple Gaussian classifier and a 1d **linear** Support Vector Machine (SVM) with $C = 1$ as classifiers. The results we obtained were the same using these two classifiers.

We started (naturally) by evaluating FDA on each dysarthric group (H, A, S) individually. The idea here was to check whether a linear combination of a group feature could discriminate between MSA and PSP. Table 5.8 shows the FDA classification scores between PSP and MSA patients by individual dysarthric group. This result showed that FDA on A and S are extremely low and even worse than chance. This suggested that (linear combination of) the ataxic and spastic features we consider cannot be individually disease-specific. The same argument holded for hypokinetic features, though the score is significantly higher.

We then proceeded to evaluate the combination of dysarthric groups. Table 5.9 presents the FDA classification scores for the 4 possible combinations. This result showed that the same argument we made for individual group holds also for H+S+A, H+S and S+A. Interestingly the score obtained by H+A was significantly high, even higher than the score in our recent work. This was a very good indication that measuring the "mutual amount" of hypokinetic and ataxic dysarthria can allow discrimination between MSA and PSP. Moreover, our result showed that a relatively simple (weighted) averaging can measure this amount.

We then went deeper in the analysis of this averaging. Table 5.10 shows the weights obtained by FDA when training using all data (no LOSO). By looking at the weights, one observes that intra-word pause and no. of pauses have lower weights compared to the

other feature. This suggested that they can be discarded in the linear FDA projection. We did so and rerun a LOSO training by discarding intra-word pause and no. of pause. This yielded a classification score of **88%**. This meant that the arguments we stretched about hypokinetic and ataxic features (H+A) hold even better when discarding these two hypokinetic features. Figure 5.3 shows the values of the resulting variable for PSP (blue) and MSA (green) patients. The latter shows that a very good separation is indeed obtained. These classification scores should however be considered with precaution because of the LOSO bias (different weights at each iteration).

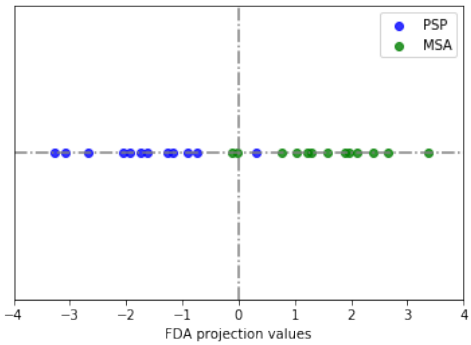


FIGURE 5.3 – Values of the new speech feature for each patient

5.7.4 Discussion and conclusion

We addressed the difficult problem of defining disease-specific speech features which is crucial in the perspective of early differential diagnosis in Parkinsonism. We focused on MSA and PSP and investigated this problem under the constraint of small dataset machine learning. Using FDA, we ended up defining a new scalar variable which can be considered as a disease-specific feature. This variable was learned from data as a linear combination of some hypokinetic and ataxic features. Using Gaussian or 1d SVM classifiers, we obtained a threshold which measures a certain degree of hypokinetic and ataxic "impairment". Above (resp. below) this threshold, the patient is classified as MSA (resp. PSP). Using LOSO training, we achieved 88% classification accuracy, which is an improvement w.r.t the 80% obtained in our previous work [222]. This achievement can be considered very successful

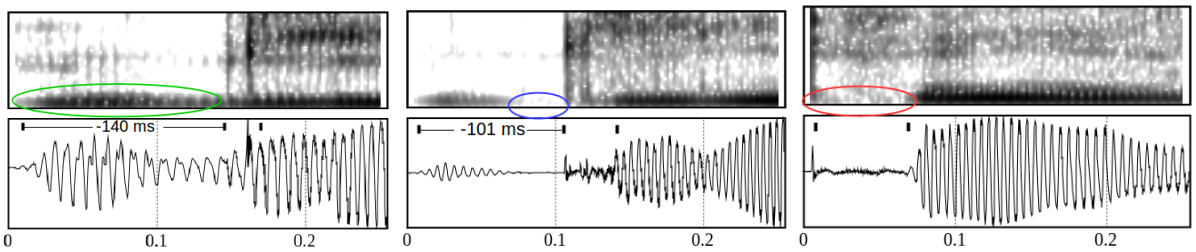
as PSP and MSA are hardly perceptually distinguishable [211]. This result also suggests that hypokinetic and ataxic dysarthria convey considerable discriminative information when mutually considered. Moreover, it suggests that the hypokinetic and ataxic features we retained can be used as a vehicle to capture this information. We showed indeed that an appropriate weighted averaging of these features (the new variable) led to a high classification accuracy. It must be clear however that these arguments are not a "hard" conclusion. They need indeed to be confirmed by additional data and studies.

5.8 Distortion of voiced obstruents for differential diagnosis between PD and MSA-P

The aim of this study was to investigate distinctive patterns in the distortion of voiced obstruents (plosives and fricatives). It is the first study which attempts to examine such distortions in the French language for the purpose of the differential diagnosis between PD and MSA-P (and among the very few studies if we consider all languages). We carried out a perceptual and objective analysis of voiced obstruents extracted from isolated pseudo-words initials. We first showed that devoicing is a significant impairment which predominates in MSA-P. We then showed that voice onset time (VOT) of voiced plosives (prevoicing duration) can be a complementary feature to improve the accuracy in discrimination between PD and MSA-P. I present the details of this work in the following.

5.8.1 Introduction

Dysarthria can manifest in all the levels of speech production [180]. In particular, the articulatory mechanism can be affected which causes deficits in range, strength, timing, stability and precision of articulators [200]. One of the most common manifestation of such deficits is imprecise consonant articulation. In the pioneer work [197], consonant realization was perceptually found to be one of the most deviant speech dimensions in PD.



(a) HC : syllable /gi/ (b) PD : syllable /gi/ (c) MSA-P : syllable /gi/

FIGURE 5.4 – Example of no/partial/total devoicing of /g/ in a HC/PD/MSA-P (top). Example of normal/shorter/vanishing VOT of /g/ for the same HC/PD/MSA-P (bottom)

Consonants distortion across various diseases have been typically assessed using perceptual evaluation [200, 202, 231, 232, 206, 216, 211]. During the last 2 decades, a considerable effort has been produced to develop objective measures that assess consonant distortions in PD [203, 233, 226, 234, 235, 219, 221, 236]. In these studies, voice onset time (VOT) has been the most analyzed feature but with rather contradictory outcomes [237, 238, 226, 219]. On the other hand, only few studies have addressed consonants distortion in differential diagnosis between PD and APD [216, 217, 219, 220, 221].

As for the French language, to the best of my knowledge, there exists no study comparing consonant production between PD and MSA. In [216], a comparison has been subjectively performed (spectrogram visual inspection) but between PD, Amyotrophic Lateral Sclerosis and Cerebellar Ataxia. This work is thus the first study on French consonant distortion for the purpose of differential diagnosis between PD and MSA-P. We carried out a subjective and objective analysis of word-initial consonants using pseudo-words, called Logatomes [239]. Among all the consonants, obstruents (plosives and fricatives) yielded the most interesting results. In particular, we showed that voiced obstruents manifest appealing distinctive impairments, in term of devoicing and VOT duration. We then provided a 2-dimensional analysis over these two deviant speech dimensions. This led us to build a decision model which discriminates between PD and MSA-P with a good accuracy.

5.8.2 Database

From 2018 to the time of writing this thesis, a total of 43 French speakers were recruited in the framework of the Voice4PD-MSA. 20 patients (5 females and 15 males) were diagnosed with idiopathic PD, with a mean age of 60 and a mean symptom duration of 4 years. 12 subjects (8 females and 4 males) were diagnosed with MSA-P, with a mean age of 67 and a mean symptom duration of 3.5 years. 11 healthy controls (HC) with a mean age of 56 (6 female and 5 male) with no history of neurological or communication disorders were recruited. ENTs carried out all the recording sessions for all participants. Each participant performed several speech tasks including sustained phonations, syllables repetition, a reading task, a monologue and a set of 25 isolated pseudo-words called Logatomes (other non-speech biosignals are also recorded). We used only the Logatomes dataset in this study. The speech signals were recorded with 48kHz sampling frequency and 16 bit resolution by a headmount condenser microphone (t.bone HC 444 TWS) placed at a distance of approximately 5cm from the speaker's mouth. Ethics approval was obtained prior to recruitment and all participants gave written informed consent.

5.8.3 Method and results

This work work dedicated to the analysis of the production of consonants extracted from the Logatomes. The latter have the advantage of being an easy speech task for patients (independently on their native language) and easy to process (even manually). We performed an auditory and visual examination of the 25 Logatomes produced by each participant. Auditory examination was performed by listening carefully several times to all the audio files. Visual examination was carried out by inspection of waveform and wide-band spectrogram using Praat [228]. As a consequence of this processing, we manually annotated all the speech signals. We followed the criteria of [238] to set the boundaries of the different phonetic units.

During the examination, we assessed the distortion of consonants (and vowels) when auditorily or/and visually perceived. This analysis showed that, among the 25 Logatomes,

voiced obstruents were significantly distorded, in term of the manifestation of devoicing. We mention that this assessment was based on word-initials only (in CV format) in order to avoid co-articulation and speaking rate effects. We thus focused on the 3 voiced plosives /b/, /d/ and /g/ extracted from the Logatomes "berdo", "dirou" and "guizant", and the 3 voiced fricatives /v/, /z/, and /Z/ extracted from the Logatomes 'vonia", "zacu", and "jinin". After devoicing analysis, we performed the traditional voice onset time (VOT) analysis of the plosives using the manually segmented units. Finally we combined the two analysis, devoicing and VOT, in order to build a decision model for differential diagnosis.

For the statistical analysis, data normality of each acoustic feature was evaluated by the one-sample Kolmogorov-Smirnov test. If data was normally distributed we used the t-test, otherwise the Kruskal-Wallis test to measure group difference between groups was used. Statistical significance was set at a p-value $p < 0.05$.

5.8.3.1 Devoicing analysis

By visual inspection of the spectrograms, we assessed devoicing by the total or partial absence of voicing bars in the realization of voiced obstruents. Figure 5.4 shows an example of spectrograms of the consonant /g/ pronounced normally by a HC, with a partial devoicing by a PD and with a total devoicing by an MSA-P patient. We observed an other phenomenon which could be considered as partial devoicing, the occurrence of voicing bars with weak energy. However, for sake of clarity, reproducibility and to reduce subjectivity effects, we did not use this criterion in our assessment of devoicing.

We found that 67% of MSA-P and 15% of PD presented devoicing in at least one obstruent. In particular, we did not observe any devoicing of /b/ nor /d/ in PD while 42% of MSA-P showed devoicing in these consonants. This suggests that devoicing of /b/ or/and /d/ could be a signature of MSA-P. We mention here however that [216] reported that 37% of PD presented devoicing in /d/ or /g/. In our data, only 10% of PD showed devoicing in /g/ (and thus in /d/ or /g/), while 33% of MSA-P showed devoicing in /d/ or /g/. This difference in PD is might be due to the relatively small size of our dataset

as compared to the one of [216]. As for voiced fricatives, 33% of MSA-P and 15% of PD presented devoicing in at least one of /v/, /z/ and /Z/.

We then provided an objective measure to detect devoicing as follows. Given a labeled consonant, a simple way to assess total or partial absence of voicing is to consider the degree of voicing measure :

$$DV = \frac{number\ of\ voiced\ frames}{total\ number\ of\ frames}(\%)$$

Using the very soft threshold 50%, one can fairly consider that an obstruent (or a consonant in general) is devoiced if $DV < 50\%$ (we use Praat to compute DV). It must be emphasized here that other objective criteria can be used to define and detect devoicing, our measure is however easy to interpret and to reproduce. Given a speaker, we defined his/her total degree of voicing DVT as the average DV over all devoiced obstruents or over all the obstruents if none is devoiced. DVT is a simple quantification of the global amount of voicing in the obstruents devoiced by a given speaker (if any). DVT is thus always less (resp. higher) than 50% in the presence (resp. absence) of devoicing. Using this measure, the assessement of devoicing matched perfectly with our visual observations, that is, 67% of MSA-P and 15% of PD presented devoicing. The value of DVT is shown in Figure 5.5 for each participant (see the projection over the DVT dimension). Along this dimension, one can note the large margin between subjects manifesting devoicing and the others. This suggests that devoicing is generally strong when it occurs, and thus easy to detect objectively by the standard tool Praat.

Overall, these results showed that devoicing can be a valuable cue for differential diagnosis between PD and MSA-P. However, this cue alone is not sufficient to achieve this diagnosis with a high accuracy.

5.8.3.2 VOT analysis of voiced plosives

As mentioned earlier, VOT is among the most studied features in consonant distortion. VOT is generally associated with plosives and is defined as the duration between the vocal fold vibration starts relative to the release of the plosive (there exist however VOT

Feature	Consonant	HC vs PD	HC vs MSA-P	PD vs MSA-P
		p-value		
VOT	/b/	0.36	0.15	0.41
	/d/	0.8	0.9	0.66
	/g/	0.07	**0.004**	**0.03**
VOTR	/b/	0.08	**0.001**	**0.029**
	/d/	0.26	**0.03**	0.161
	/g/	**0.002**	**0.0009**	**0.014**

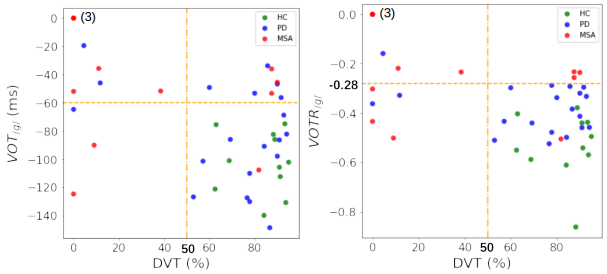
TABLE 5.11 – Results of acoustic speech analyses for three voiced plosives including /b/,/d/ and /g/. Bold numbers indicate group difference ($p < 0.05$)

definitions for other consonant types [240]). In the case of voiced plosive, vibration begins before the release, VOT is thus considered as negative. When negative VOT tends to 0, it actually corresponds to a total devoicing. In order to avoid a potential dependency on speaking rate, VOT ratio (VOTR) is sometimes considered. VOTR is defined as VOT divided by the duration of whole syllable [238].

Our purpose here was to determine whether VOT analysis of voiced plosives (/b/, /d/ and /g/) can yield another distinctive cue (hopefully complementary to devoicing). Using our manual segmentation, we computed the VOT and VOTR statistical group difference between HC and PD, HC and MSA-P, PD and MSA-P. Table 5.11 shows the obtained p-value of each group difference. We first observed that, for VOT, statistical significance between MSA and the other groups was achieved only for /g/. More interestingly, this impairment was more severe in MSA-P than in PD. The waveforms of Figure 5.4 shows an example of such a distortion. This trend was confirmed by VOTR with an additional group difference between PD and HC. Globally, this is in accordance with the findings of [219] which reported shorter VOT and lower VOTR for MSA averaged on all voiced plosives (with Czech patients). We could not however confirm the same statement for /b/ and /d/. On the other hand, one can confidently consider that VOT/VOTR of /g/ is a valuable cue for the differential diagnosis. However, as devoicing, this cue alone is not sufficient to achieve this diagnosis with a high accuracy.

5.8.3.3 Classification

Given the findings of the previous sections, it was natural to proceed with an analysis over the 2 deviant speech dimensions, devoicing and VOT of /g/ ($VOT_{/g/}$). Figure 5.5a shows the biplot of DVT w.r.t to $VOT_{/g/}$. Using our HC data, the mean/standard deviation of the VOT of /g/ is $-103/22(ms)$. This is in accordance with the $-109/32(ms)$ reported in [219] and [238] (the latter reported the mean only). As we did for devoicing, we can set a very soft threshold at $-60ms$ above which we confidently consider that a VOT impairment of /g/ is occurring.



(a) (b)

FIGURE 5.5 – Biplot of $DVT(\%)$ w.r.t to $VOT_{/g/}$ and $VOTR_{/g/}$ (dotted line represent decision thresholds) ; (3) means that 3 MSA-P patients have same coordinates (total devoicing)

We observed that all but one MSA-P manifested devoicing or/and short VOT. Thus using the simple decision tree of Figure 5.6, with the soft thresholds $DVT = 50\%$ and $VOT_{/g/} = -60ms$, we obtain an accuracy of 72%, with a high sensitivity (correctly classified MSA-P) of 92% but a low specificity of 60%. This means that a mis-diagnosis of MSA-P presenting devoicing or short VOT of /g/ is unlikely. This statement does not hold for PD.

Figure 5.5b shows the biplot of DVT w.r.t to $VOTR_{/g/}$. We see now that, along the $VOTR_{/g/}$ dimension, a separation appears between the 3 MSA-P and 5 PD which were confused using $VOT_{/g/}$ (right top rectangle of 5.5b). If we target only classification score and replace the threshold $VOT_{/g/} = -60ms$ by $VOTR_{/g/} = -0.28$ in the decision tree,

then the specificity increased to 85% and the accuracy to 87.5%. However, the threshold −0.28 is likely overfitted to our data. Thus, given the small amount of instances of /g/ and its following vowels, we cannot confidently claim that VOTR is a better feature for discrimination than VOT. These results show however that the prevoicing duration of /g/ (and probably all voiced plosives) could be a complementary cue to devoicing of obstruents in order to achieve a high accuracy differential diagnosis between PD and MSA-P.

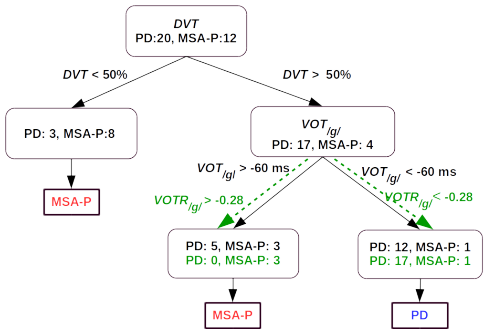


FIGURE 5.6 – Decision tree using DVT and $VOT_{/g/}$ or $VOTR_{/g/}$ (in green) dimensions for discrimination between PD and MSA-P

Overall, the results (along with literature reporting) show that devoicing of voiced obstruents and VOT of /g/ are 2 distinctive and deviant speech dimensions which are worth considering in the differential diagnosis between PD and MSA-P.

5.8.4 Discussion and conclusion

This work constitutes the first study that attempts to highlight distinctive cues in the distortion of French voiced obstruents realization in PD and MSA-P. Our results partially confirmed previous findings on VOT with other languages [217, 219]. Indeed, we found that VOT of the voiced plosive /g/ was significantly reduced in MSA-P while it was natural for most PD. On the other hand, VOT in not the only factor of the distortion of voiced plosives (and obviously fricatives). We showed indeed that the absence of voicing leads was the main factor of voiced obstruents distortions and is the most distinctive cue between PD and MSA-P (in the production of voiced obstruents). Moreover, there was a perfect correlation in devoicing assessment between perceptual and objective evaluations.

This supports a potential use of devoicing in clinical practice as an additional tool for the examination of patients with a suspicion of MSA-P. We also showed that the combination of VOT and devoicing can significantly improve the differential diagnosis accuracy.

VOT impairment can be explained by a difficulty in initiating articulation resulting from a deficit in maintaining the speech motor program [219]. The latter is a characteristic of hypokinetic dysarthria which is a known feature in both PD and MSA. An accurate production of word-initial voiced plosives requires a precise coordination between glottal opening and articulatory closure. Devoicing is a manifestation of an impairment of such coordination. This is a characteristic of ataxic dysarthria which is known to manifest in MSA. Our results are thus in accordance with the consensus that PD develop essentially hypokinetic dysarthria while MSA develop a mixed type dysarthria. More importantly, since ataxia seems to be responsible for devoicing, the latter might manifest in early disease stages. If proven, devoicing would thus constitute a valuable deviant speech dimension to consider in early differential diagnosis.

There are some limitations to our study. The most significant one is obviously the relatively small size of the dataset due to the difficulty of recruiting patients, particularly with a rare disease such as MSA-P. We are however still continuing the effort of recruitment. Moreover, the dataset is unbalanced in gender, we cannot thus exclude that our findings are biased by gender-specific effects. However, no gender-specific characteristics of VOT have been reported in healthy speech. Another limitation is that we used only one consonant instance per speaker, we do not thus know how the results stand to intra-speaker pronunciation variability. We can expect however that the effect of such variability is reduced by the restriction to word-initials. From this perspective, our study should be considered as a promising first step in the analysis of French voiced obstruents in PD and MSA-P. Our findings need to be confirmed by additional data. This is the purpose of our on going research.

Conclusion et perspectives

Sommaire

I . Parole et troubles respiratoires 178

I . I . Le projet VocaPnée 178

I . II . Justification scientifique 180

I . III . Méthodologie 183

II . Étude la dysarthrie 185

II . I . Poursuite de Voice4PD-MSA 185

II . II . Passage à la médecine de ville 185

II . III . Méthodologie 186

À l'horizon des 5 prochaines années (au moins), mes recherches futures s'inscriront exclusivement dans le cadre de la parole pathologique. La page "image" est tournée même si, comme je l'expliquerai plus tard, certaines des connaissances théoriques que j'ai acquises dans ce domaine me seront très utiles. Alors que je m'orientais tout naturellement vers une spécialisation dans l'étude de la dysarthrie, la pandémie Covid-19 a ajouté une nouvel axe de recherche majeur à la perspective de mes recherches actuelles et futures, l'étude des troubles de la parole causés par les affections respiratoires. Je commence par décrire mes perspectives pour cette dernière. Je présenterai ensuite mes perspectives pour l'étude de la dysarthrie.

I . Parole et troubles respiratoites

Dès le début de la crise Covid-19, j'ai identifié le potentiel du traitement de la parole pour la gestion des patients en télé-médecine ainsi que dans l'aide au triage des flux téléphoniques aux urgences. Ainsi, dès premier confinement, j'ai proposé ces idées à la DG d'Inria. Cette proposition a immédiatement suscité l'intérêt d'Inria, du MESRI et du corps médical. Inria a mis à disposition des moyens exceptionnels et nous avons ainsi commencé à travailler sur la réalisation d'un outil pour aider au suivi à domicile de patients atteints du Covid, le projet CovidVoice était né. Alors que l'ambition initiale était de fournir très rapidement un tel outil pour participer à l'effort national, les rouages de l'administration ont posé de nombreux obstacles qui ont empêché la réalisation de cette ambition. Cependant, l'intérêt scientifique et clinique d'une telle approche existe aussi pour d'autres maladies respiratoires. Je dirige ainsi depuis, avec Thomas Similowski responsable du service de pneumologie et de réanimation de la Pitié-Salpêtrière et de l'UMR-S 1158, le projet VocaPnée (le terme CovidVoice n'était plus d'actualité !).

I . I . Le projet VocaPnée

L'objectif de ce projet d'envergure est le développement d'un biomarqueur vocal de la fonction respiratoire et de son évolution au cours du télé-suivi à domicile après une affection respiratoire (Covid ou autre). Ce projet ambitieux requière des développements scientifiques, technologiques et organisationnels importants. Le volume de ce projet est ainsi très grand puisqu'il implique, outre les partenaires cliniciens et informaticiens d'AP-HP, plusieurs intervenants Inria (DGS, DGT, SED, équipe Tau et potentiellement d'autre équipes Inria).

VocaPnée se décline en 2 études cliniques longitudinales en parallèle :

• Une étude hospitalière "VocaPnée-Hospit", avec 100 patients, dont le but est le développement et validation d'un biomarqueur vocal de la fonction respiratoire et de son évolution au cours des affections respiratoires aiguës prises en charge en hospitalisation de pneumologie.

• Une étude en télé-médecine ”VocaPnée-Dom”, avec 500 patients, dont le but est le développement et validation d'un biomarqueur vocal de la fonction respiratoire et de son évolution au cours du suivi à domicile après une affection respiratoire aiguë ou de la surveillance évolutive des maladies respiratoires chroniques à risque d'exacerbations.

L'intérêt de la première est de collecter des données de haute qualité dans un environnement controlé et avec des relevés cliniques importants mais inaccessibles à distance. L'intérêt de la seconde est de collecter des données dans des conditions réelles d'utilisation (via la plateforme ORTIF ou COVIDOM dans le cas Covid).

Une maquette de la plateforme qui sera utilisée pour collecter les données des patients a déjà été développée par une équipe d'ingénieurs d'Inria Sophia Antipolis (https ://-dream.inria.fr/vocapnee/). Cette maquette est toutefois utilisée depuis quelques mois pour collecter les enregistrement vocaux de sujets témoins (de façon indépendante des 2 études cliniques). Dans ce cadre, ma première tâche sera d'effectuer une étude « normative » dans le but d'établir des valeurs de références sur les paramètres acoustiques pertinents.

Une fois la phase clinique lancée, outre ma contribution scientifique, une activité cruciale sera d'établir un cadre et un outil pour l'évaluation, et éventuellement la fusion, des différentes techniques qui seront proposées. Ensuite, en cas de succès du projet VocaPnée, je devrais planifier et mettre en place le passage à la phase validation et production. Après presque une année d'efforts considérables, VocaPnée-Hospit vient d'être soumis en Janvier 2021 au Comité de Protection des Personnes (CPP), VocaPnée-Dom. Nous pouvons ainsi envisager l'avenir de façon positive, même si des développements technologiques importants sont encore à faire pour permettre le recueil et le traitement des données sur les serveurs d'AP-HP. Les DSI d'AP-HP et d'Inria travaillent actuellement conjointement sur cette tâche.

Je signale aussi VocaPnée devrait servir aussi pour ma deuxième proposition, l'aide au triage des appels aux urgences. Ce projet à long terme est sur la table et devrait être mené en partenariat avec l'hôpital Necker et le LIMSI, avec qui j'ai déjà entamé des discussions.

I . II . Justification scientifique

I . II ..1 La respiration

La respiration est une fonction végétative qui a pour spécificité de ne pas disposer d'une commande intrinsèque (comme c'est le cas de la circulation -un cœur explanté peut battre tout seul-), mais de dépendre d'une commande nerveuse extrinsèque qui provient du système nerveux central : des oscillateurs neuronaux situés dans le tronc cérébral envoient une commande cyclique à des motoneurones spinaux, qui gouvernent la contraction de muscles respiratoires. Cette organisation anatomo-physiologique donne à la respiration la particularité de pouvoir répondre non seulement à sa commande automatique, permanente, et autorégulée en fonction des besoins métaboliques de l'organisme (circuit bulbospinal), mais aussi à des commandes volontaires, comportementales, dissociées de l'homéostasie (circuit corticospinal). Il est ainsi possible de réaliser des manoeuvres respiratoires volontaires, et d'utiliser la respiration pour des actions non respiratoires (souffler les bougies du gâteau d'anniversaire, jouer du saxophone, parler).

I . II ..2 Parole et respiration

La parole est ainsi l'exemple principal d'utilisation de l'appareil respiratoire et de sa commande neurologique pour une fonction "non métabolique". La parole, qui ne peut prendre place que pendant la phase expiratoire du cycle respiratoire, implique en effet une relégation temporaire du contrôle automatique de la ventilation au second plan. Ainsi, la production d'une phrase exige d'une part une préparation pré-phonatoire (inspiration rapide d'un volume d'air qui correspond, au travers d'un mécanisme de type "feed-forward", à l'intention du locuteur au regard de la longueur de la phrase et du volume sonore ; il s'agit d'un contrôle "excitateur") puis, d'autre part, l'inhibition de l'inspiration automatique tant que la vocalisation de la phrase n'est pas terminée. Il n'y a donc pas de parole possible sans contrôle cortical de la respiration.

Chez un sujet normal, indemne de maladie respiratoire, le système respiratoire dispose d'une marge très largement suffisante pour permettre une parole fluide et adaptable sur

une très large gamme phonique et prosodique. Ainsi, les volumes pulmonaires à mobiliser pour la production des phrases sont très inférieurs aux volumes maximaux accessibles, et la commande respiratoire automatique est suffisamment peu intense pour "accepter" des perturbations importantes (prononcer une phrase correspond grossièrement à réaliser une courte apnée volontaire).

Au contraire, en présence d'anomalies respiratoires aiguës ou chroniques, la parole peut être altérée du fait de l'incapacité du système respiratoire à "assumer" les contraintes correspondantes. Deux principaux facteurs sont en cause. Le premier est une limitation "mécanique", consistant en une réduction de la capacité à faire entrer de l'air dans les poumons et à le faire rapidement. Ceci impose une contrainte sur la préparation respiratoire pré-phonatoire. Un patient souffrant d'une affection respiratoire ne peut pas prendre une grande inspiration pour faire une phrase longue et/ou "forte". Le second facteur est une limitation "dynamique", consistant en une augmentation de l'intensité du contrôle automatique (bulbospinal) de la respiration, qui explique que les patients atteints d'affections respiratoires génératrices de dyspnée respirent rapidement. Pour accélérer la respiration, une "solution" consiste à rendre l'expiration active, alors qu'elle est normalement passive et uniquement due à la rétraction élastique des poumons préalablement distendus par l'inspiration. Ceci supprime la "souplesse expiratoire" liée à l'absence de contrôle moteur de l'expiration normale. Un patient souffrant d'une affection respiratoire aigüe ou chronique ne peut pas moduler à volonté la durée de son expiration, parce que l'augmentation de l'intensité de la commande automatique rend plus difficile la prise de contrôle par la commande corticale. Ces mécanismes expliquent pourquoi les patients souffrants de maladies respiratoires ont tendance à produire des phrases courtes, de faible intensité, et à avoir une prosodie (intonations, accents, modulation de la parole) instable. Des altérations de diverses variables vocales ont ainsi été mises en évidence chez les patients atteints de maladies respiratoires.

I . II ..3 La dyspnée

La dyspnée, symptôme principal de toutes les maladies respiratoires aigües et chroniques, se définit comme une plainte liée à la perception consciente de l'activité respiratoire assortie d'un affect négatif (principalement peur ou anxiété). Source de handicap physique, psychologique et social, la dyspnée joue également le rôle clé de signal d'alarme, en permettant au clinicien de repérer la survenue d'anomalies respiratoires aigües chez des patients sans antécédents, mais aussi l'exacerbation des pathologies respiratoires chroniques. Il importe de souligner que pour que la dyspnée puisse jouer ce rôle, il est nécessaire que le patient perçoive les modifications physiologiques liés à la maladie respiratoire, que son cerveau en effectue le traitement cognitif et affectif, et qu'il soit capable de décrire les sensations et émotions qui en résultent. Il importe également de souligner la nature subjective du symptôme dyspnéique, dont la proportionnalité avec les perturbations physiologiques sous-jacente est incertaine et variable.

Ce sont exactement les mêmes anomalies respiratoires qui perturbent la parole (cf. supra, limitation "restrictive" de la capacité inspiratoire, augmentation de la commande ventilatoire centrale) qui sont à l'origine de la dyspnée. Il n'est ainsi pas surprenant de constater que les anomalies de la parole décrites en relation avec les pathologies respiratoires (cf. supra) ont généralement un lien statistique avec la dyspnée. Ceci fait de ces anomalies des candidats raisonnables à une évaluation en tant que biomarqueurs substitutifs de la dyspnée.

I . II ..4 Déficit de perception de la dyspnée

La mauvaise perception par le patient des variations de son état respiratoire est un facteur de gravité, en ce qu'il complique la prise en charge et expose les patients à des retards thérapeutiques potentiellement catastrophiques. Une telle absence peut exister au cours de l'asthme, au cours duquel on sait que les patients qui ne perçoivent pas ou mal les variations de leur état bronchique sont davantage victimes de crises graves et parfois mortelles. C'est également le cas au cours de la bronchopneumopathie chronique obstructive

(BPCO). Récemment, ce phénomène a été décrit au cours de l'infection par le SARS-CoV-2 qui peut entraîner des insuffisances respiratoires aiguës graves, survenant parfois de façon différée. Dans ce contexte, le constat clinique d'une absence ou quasi-absence de dyspnée malgré des anomalies physiologiques majeures a été fait par de multiples équipes.

I . II ..5 Biomarqueurs physiologiques de la dyspnée, biomarqueurs vocaux

L'absence de dyspnée en réponse aux variations de l'état respiratoire justifie l'intérêt porté aux développements de biomarqueurs objectifs de ces variations. Pour être utilisables et efficaces, ces biomarqueurs doivent idéalement être implémentables à distance (télé-médecine), ne pas nécessiter le recours à un appareillage spécifique, et reposer le moins possible sur la coopération des patients. Une analyse vocale effectuée par téléphone/PC répond à la fois aux critères de pertinence (relation fonction respiratoire - voix - dyspnée) et de faisabilité. Ce type de biomarqueur vocal fournirait une solution appropriée et applicable rapidement à grande échelle.

I . III . Méthodologie

Du point de vue méthodologique , il existe relativement peu de travaux sur l'analyse des troubles de la parole résultant de maladies respiratoires. La récente crise sanitaire a donné lieux à quelques travaux sur les « sons Covid », mais ces travaux restent assez obscures et sans impact clinique, à mon avis. Mon objectif est d'abord d'explorer les méthodes existantes en traitement de la parole pathologique pour en extraire celles susceptibles d'être utiles pour la problématique. Ensuite, il s'agira de développer de nouvelles techniques pour les améliorer ou les compléter. Dans ce cadre, mon premier objectif est de développer une méthode de détection des intervalles respiratoires lors la production de la parole qui soit robuste aux conditions d'enregistrement et de la variabilité intra-patient. Une solution à ce problème difficile est déterminante pour l'identification d'une aggravation ou détresse respiratoire.

Un autre volet est la méthodologie statistique pour développer des modèles descriptifs

et prédictifs de l'état respiratoire du patient à partir des données de surveillance vocale et des données de surveillance physiologiques (fréquence cardiaque, fréquence respiratoire, température, saturation en oxygène, échelle de dyspnée). Étant donné le caractère exploratoire de la recherche, nous allons privilégier deux directions, chacune quantifiant un aspect spécifique des données acquises.

Nous allons d'abord utiliser des méthodes de classification non-supervisées (classification hiérarchique, k-means, ACP, ...) pour identifier les clusters potentiels des patients, en fonction des paramètres acoustiques de la voix. Ces analyses ne tiennent pas compte de l'état des patients au cours de l'analyse. Elles consistent plutôt en une analyse exploratoire des paramètres qui mettraient en évidence, par exemple, le comportement relatif des paramètres obtenus par l'analyse de la parole.

Ensuite, la partie la plus importante de l'analyse statistique se concentrera sur des méthodes de classification supervisée. Contrairement à ce qui aura été fait en utilisant des techniques non-supervisées, les méthodes ici dépendront de l'état des patients. Dans ce cadre, trois approches seront utilisées :

• Des analyses univariées basées sur des tests statistiques standards seront effectuées pour déterminer des différences inter-groupes sur les variables acoustiques pertinentes.

• Des analyses de corrélation et de régression permettront d'étudier l'évolution des paramètres acoustiques en fonction des variables physiologiques.

• Les méthodes d'apprentissage statistique seront aussi utilisées dans le cadre supervisé. Différentes méthodes de construction de modèles statistiques seront testées, et éventuellement combinées : analyse discriminante, régression logistique, arbres de décision, SVM et réseaux de neurone profonds. Ces méthodes permettront d'une part la comparaison avec des données normatives appariées en âge et en sexe, d'autre part la séparation entre les patients stables et ceux présentant une détérioration respiratoire.

II . Étude la dysarthrie

II . I . Poursuite de Voice4PD-MSA

Les premiers résultats du projet Voice4PD-MSA sont très prometteurs (nous sommes d'ailleurs entrain de travailler sur un article de revue qui les décrit). Nous serons bientôt en mesure de fournir certains critères de jugement principaux nécessaires pour passer à la deuxième phase du projet. Cette phase, cruciale pour la validité de l'étude, nécessitera le développement d'un biomarqueur vocal robuste capable de fournir une haute performance de discrimination entre PD et MSA-P (sensitivité et spécificité $> 80\%$). Outre cet aspect « statistique », mon objectif est aussi de fournir des indicateurs qui peuvent renseigner sur l'origine des troubles observées, en terme des sous-systèmes de production de la parole et leur interaction. Cet aspect est non seulement déterminant pour le diagnostic précoce, objectif finale de l'étude, mais pourrait aussi ouvrir des pistes aider à comprendre l'étiologie de la MSA-P. Le travail sur la réalisation des consonnes est un premier pas dans cette direction et il sera étendu et approfondi puis valorisé dans un futur très proche.

II . II . Passage à la médecine de ville

Une fois l'étude pilote complétée, l'étape suivante est de passer à la validation du biomarqueur proposé pour le diagnostic précoce. Ce projet de recherche, encore à mettre en place, impliquera la médecine de ville. Un nombre important de cabinets de neurologies participeront au projet pour collecter des échantillons de voix de patients présentant des syndromes Parkinsoniens dès leurs premières consultations. Il s'agira ensuite de mener une étude longitudinale, sur 2 ou 3 ans, pour suivre l'évolution des dysarthries et de leur corrélation avec notre biomarqueur. Je souligne que, pour la collecte des données, je prévois d'utiliser ici une version adaptée de la plateforme de collecte développée dans le cadre du projet VocaPnée.

II . III . Méthodologie

Du point de vue méthodologique, à la lumière de nos résultats, je compte suivre la direction suivante pour réaliser les objectifs. Dans un premier temps, détecter et caractériser les dysarthries par les déficiences des sous-systèmes de production de la parole : respiration, phonation, articulation et prosodie. Ceci permettra notamment de fournir aux phoniatres un moyen d'identifier et étudier les sources possibles des dérèglements dans la planification et/ou l'exécution de la parole. Dans un deuxième temps, utiliser cette caractérisation pour dresser un tableau des déficiences par type de dysarthrie : hypokinétique, ataxique et spatique. Ceci permettra notamment de fournir aux neurologues un moyen pour corréler l'évaluation acoustique avec les observations ou les connaissances récemment acquises par l'utilisation de l'imagerie IRM du cerveau. Outre sa pertinence scientifique, une telle approche a aussi l'avantage de fournir des éléments de langage compréhensibles par les 3 acteurs de cette recherche (acousticiens, phoniatres et neurologues), très utiles pour une telle recherche pluri-disciplinaire.

Du point de vue fondamental, l'approche non-linéaire est encore plus d'actualité. En effet, d'une part, on dispose maintenant de données assez riches (en particulier l'électroglottographie), pour évaluer (et améliorer le cas échéant), dans le cadre pathologique, certaines techniques que nous avons développées (détection GCI, estimation du signal source). D'autre part, nos expérimentations montrent effectivement que les outils d'analyse standard, développés pour la parole saine, se retrouvent souvent désarmés face à la parole Parkinsonienne. Je compte ainsi développer une version « pathologique », ou au moins « Parkinsonienne, d'une partie de ces outils. Pour ce faire, je compte m'appuyer sur certaines connaissances que j'ai acquises dans le cadre du transfert technologique à la société I2S. En effet, les fondements théoriques sur lesquels repose ce transfert sont basées sur des techniques d'optimisation sparse non-convexe, développées dans le cadre de la thèse de Hicham Badri (hal.inria.fr/tel-01239958) qui reçu le prix de thèse ARFIF 2016. Déjà durant cette thèse, j'avais identifié le grand potentiel de cette approche pour l'analyse de la parole, mais nous n'avons jamais eu le temps d'aborder ce sujet. Grâce

aux travaux effectués dans le cadre du Labcom Inria-I2S, je suis encore plus convaincu du potentiel considérable de ces techniques surtout pour l'analyse de la parole dysarthrique. Le premier chantier que je vais commencer est d'appliquer cette approche pour la généralisation du modèle de prédiction linéaire. Ceci permettra (je pense) le développement, sous un même formalisme, de nouvelles versions d'algorithmes standards mieux adaptés pour la parole dysarthrique (estimation spectrale, segmentation phonétique...).

Sur un autre plan, celui de l'apprentissage statistique, je vais explorer les approches émergentes en "small data deep learning" qui pourraient s'avérer utile pour l'étude de la dysarthrie (ainsi que pour les troubles respiratoires). Contrairement aux applications "usuelles" de l'apprentissage profond, ces deux problématiques ne peuvent pas disposer de corpus de données assez grands (comme beaucoup d'applications biomédicales). Dans ce cas, un pré-traitement particulier doit être appliqué, notamment avec des techniques d'augmentation de données ou de "transfer learning". Cependant, il ne s'agira pas d'appliquer ces techniques de façon brutale et boîte noire, mon objectif est en effet de développer des modèles "profonds" qui gardent un certain degré d'interprétabilité, essentiel du point de vue clinique. Dans ce cadre, une doctorante, Soukaina Wakrim, vient d'être recrutée début 2021 à l'université de Rabat par mon collaborateur K. Minaoui (avec qui j'ai collaboré sur l'Upwelling). S. Wakrim travaillera sur l'apprentissage profond en dysarthrie sous ma co-direction, avec la perspective d'une thèse en co-tutelle à partir de 2022.

Bibliographie

[1] H. Bourlard and S. Dupont. A new ASR approach based on independent processing and recombination of partial frequency bands. International Conference on Spoken Language Processing (ICSLP), 1996.

[2] S. Dupont and H. Bourlard. Multiband approach for speech recognition. Workshop on Circuits, Systems, and Signal Processing 1996.

[3] H. Hermansky, S. Tibrewala, and M. Pavel. Towards ASR on partially corrupted speech. International Conference on Spoken Language Processing (ICSLP), 1996.

[4] H. Fletcher. *Speech and hearing in communication*. Krieger, New-York, 1953.

[5] J. Allen. How do humans process and recognize speech. *IEEE Trans. Speech and Audio Processing*, 2(4) :567–576, 1994.

[6] J. Pearl. *Probabilistic reasoning in intelligent systems : Networks of plausible inference*. Morgan Kaufmann Publishers, San Mateo, CA, 1988.

[7] J. A. Bilmes. Data-driven extensions to hmm statistical dependencies. In *International Conference on Spoken Language Processing (ICSLP)*, 1998.

[8] J. A. Bilmes. Buried markov models for speech recognition. In *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 1999.

[9] J. A. Bilmes. *Natural Statistical Models for Automatic Speech Recognition*. PhD thesis, International Compute Science Institute, Berkeley, California, 1999.

[10] K. Daoudi, D. Fohr, and C. Antoine. A new approach for multi-band speech recognition based on probabilistic graphical models. In *International Conference on Spoken Language Processing (ICSLP), Beijing, China*, October 2000.

[11] K. Daoudi, D. Fohr, and C. Antoine. Continuous Multi-Band Speech Recognition using Bayesian Networks. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Terento, Italy*, December 2001.

[12] M. Deviren and K. Daoudi. Structural Learning of Dynamic Bayesian Networks in Speech Recognition. In *EUROSPEECH, Alborg, Denemark*, September 2001.

[13] G. G. Zweig and S. Russell. Speech recognition with dynamic bayesian networks. In *Proceedings Fifteenth National Conference on Artificial Intelligence*, Madison, Wisconsin, 1998.

[14] G.G. Zweig. *Speech Recognition with Dynamic Bayesian Networks.* PhD thesis, University of California, Berkeley, 1998.

[15] G. G. Zweig and S. Russell. Probabilistic modeling with bayesian networks for automatic speech recognition. In *International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia, 1998.

[16] G. G. Zweig and S. Russell. Probabilistic modeling with bayesian networks for automatic speech recognition. *Australian Journal of Intelligent Information Processing Systems*, 5(4) :253–260, 1999.

[17] R. Cowell, A. Dawid, S. Lauritzen, and D. Spiegelhalter. Probabilistic Networks and Expert Systems. *Springer-Verlag*, 1999.

[18] E. Castillo, J.M. Gutiez, and A.S. Hadi. *Expert Systems And Probabilistic Network Models.* Springer-Verlag, New York, 1997.

[19] M. Jordan, editor. Learning in graphical models. *MIT Press*, 1999.

[20] P. Smyth, D. Heckerman, and M.I. Jordan. Probabilistic independence networks for hidden Markov probability models. *Neural Computation*, 9(2) :227–269, 1997.

[21] Z. Ghahramani. An introduction to hidden markov models and bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(1) :9–42, 2001.

[22] K. Murphy. *Dynamic Bayesian Networks : Representation, Inference and Learning*. PhD thesis, UC Berkeley, Computer Science Division, July 2002.

[23] F. Jensen, S. Lauritzen, and K. Olsen. Bayesian updating in recursive graphical models by local computations. *Computational Statistics and Data Analysis*, (4) :269–282, 1990.

[24] A. Dawid. Applications of a general propagation algorithm for probabilistic expert systems. *Statistics and Computing*, (2) :25–36, 1992.

[25] M. Yannakakis. Computing fill-in is NP-complete. *SIAM Journal of Algebraic Discrete Mathods*, 2 :77–79, 1981.

[26] R. E. Tarjan and M. Yannakakis. Simple linear-time algorithms to test chordality of graphs, test acyclity of hypergraphs and selectively reduce acyclic hypergraphs. *SIAM Journal of Computing*, 13 :566–579, 1984.

[27] A. Morris, A. Hagen, H. Glotin, and H. Bourlard. Multi-stream adaptive evidence combination for noise robust asr. *Speech Communication*, 2001.

[28] C. Cerisara and D. Fohr. Multi-band automatic speech recognition. *Computer Speech and Language*, 15(2) :151–174, 2001.

[29] D. Heckerman. A tutorial on learning with bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Advanced Technology Division, March 1995.

[30] M. Deviren and K. Daoudi. Continuous speech recognition using structural learning of dynamic Bayesian networks. In *EUSIPCO*, Toulouse, France, 2002.

[31] D. Heckerman, D. Geiger, and D.M. Chickering. Learning bayesian networks : The combination of knowledge and statistical data. *Machine Learning*, 20 :197–243, 1995.

[32] S. Lauritzen. Propagation of probabilities, means, and variances in mixed graphical association models. *Jour. Amer. Stat. Ass.*, 87(420) :1098–1108, 1992.

[33] G. Gravier, M. Sigelle, and G. Chollet. Markov Random Field modeling for Speech Recognition. *Australian Journal of Intelligent Information Processing Systems*, 5(4) :245–252, 1999.

[34] G. Gravier. *Analyse statistique ux dimensions pour la modsation segmentale du signal de parole : Application reconnaissance*. PhD thesis, ENST Paris, 2000.

[35] K. Weber, S. Bengio, and H. Bourlard. Increasing speech recognition noise robustness with hmm2. In *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2002.

[36] K. Weber, S. Bengio, and H. Bourlard. Speech recognition using advanced hmm2 features. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2001.

[37] K. Weber, S. Bengio, and H. Bourlard. Hmm2- extraction of formant features and their use for robust asr. In *EUROSPEECH*, 2001.

[38] N. Mirghafori and N. Morgan. Transmissions and transitions : A study of two common assumptions in multi-band asr. In *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 1998.

[39] N. Mirghafori and N. Morgan. Sooner or later : Exploring asynchrony in multi-band speech recognition. In *EUROSPEECH*, 1999.

[40] R.K. Moore. A dynamic programming algorithm for the distance between two finite areas. *IEEE Trans. on PAMI*, 1(1) :86–88, 1979.

[41] D. Reynolds, T. Quatieri, and R. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10 :19–41, 2000.

[42] M. Schmidt and H. Gish. Speaker identification via support vector machines. In *Proc. ICASSP*, 1996.

[43] Z. Lei, Y. Yang, and Z. Wu. Mixture of support vector machines for text-independent speaker recognition. In *Proc. Interspeech*, 2005.

[44] J. Kharroubi, D. Petrovska-Delacretaz, and G. Chollet. Combining GMM's with suport vector machines for text-independent speaker verification. In *Proc. Eurospeech*, 2001.

[45] V. Wan. *Speaker Verification using Support Vector Machines*. PhD thesis, University of Sheffield, 2003.

[46] P. Ho and P Moreno. SVM kernel adaptation in speaker classification and verification. In *Proc. ICSLP*, 2004.

[47] N. Dehak and G. Chollet. Support vector GMMs for speaker verification. In *Proc. IEEE Odyssey*, 2006.

[48] W.M. Campbell, J.P. Campbell, D. Reynolds, E. Singer, and P. Torres-Carrasquillo. Support vector machines for speaker and language recognition. *Computer Speech and Language*, 20 :210–229, 2006.

[49] J.-W. Xu, P. Pokharel, K.-H. Jeong, and J. C.Principe. An explicit construction of a reproducing gaussian kernel hilbert space. In *Proc. ICASSP*, 2006.

[50] V. Wan and S. Renals. Speaker verification using sequence discriminant support vector machines. *IEEE Trans. on Speech and Audio Processing*, 2004.

[51] W.M. Campbell, D.E. Sturim, D.A. Reynolds, and A. Solomonoff. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In *Proc. ICASSP*, 2006.

[52] N. Krause and R. Gazit. SVM-based speaker classification in the GMM models space. In *Proc. IEEE Odyssey*, 2006.

[53] T.S. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. *Advances in Neural Information Processing Systems*, 11, 1998.

[54] M. Seeger. Covariance kernels from bayesian generative models. *Advances in Neural Information Processing Systems*, 14, 2002.

[55] T. Jebara, R. Kondor, and A. Howard. Probability product kernels. *Journal of Machine Learning Research*, 5, 2004.

[56] S. Lyu. A kernel between unordered sets of data : the gaussian mixture approach. In *Proc. ECML*, 2005.

[57] T. Jebara and R. Kondor. Bhattacharyya and expected likelihood kernels. In *Proc. Annual Conference on Computational Learning Theory and Kernel Workshop*, 2003.

[58] P. Moreno and P. Ho. A new SVM approach to speaker identification and verification using probabilistic distance kernels. In *Proc. Eurospeech*, 2003.

[59] R. Kondor and T. Jebara. A kernel between sets of vectors. In *Proc. ICML*, 2003.

[60] S. Zhou and R. Chellappa. From sample similarity to ensemble similarity : Probabilistic distance measures in reproducing kernel hilbert space. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(6) :917–929, 2006.

[61] H. Shimodaira, K.I. Noma, M. Nakai, and S. Sagayama. Support vector machine with dynamic time-alignment kernel for speech recognition. In *Proc. Interspeech*, 2001.

[62] V. Wan and J. Carmichael. Polynomial dynamic time warping kernel support vector machines for dysarthric speech recognition with sparse training data. In *Proc. Interspeech*, 2005.

[63] C. Wallraven, B. Caputo, and A. Graf. Recognition with local features : the kernel recipe. In *Proc. ICCV*, 2003.

[64] J. Mariéthoz and S. Bengio. A max kernel for text-independent speaker verification systems. In *Proc. MMUA*, 2006.

[65] S. Boughorbel, J.-P. Tarel, and F. Fleuret. Non-mercer kernels for SVM object recognition. In *Proc. British Machine Vision Conference (BMVC)*, 2004.

[66] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[67] G.H. Golub and C.F. Van Loan. *Matrix Computation*. The John Hopkins Univ. Press, 1996.

[68] B. Schölkopf, S. Mika, C. J.C.Burges, P. Knirsch, K.-R. Mller, G. Rch, and A. J. Smola. Input space versus feature space. *IEEE Trans. Neural Networks*, 100(5) :1000–1017, 1999.

[69] M.J. Daniels and R.E. Kass. Shrinkage estimators for covariance matrices. *Biometrics*, 57(4) :1173–1184, 2001.

[70] Y.-S. Choi, H.-C. Shin, and W.-J. Song. Affine projection algorithms with adaptative regularization matrix. In *Proc. ICASSP*, 2006.

[71] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[72] B. Schölkopf, A.J. Smola, and K.R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5) :1299–1319, 1998.

[73] A.J. Smola and B. Schölkopf. Sparse greedy matrix approximation for machine learning. In *Proc. ICML*, 2000.

[74] S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2 :243–264, 2001.

[75] F.R. Bach and M.I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3 :1–48, 2002.

[76] F.R. Bach and M.I. Jordan. Predictive low-rank decomposition for kernel methods. In *Proc. ICML*, 2005.

[77] G. Blanchard, P. Massart, R. Vert, and L. Zwald. Kernel projection machine : a new tool for pattern recognition. In *Proc. NIPS*, 2004.

[78] C.K.I. Williams and M. Seeger. *Advances in Neural Information Processing Systems*, volume 13, chapter Using the Nystrm Method to Speed Up Kernel Machines. MIT Press, 2001.

[79] C.K.I. Williams and M. Seeger. Effect of the input density distribution on kernel-based classifiers. In *Proc. ICML*, 2000.

[80] S. Mika, B. Schölkopf, A.J. Smola, K.R. Mller, Scholz M., and Rch G. *Advances in Neural Information Processing Systems*, volume 11, chapter Kernel PCA and de-noising in feature spaces, pages 536–542. MIT Press, 1999.

[81] M.E. Tipping. Sparse kernel principal component analysis. *Advances in Neural Information Processing Systems*, 13, 2001.

[82] R. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 :179–188, 1936. Academic Press.

[83] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35 :99–109, 1943.

[84] The NIST year 2005 speaker recognition evaluation plan. `http://www.nist.gov/speech/tests/spk/2005/sre-05_evalplan-v5.pdf`, 2005.

[85] Biosecure network of excellence : Biometrics for secure authentification. `http://www.biosecure.info`, 2005.

[86] I. Magrin-Chagnolleau, G. Gravier, and R. Blouet. Overview of the 2000-2001 ELISA consortium research activities. In *Proc. IEEE Odyssey*, 2001.

[87] J. Pelecanos and S. Sridharan. Feature warping for robust speaker verification. In *Proc. IEEE Odyssey*, 2001.

[88] Y. Linde, A. Buzo, and R.M. Gray. An algorithm for vector quantization design. *IEEE Trans. on Communications*, 28(1) :84–95, 1980.

[89] J. Louradour and K. Daoudi. Conceiving a new sequence kernel and applying it to SVM speaker verification . In *Proc. Interspeech*, 2005.

[90] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10(1-3) :42–54, 2000.

[91] C. Barras and J.-L. Gauvain. Feature and score normalization for speaker verification of cellular data. In *Proc. ICASSP*, 2003.

[92] J.-F. Bonastre, F. Wils, and S. Meignier. Alize, a free toolkit for speaker recognition. In *Proc. ICASSP*, 2005.

[93] N. Scheffer and J.-F. Bonastre. Fusing generative and discriminative UBM-based systems for speaker verification. In *Proc. of the 2nd international workshop on MMUA (MultiModal User Authentification)*, 2006.

[94] T. F. Quatieri. *Discret-time speech signal processing principles and practice.* Prentice-Hall, 2001.

[95] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[96] J. F. Kaiser. Some observations on vocal tract operation from a fluid flow point of view. In I. R. Titze and R. C. Scherer, editors, *Vocal Fold Physiology : Biomechanics, Acoustics, and Phonatory Control*, pages 358–386. The Denver Center for the Performing Arts, 1983.

[97] Gernot Kubin. *Speech coding and synthesis*, chapter Chapter 16 : Nonlinear processing of speech, pages 557–610. Elsevier, 1995.

[98] M. Faundez-Zanuy, G. Kubin, W. B. Kleijn, P. Maragos, S. McLaughlin, A. Esposito, A. Hussain, and J. Schoentgen. Nonlinear speech processing : Overview and applications. *Control and intelligent systems*, 30 :1–10, 2002.

[99] P. Maragos and A. Potamianos. Fractal dimensions of speech sounds : Computation and application to automatic speech recognition. *Journal of Acoustic Society of America*, 105 :1925–1932, March 1999.

[100] H. M. Teager and S. M. Teager. Evidence for nonlinear sound production mechanisms in the vocal tract. In W.J. Hardcastle and A. Marchal, editors, *Speech Production and Speech Modelling*. NATO Advanced Study Institute Series D, 1989.

[101] S. McLaughlin and P. Maragos. *Nonlinear methods for speech analysis and synthesis, in Advances in Nonlinear Signal and Image Processing*. Hindawi Publ. Corp., 2006.

[102] A. Esposito and M. Marinaro. Nonlinear speech modeling and applications, chapter : Some notes on nonlinearities of speech. pages 1–14. Springer-Verlag, 2005.

[103] T. Koizumi, S. Taniguchi, and S. Hiromitsu. Glottal source - vocal tract interaction. *Journal of Acoustic Society of America*, 78 (5) :1541–1547, 1985.

[104] M. D. Plumpe. Modeling of the glottal flow derivative waveform with application to speaker identification. Master's thesis, Massachusetts Institute of Technology, 1997.

[105] M. A. Little. Mathematical foundations of nonlinear, non-gaussian, and time-varying digital speech signal processing. In *NOLISP*, pages 9–16, 2011.

[106] M.A. Little. *Biomechanically Informed Nonlinear Speech Signal Processing*. PhD thesis, Oxford University, 2007.

[107] M. A. Little, P. E McSharry, I.M. Moroz, and S.J. Roberts. Testing the assumptions of linear prediction analysis in normal vowels. *Journal of the Acoustical Society of America*, 119 :549–558, January 2006.

[108] G. Chollet, A. Esposito, M. Faundez-Zanuy, and M. Marinaro. *Nonlinear Speech Modeling and Applications : Advanced Lectures and Revised Selected Papers*. Lecture notes in computer science : Tutorial. Springer, 2005.

[109] J. Sole-Casals and V. Zaiats. *Advances in Nonlinear Speech Processing : International Conference on Nonlinear Speech Processing, NOLISP 2009, Vic, Spain, June 25-27, 2009, Revised Selected Papers*. Lecture Notes in Artificial Intelligence. Springer, 2010.

[110] Y. Stylianou, M. Faundez-Zanuy, and A. Eposito. *Progress in Nonlinear Speech Processing*. Lecture Notes in Computer Science. Springer, 2007.

[111] C.M. Travieso-González and J. Alonso-Hernández. *Advances in Nonlinear Speech Processing : International Conference on Nonlinear Speech Processing, NoLISP 2011, Las Palmas de Gran Canaria, Spain, November 7-9, 2011, Proceedings*. Lecture Notes in Computer Science / Lecture Notes in Artificial Intelligence. Springer, 2011.

[112] M. A. Little, P. McSharry, S. Roberts, D. Costello, and I. Moroz. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *BioMedical Engineering OnLine*, 6(1) :23, 2007.

[113] V. Pitsikalis and P. Maragos. Analysis and classification of speech signals by generalized fractal dimension features. *Speech Communication*, 51, Issue 12 :1206–1223, December 2009.

[114] H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. Cambridge nonlinear science series. Cambridge University Press, 2003.

[115] I. Kokkinos and P. Maragos. Nonlinear speech analysis using models for chaotic systems. *IEEE Transactions on Speech and Audio Processing*, 13(6) :1098–1109, Jan. 2005.

[116] PE. McSharry, LA. Smith, and L. Tarassenkeo. Prediction of epileptic seizures : are nonlinear methods relevant ? *Nature Medicine*, 9(3) :241–242, 2009.

[117] K. Hu, P.C. Ivanov, Z. Chen, P. Carpena, and H. E. Stanley. Effect of trends on detrended fluctuation analysis. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 64(1 Pt 1), July 2001.

[118] M. A. Little, P. McSharry, I. Moroz, and S. Roberts. Nonlinear, biophysically-informed speech pathology detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, page II, may 2006.

[119] M. Banbrook, S. McLaughlin, and I. Mann. Speech characterization and synthesis by nonlinear methods. *Speech and Audio Processing, IEEE Transactions on*, 7 :1 – 17, Jan 1999.

[120] T. Takagi and M. SUGENO M. Fuzzy identification of systems and its applications to modeling and control. *IEEE transactions on systems, man, and cybernetics*, 15 :116–132, 1985.

[121] J. Tailleur. *Grandes déviations, physique statistique et systèmes dynamiques*. PhD thesis, Universitè Pierre et Marie Curie - Paris VI, 2007.

[122] S. C. Shadden. Lagrangian Coherent Structures Tutorial[Online], http ://mmae.iit.edu/shadden/LCS-tutorial/, 2009.

[123] A. Arneodo, F. Argoul, E. Bacry, J. Elezgaray, and J. F. Muzy. *Ondelettes, multifractales et turbulence*. Diderot Editeur, Paris, France, 1995.

[124] U. Frisch. *Turbulence : The legacy of A.N. Kolmogorov*. Cambridge University Press, 1995.

[125] Hicham Badri. Computer graphics effects from the framework of reconstructible systems. Master's thesis, Rabat faculty of science-INRIA Bordeaux Sud-Ouest, 2012.

[126] J. Grazzini, A. Turiel, H. Yahia, and I. Herlin. Edge-preserving smoothing of high-resolution images with a partial multifractal reconstruction scheme. In *International Society for Photogrammetry and Remote Sensing (ISPRS)*, 2004.

[127] S. K. Maji, H. M. Yahia, O. Pont, J. Sudre, T. Fusco, and V. Michau. Towards multiscale reconstruction of perturbated phase from hartmann-shack acquisitions. In *AHS*, pages 77–84, 2012.

[128] A. Turiel and A. del Pozo. Reconstructing images from their most singular fractal manifold. *IEEE Transactions on Image Processing*, 11 :345–350, 2002.

[129] A. Turiel and N. Parga. The multi-fractal structure of contrast changes in natural images : from sharp edges to textures. *Neural Computation*, 12 :763–793, 2000.

[130] H. Yahia, J. Sudre, V. Garçon, and C. Pottier. High-resolution ocean dynamics from microcanonical formulations in non linear complex signal analysis. In *AGU FALL MEETING*, San Francisco, États-Unis, December 2011. American Geophysical Union.

[131] H. Yahia, J. Sudre, C. Pottier, and V. Garçon. Motion analysis in oceanographic satellite images using multiscale methods and the energy cascade. *Journal of Pattern Recognition*, 2010, to appear. doi :10.1016/j.patcog.2010.04.011.

[132] H. Eugene Stanley. *Introduction to Phase Transitions and Critical Phenomena (International Series of Monographs on Physics)*. Oxford University Press, USA, July 1987.

[133] A. Turiel, H. Yahia, and C.J Pérez-Vicente. Microcanonical multifractal formalism : a geometrical approach to multifractal systems. part 1 : singularity analysis. *Journal of Physics A : Mathematical and Theoretical*, 41 :015501, 2008.

[134] A. Turiel. Method and system for the singularity analysis of digital signals, patent registered under number pct/es2008/070195, 2008.

[135] O. Pont, A. Turiel, and C. J. Pérez-Vicente. Description, modeling and forecasting of data with optimal wavelets. *Journal of Economic Interaction and Coordination*, 4(1), June 2009.

[136] I. R. Titze. *The Myoelastic Aerodynamic Theory of Phonation*. The national center for voice & speech, 2007.

[137] B. H. Story. An overview of the physiology, physics, and modeling of the sound source for vowels. *Acoustical Science and Technology*, 23(4) :195–206, 2002.

[138] K.S.R. Murty and B. Yegnanarayana. Epoch extraction from speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 16 :1602–1613, 2008.

[139] T. Ananthapadmanabha and B. Yegnanarayana. Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27 (4) :309–319, 1979.

[140] M.R.P Thomas amd J. Gudnason and P.A. Naylor. Estimation of glottal closing and opening instants in voiced speech using the yaga algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 20 (1) :82–97, 2012.

[141] E. N. Pinson. Pitch synchronous time domain estimation of formant frequencies and bandwidths. *Journal of the Acoustical Society of America*, 35 (8) :1264–1273, 1963.

[142] K. Steiglitz and B. Dickinson. The use of time-domain selection for improved linear prediction. *IEEE Transactions on Acoustics, Speech and Signal Processing,*, 25 (1) :34–39, 1977.

[143] T. Ewender and B. Pfister. Accurate pitch marking for prosodic modification of speech segments. In *Proceedings of INTERSPEECH*, 2010.

[144] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6) :453 – 467, 1990.

[145] T. Drugman, G. Wilfart, and T. Dutoit. A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis. In *Interspeech conference*, 2010.

[146] Y. Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing*, 9 (1) :21–29, 2001.

[147] N.D. Gaubitch and P.A. Naylor. Spatiotemporal averagingmethod for enhancement of reverberant speech. In *15th International IEEE Conference on Digital Signal Processing*, 2007.

[148] B. Bozkurt and T. Dutoit. Mixed-phase speech modeling and formant estimation, using differential phase spectrums. In *ISCA Voice Quality : Functions, Analysis and Synthesis*, 2003.

[149] T. Drugman. *Advances in Glottal Analysis and its Applications*. PhD thesis, University of Mons, 2011.

[150] D. Wong, J. Markel, and A. Jr. Gray. Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(4) :350–355, aug 1979.

[151] A. Krishnamurthy and. Two channel speech analysis. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34 (4) :730–743, 1986.

[152] B. S. Atal and S. L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *Journal of Acoustic Society of America*, 50 (2B) :637–655, 1971.

[153] P. A. Naylor. Estimation of glottal closure instants in voiced speech using the dypsa algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 15 (1) :34–43, 2007.

[154] R. Smits and B. Yegnanarayana. Determination of instants of significant excitation in speech using group delay function waveform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 3 (5) :325–333, 1995.

[155] C. d'Alessandro and N. Sturmel. Glottal closure instant and voice source analysis using time-scale lines of maximum amplitude. *Sadhana*, 36 :601–622, 2011.

[156] N. Sturmel, C. d'Alessandro, and F. Rigaud. Glottal closure instant detection using lines of maximum amplitudes (loma) of thewavelet transform. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4517–4520, april 2009.

[157] V.N. Tuan and C. d'Alessandro. Robust glottal closure detection using the wavelet transform. In *in Proceedings of the European Conference on Speech Technology*, pages 2805–2808, 1999.

[158] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit. Detection of glottal closure instants from speech signals : A quantitative review. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3) :994 –1006, march 2012.

[159] Aicha Bouzid and Noureddine Ellouze. Glottal opening instant detection from speech signal. In *EUropean Signal Processing COnference (EUSIPCO)*, 2004.

[160] T. Drugman and T. Dutoit. Glottal closure and opening instant detection from speech signals. In *INTERSPEECH*, 2009.

[161] CMU ARCTIC speech synthesis databases. [Online], http ://festvox.org/cmu_arctic.

[162] KED TIMIT database. [Online], http ://festvox.org/.

[163] T. drugman. Gloat toolbox. [Online], http ://tcts.fpms.ac.be/ drugman/.

[164] Noisex-92. [Online], www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html.

[165] D. Giacobello. *Sparsity in Linear Predictive Coding of Speech*. PhD thesis, Multimedia Information and Signal Processing, Department of Electronic Systems, Aalborg University, 2010.

[166] D. Meng, Q. Zhao, and Z. Xu. Improved robustness of sparse pca by l1-norm maximization. *Pattern Recognition Elsevier*, 45 :487–497, 2012.

[167] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[168] Etienne Denoel and Jean Philippe Solvay. Linear prediction of speech with a least absolute error criterion. *IEEE Transactions on Acoustics, Speech and Signal Processing,*, 33 :1397–1403, 1985.

[169] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen. Sparse linear predictors for speech processing. In *Proceedings of the INTERSPEECH*, 2009.

[170] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen. Enhancing sparsity in linear prediction of speech by iteratively reweighted 1-norm minimization. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010.

[171] E. J. Candès and M. B. Wakin. Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier Analysis and Applications*, 14 :877–905, 2008.

[172] E. J. Candès and J. Romberg. l1-magic : Recovery of sparse signals via convex programming. 2005.

[173] D. Giacobello, M. G. Christensen, M. N. Murth, Søren Holdt Jensen, and Fello Marc Moonen. Sparse linear prediction and its applications to speech processing. *IEEE Transactions on Audio, Speech and Language Processing*, 20 :1644–1657, 2012.

[174] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue. DARPA TIMIT acoustic-phonetic continuous speech corpus. Technical report, U.S. Dept. of Commerce, NIST, Gaithersburg, MD, 1993.

[175] N. Hurley and S. Rickard. Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55 :4723–4740, 2009.

[176] B. S. Atal and J. Remde. A new model of lpc excitation for producing natural-sounding speech at low bit rates. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1982.

[177] S. Singhal and B.S. Atal. Amplitude optimization and pitch prediction in multipulse coders. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37 :317 – 327, 1989.

[178] D. Giacobello, M.G. Christensen, M.N. Murthi, S.H. Jensen, and M. Moonen. Retrieving sparse patterns using a compressed sensing framework : Applications to speech coding based on sparse linear prediction. *IEEE Signal Processing Letters*, 17, 2010.

[179]

[180] J.R. Duffy. *Motor Speech Disorders Substrates, Differential Diagnosis, and Management*. Elsevier, 2013.

[181] Lorraine A. Ramig, Ingo R. Titze, Ronald C. Scherer, and Steven P. Ringel. Acoustic analysis of voices of patients with neurologic disease rationale and preliminary data. *Annals of Otology, Rhinology & Laryngology*, 97(2) :164–172, 1988.

[182] F. Tison, J. F. Dartigues, L. Dubes, M. Zuber, A. Alperovitch, and P. Henry. Prevalence of parkinson's disease in the elderly a population study in gironde, france. *Acta Neurologica Scandinavica*, 90(2) :111–115, 1994.

[183] A J Hughes, S E Daniel, L Kilford, and A J Lees. Accuracy of clinical diagnosis of idiopathic parkinson's disease a clinico-pathological study of 100 cases. *Journal of Neurology, Neurosurgery & Psychiatry*, 55(3) :181–184, 1992.

[184] Y Winter, M Balzer-Geldsetzer, Annika Spottke, Jens Reese, Erika Baum, Jens Klotsche, J Rieke, A Simonow, K Eggert, W H Oertel, and R Dodel. Longitudinal study of the socioeconomic burden of parkinson's disease in germany. 17 :1156–63, 03 2010.

[185] A Schrag, Y Ben-Shlomo, and NP Quinn. Prevalence of progressive supranuclear palsy and multiple system atrophy a cross-sectional study. *The Lancet*, 354(9192) :1771 – 1775, 1999.

[186] F Tison, F Yekhlef, V Chrysostome, and C Sourgen. Prevalence of multiple system atrophy. *The Lancet*, 355(9202) :495 – 496, 2000.

[187] Y. Ben-Shlomo, G. K. Wenning, F. Tison, and N. P. Quinn. Survival of patients with pathologically proven multiple system atrophy. *Neurology*, 48(2) :384–393, 1997.

[188] Schrag A., Wenning G. K., Quinn N., and Ben-Shlomo Y. Survival in multiple system atrophy. *Movement Disorders*, 23(2) :294–296.

[189] Gregor Wenning, Y Ben Shlomo, Marina Magalhães, S.E. Daniel, and N.P. Quinn. Clinical features and natural history of multiple system atrophy. an analysis of 100 cases. 117 (Pt 4) :835–45, 09 1994.

[190] S. Gilman, G. K. Wenning, P. A. Low, D. J. Brooks, C. J. Mathias, J. Q. Trojanowski, N. W. Wood, C. Colosimo, A. Dürr, C. J. Fowler, H. Kaufmann, T. Klockgether, A. Lees, W. Poewe, N. Quinn, T. Revesz, D. Robertson, P. Sandroni, K. Seppi, and M. Vidailhet. Second consensus statement on the diagnosis of multiple system atrophy. *Neurology*, 71(9) :670–676, 8 2008.

[191] Osaki Yasushi, Ben-Shlomo Yoav, Lees Andrew J., Wenning G. K., and Quinn Niall P. A validation exercise on the new consensus criteria for multiple system atrophy. *Movement Disorders*, 24(15) :2272–2276, 2009.

[192] Wenning G. K., Litvan I., and Tolosa E. Milestones in atypical and secondary parkinsonisms. *Movement Disorders*, 26(6) :1083–1095.

[193] A. Schrag, Y. Ben-Shlomo, and N.P. Quinn. Prevalence of progressive supranuclear palsy and multiple system atrophy a cross-sectional study. *The Lancet*, 354 :1771–1775, November 1999.

[194] Yasushi Osaki, Yoav Ben-Shlomo, Andrew Lees, Susan Daniel, Carlo Colosimo, Gregor Wenning, and Niall Quinn. Accuracy of clinical diagnosis of progressive supranuclear palsy. *Movement disorders official journal of the Movement Disorder Society*, 19 :181–9, 02 2004.

[195] U. Nath, Y. Ben-Shlomo, R.G. Thomson, A.J. Lees, and D.J. Burn. Clinical features and natural history of progressive supranuclear palsy. *Neurology*, 60(6) :910–916, 2003.

[196] Frederic L Darley, 1928-(joint author.) Aronson, Arnold E. (Arnold Elvin), and 1911-(joint author.) Brown, Joe Robert. *Motor speech disorders.* Philadelphia Saunders, 1975. References interspersed.

[197] F. L. Darley, A. E. Aronson, and J. R. Brown. Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research*, 12(2) :246–269, 1969.

[198] J. Rusz, C. Bonnet, J. Klempir, T. Tykalova, E. Baborov, M. Novotny, A. Rulseh, and E. Ruzicka. Speech disorders reflect differing pathophysiology in parkinson's disease, progressive supranuclear palsy and multiple system atrophy. *Journal of neurology*, 262 :992–1001, 2015.

[199] Kluin K.J., Gilman S., Lohman M., and Junck L. Characteristics of the dysarthria of multiple system atrophy. *Archives of Neurology*, 53(6) :545–548, 1996.

[200] J. A. Logemann and H. B. Fisher. Vocal tract control in parkinson's disease phonetic feature analysis of misarticulations. *The Journal of speech and hearing disorders*, 46 :348–52, 12 1981.

[201] L. Brabenec, J. Mekyska, Z. Galaz, and I. Rektorova. Speech disorders in parkinson's disease early diagnostics and effects of medication and brain stimulation. *Journal of Neural Transmission*, March 2017.

[202] L. Hartelius, H. Gustavsson, M. Astrand, and B. Holmberg. Perceptual analysis of speech in multiple system atrophy and progressive supranuclear palsy. *Journal of Medical Speech-Language Pathology*, 14 :241–248, 2006.

[203] H. Ackermann and I. Hertrich. Voice onset time in ataxic dysarthria. *Brain and Language*, 56(3) :321 – 333, 1997.

[204] S Sachin, Garima Shukla, Vinay Goyal, Shivangi Singh, Vijay Aggarwal, and Madhuri Behari. Clinical speech impairment in parkinson's disease, progressive supranuclear palsy, and multiple system atrophy. 56 :122–6, 04 2008.

[205] N. Chakraborty, T. Roy, A. Hazra, A. Biswas, and K. Bhattacharya. Dysarthric bengali speech a neurolinguistic study. *Journal of postgraduate medicine*, 54 :268–72, 10 2008.

[206] Y. Kim, R.D. Kent, J. Kent, and J.R. Duffy. Perceptual and acoustic features of dysarthria in multiple system atrophy. *Journal of Medical Speech-Language Pathology*, 18 :66–71, October 2010.

[207] S. Skodda, W. Visser, and U. Schlegel. Acoustical analysis of speech in progressive supranuclear palsy. *Journal of Voice*, 25 :725–731, November 2011.

[208] Y.E. Huh, J. Park, M.K Suh, S.E. Lee, J. Kim, Y. Jeong, H. Kim, and J.W. Cho. Differences in early speech patterns between parkinson variant of multiple system atrophy and parkinson's disease. *Brain and language*, 147 :14–20, May 2015.

[209] J. Rusz, C. Bonnet, J. Klempir, T. Tykalova, E. Baborov, M. Novotny, A. Rulseh, and E. Ruzicka. Speech disorders reflect differing pathophysiology in parkinson's disease, progressive supranuclear palsy and multiple system atrophy. *Journal of neurology*, 262 :992–1001, 2015.

[210] J. Hlavnička, R. Čmejla, T. Tykalová, K. Sonka, E. Růžička, and J. Rusz. Automated analysis of connected speech reveals early biomarkers of parkinson's disease in patients with rapid eye movement sleep behaviour disorder. *Scientific Reports*, 7, 12 2017.

[211] N. Miller, U. Nath, E. Noble, and D. Burn. Utility and accuracy of perceptual voice and speech distinctions in the diagnosis of parkinson's disease, psp and msa-p. *Neurodegenerative Disease Management*, 7 :191–203, June 2017.

[212] M. Brückl, A. Ghio, and F. Viallet. Measurement of tremor in the voices of speakers with parkinson's disease. *Procedia Computer Science*, 128 :47 – 54, 2018. 1st International Conference on Natural Language and Speech Processing.

[213] G. Li, K. Daoudi, J. Klempír, and J. Rusz. Linear classification in speech-based objective differential diagnosis of parkinsonism. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pages 5999–6003. IEEE, 2018.

[214] B. Das, K. Daoudi, J. Klempír, and J. Rusz. Towards disease-specific speech markers for differential diagnosis in parkinsonism. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 5846–5850. IEEE, 2019.

[215] R. Kowalska-Taczanowska, A. Friedman, and D. Koziorowski. Parkinson's disease or atypical parkinsonism ? the importance of acoustic voice analysis in differential diagnosis of speech disorders. *Brain and Behavior*, 10(8), 2020.

[216] T.K. Antolík and C. Fougeron. Consonant distortions in dysarthria due to parkinson's disease, amyotrophic lateral sclerosis and cerebellar ataxia. pages 2152–2156, 2013.

[217] M. Saxena, M. Behari, S. S. Kumaran, V. Goyal, and V. Narang. Assessing speech dysfunction using bold and acoustic analysis in parkinsonism. *Parkinsonism & Related Disorders*, 20(8) :855 – 861, 2014.

[218] Ian Lalich, Dale Ekbom, Sidney Starkman, Diana Orbelo, and Timothy Morgenthaler. Vocal fold motion impairment in multiple system atrophy. *The Laryngoscope*, 124, 03 2014.

[219] T. Tykalova, J. Rusz, J. Klempir, R. Cmejla, and E. Ruzicka. Distinct patterns of imprecise consonant articulation among parkinson's disease, progressive supranuclear palsy and multiple system atrophy. *Brain and language*, 165 :1–9, February 2017.

[220] J. Hlavnička, T. Tykalová, R. Čmejla, J. Klempir, E. Růžička, and J. Rusz. Dysprosody differentiate between parkinson's disease, progressive supranuclear palsy, and multiple system atrophy. pages 1844–1848, 08 2017.

[221] J. Rusz, T. Tykalová, G. Salerno, S. Bancone, J. Scarpelli, and M. Pellecchia. Distinctive speech signature in cerebellar and parkinsonian subtypes of multiple system atrophy. *Journal of Neurology*, 266, 03 2019.

[222] G. Li, K. Daoudi, J. Klempir, and J. Rusz. Linear classification in speech-based objective differential diagnosis of parkinsonism. In *IEEE-ICASSP - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, Canada, Apr 2018.

[223] I. Litvan, Y. Agid, D. Calne, G. Campbell, B. Dubois, R.C. Duvoisin, C.G. Goetz, L. Golbe, J. Grafman, J.H. Growdon, et al. Clinical research criteria for the diagnosis of progressive supranuclear palsy (steele-richardson-olszewski syndrome) report of the ninds-spsp international workshop. *Neurology*, 47 :1–9, 1996.

[224] S. Gilman, G.K. Wenning, P.A. Low, D.J. Brooks, C.J. Mathias, et al. Second consensus statement on the diagnosis of multiple system atrophy. *Neurology*, 71 :670–676, August 2008.

[225] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka. Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated parkinson's disease. *The journal of the Acoustical Society of America*, 129 :350–367, October 2011.

[226] M. Novotny, J. Rusz, R. Cmejla, and E. Ruzicka. Automatic evaluation of articulatory disorders in parkinson's disease. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 22 :1366–1378, 2014.

[227] J. Rusz, C. Bonnet, J. Klempir, T. Tykalova, E. Baborova, M. Novotny, and E. Ruzicka. supplementary material acoustic measurements for objective evaluation of motor speech disorders. February 2015.

[228] P. Boersma and D. Weenink. Praat doing phonetics by computer [computer program]. *Version 5.3.51*, 2013.

[229] M. Bruckl. Vocal tremor measurement based on autocorrelation of contours. In *INTERSPEECH*, pages 715–718, 2012.

[230] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[231] D. Duez. Acoustic analysis of occlusive weakening in parkinsonian french speech. *International Congress of Phonetic Sciences*, 08 2007.

[232] S. Sachin, G. Shukla, V. Goyal, S. Singh, V. Aggarwal, M. Behari, et al. Clinical speech impairment in parkinson's disease, progressive supranuclear palsy, and multiple system atrophy. *Neurology India*, 56 :122, 2008.

[233] H. Ackermann, S. Gräber, I. Hertrich, and I. Daum. Phonemic vowel length contrasts in cerebellar disorders. *Brain and Language*, 67(2) :95 – 109, 1999.

[234] J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, J. F. V. Bonilla, S. Skodda, J. Rusz, and E. Nöth. Voiced/unvoiced transitions in speech as a potential biomarker to detect parkinson's disease. pages 95–99, 2015.

[235] Y. Kim. Acoustic characteristics of fricatives /s/ and /S/ produced by speakers with parkinson's disease. *Clinical Archives of Communication Disorders*, 2, 01 2017.

[236] U. Goswami, S. Nirmala, Vikram C M, S. Kalita, and S. Prasanna. Analysis of articulation errors in dysarthric speech. *Journal of Psycholinguistic Research*, 49, 10 2019.

[237] A.J. Flint, S.E. Black, I. Campbell-Taylor, G.F. Gailey, and C. Levinton. Acoustic analysis in the differentiation of parkinson's disease and major depression. *Journal of Psycholinguistic Research*, 21(5) :383–399, Sep 1992.

[238] E. Fischer and A. M. Goberman. Voice onset time in parkinson disease. *Journal of Communication Disorders*, 43(1) :21 – 34, 2010.

[239] M. Lalain, A. Ghio, L. Giusti, D. Robert, C. Fredouille, and V. Woisard. Design and development of a speech intelligibility test based on pseudowords in french why

and how ? *Journal of Speech, Language, and Hearing Research*, 63(7) :2070–2083, 2020.

[240] A.S. Abramson and D.H. Whalen. Voice onset time (vot) at 50 theoretical and practical issues in measuring voicing distinctions. *Journal of Phonetics*, 63 :75 – 86, 2017.

Bibliographie personnelle

Sommaire

A . Revues internationales 217

A . I . Réseaux Bayésiens et RAP 217

A . II . Reconnaissance automatique du locuteur 217

A . III .Analyse non-linéaire de la parole 218

A . IV .Upwelling 218

A . V . Analyse multi-fractale 219

A . VI .Divers 220

B . Conférences internationales avec comité de lecture 220

B . I . Réseaux Bayésiens et RAP 220

B . II . Reconnaissance automatique du locuteur 223

B . III .Analyse non-linéaire de la parole 225

B . IV .Upwelling 227

B . V . Analyse multi-fractale 228

B . VI .Divers 229

C . Chapitres de livres 230

D . Autres publications internationales (posters, short papers) 231

E . Conférences nationales avec comité de lecture 232

A . Revues internationales

A . I . Réseaux Bayésiens et RAP

[P1] Multiscale Autoregressive Models and Wavelets

Khalid DAOUDI, Austin B. FRAKT and Alan S. WILLSKY

IEEE Transactions on Information Theory. Vol. 45. Num. 3. 1999. p.828-845.

[P2] Dynamic Bayesian Networks for Multi-Band Automatic Speech Recognition Khalid DAOUDI, Dominique FOHR and Christophe ANTOINE Computer Speech and Language. Vol 17, 2003. p.263-285.

[P3] Unsupervised Stream-Weights Computation in Classification and Segmentation Tasks

Eduardo Sanchez-Soto, Alexandros Potamianos and Khalid Daoudi

IEEE Transactions on Audio, Speech and Language Processing. Vol. 17, Issue 3, 2009. p. 436-445.

A . II . Reconnaissance automatique du locuteur

[P4] Feature Space Mahalanobis Sequence Kernels : Application to SVM Speaker Verification

Jérôme Louradour, Khalid Daoudi and Francis Bach

IEEE Transactions on Audio, Speech and Language Processing. Vol. 15, Iss. 8,. p.2465 – 2475, 2007

[P5] Large Margin GMM for discriminative speaker verification

Reda Jourani, Khalid Daoudi, Régine André-Obrecht, Driss Aboutajdine

Multimedia Tools and Applications, Springer Verlag, 2012

[P6] Discriminative speaker recognition using Large Margin GMM

R. Jourani, K. Daoudi, R. André-Obrecht, D. Aboutajdine

Neural Computing & Applications, (22)-7, pp 1329–1336, 2012

A . III . Analyse non-linéaire de la parole

[P7] An efficient solution to sparse linear prediction analysis of speech

Vahid Khanagha, Khalid Daoudi

EURASIP Journal on Audio, Speech, and Music Processing, SpringerOpen, (3), 2013.

[P8] Non-linear speech representation based on local predictability exponents

V. Khanagha, K. Daoudi, O. Pont, H. Yahia, A. Turiel

Neurocomputing, (132), p136-141, 2014

[P9] Phonetic segmentation of speech signal using local singularity analysis

Vahid Khanagha, Khalid Daoudi, Oriol Pont, Hussein Yahia

Digital Signal Processing, Elsevier, (35) p86-94, 2014

[P10] Detection of glottal closure instants based on the microcanonical multiscale formalism

Vahid Khanagha, Khalid Daoudi, Hussein Yahia

IEEE Transactions on Audio, Speech and Language Processing, (22)-12, p1941-1950, 2014

A . IV . Upwelling

[P11] An Efficient Tool for Automatic Delimitation of Moroccan Coastal Upwelling Using SST Images

Ayoub Tamim, Khalid Minaoui, Khalid Daoudi, Hussein Yahia, Abderrahman Atillah, Driss Aboutajdine

IEEE Geoscience and Remote Sensing Letters, 2014.

[P12] Detection of Moroccan Coastal Upwelling Fronts in SST Images using the Microcanonical Multiscale Formalism

Ayoub Tamim, Hussein Yahia, Khalid Daoudi, Khalid Minaoui, Abderrahman Atillah, Driss Aboutajdine, Mohammed Faouzi Smiej

Pattern Recognition Letters, Elsevier, (55), 2015

[P13] Automatic Detection of Moroccan Coastal Upwelling Zones using Sea Surface Temperature Images

Ayoub Tamim, Khalid Minaoui, Khalid Daoudi, Hussein Yahia, Abderrahman Atillah, Salma Fellah, Driss Aboutajdine, Mohamed Ansari

International Journal of Remote Sensing, Taylor & Francis, 2018.

[P14] Surface mixing and biological activity in the North-West African upwelling

Anass El Aouni, Khalid Daoudi, Hussein Yahia, Khalid Minaoui, Aïssa Benazzouz

Chaos, American Institute of Physics, 2019, 29 (1).

[P15] A Fourier approach to Lagrangian vortex detection

Anass El Aouni, Hussein Yahia, Khalid Daoudi, Khalid Minaoui

Chaos : An Interdisciplinary Journal of Nonlinear Science, American Institute of Physics, 2019, 29 (9), pp.093106

[P16] Physical and Biological Satellite Observations of the Northwest African Upwelling : Spatial Extent and Dynamics

Anass El Aouni, Véronique Garçon, Joël Sudre, Hussein Yahia, Khalid Daoudi, Khalid Minaoui

IEEE Transactions on Geoscience and Remote Sensing, 2019, pp.1-13.

[P17] Robust Detection of the North-West African Upwelling From SST Images

Anass El Aouni, Khalid Daoudi, Khalid Minaoui, Hussein Yahia

IEEE Geoscience and Remote Sensing Letters, 2020, pp.1-4.

[P18] Defining Lagrangian coherent vortices from their trajectories

Anass El Aouni, Khalid Daoudi, Hussein Yahia, Suman Kumar Maji, Khalid Minaoui

Physics of Fluids, American Institute of Physics, 2020, 32 (1).

A . V . Analyse multi-fractale

[P19] Fractal Modelling of Speech Signals

Jacques LEVY VEHEL, Khalid DAOUDI and Evelyne LUTTON

Fractals. Vol. 2. Num. 3. 1994. p.379-382

[P20] Construction of Continuous Functions with Prescribed Local Regularity
Khalid DAOUDI, Jacques LEVY VEHEL and Yves MEYER
Constructive Approximation. Vol. 14. Num. 3, 1998. p.349-386

[P21] Signal Representation and Segmentation based on Multifractal Stationarity
Khalid DAOUDI and Jacques LEVY VEHEL
Signal Processing. Vol. 82. Num. 12. 2002. p.2015-2024.

A . VI . Divers

[P22] Permuted Spectral and Permuted Spectral-Spatial CNN Models for PolSAR-Multispectral Data based Land Cover Classification
Gopal Phartiyal, Nicolas Brodu, Dharmendra Singh, Hussein Yahia, Khalid Daoudi
International Journal of Remote Sensing, Taylor & Francis, In pres

B . Conférences internationales avec comité de lecture

B . I . Réseaux Bayésiens et RAP

• "Unsupervised Stream Weight Estimation Using Anti-Models"
Eduardo Sanchez-Soto, Alexandros Potamianos and Khalid Daoudi
IEEE ICASSP'2007. Apr 15-20, 2007. Hawaii, USA.

• "Unsupervised Stream-Weight Computation in Segmentation Task : Application to Audio-Visual Speech Recognition"
Eduardo Sanchez-Soto, Alexandros Potamianos, Khalid Daoudi
IEEE International Conference on Signal Processing and Communications (ICSPC 2007), Dubaï, UAE.

• "Stream Weight Computation for Multi-Stream Classifiers"
A. Potamianos, E. Sánchez-Soto and K. Daoudi
IEEE ICASSP'2006. May 14-19, 2006. Toulouse, France.

• "Evaluation of the SPACE denoising Algorithm on Aurora2"

Christophe CERISARA and Khalid DAOUDI

IEEE ICASSP'2006. May 14-19, 2006. Toulouse, France.

• "An improved version of the SPACE algorithm for noise robust speech recognition"

Khalid DAOUDI and Christophe CERISARA

IEEE-EURASIP ISCCSP'2006. March 13-15, 2005. Marrakech, Morroco

• "The MAP-SPACE denoising algorithm for noise robust speech recognition"

Khalid DAOUDI and Christophe CERISARA

IEEE ASRU Workshop. Nov 27th-Dec 1st, 2005. San Jaun, Puerto Rico.

• "Rethinking Language Models Within the Framework of Dynamic Bayesian Networks"

Murat DEVIREN, Khalid DAOUDI and Kamel SMAILI

18th Canadian Conference on Artificial Intelligence (AI'05). Victoria, Canada, May 9-11, 2005.

• "Language Modeling using Dynamic Bayesian Networks"

Murat DEVIREN, Khalid DAOUDI and Kamel SMAILI

Int. Conference on Language Resources and Evaluation (LREC). Lisbon, Portugal, June 26-28, 2004.

• "A New Supervised-Predictive Compensation Scheme for Noisy Speech Recognition"

Khalid DAOUDI and Murat DEVIREN

EUROSPEECH. Geneva, Switzerland, Sept 1-4, 2003.

• "Frequency Filtering or Wavelet Filtering ?"

Murat DEVIREN and Khalid DAOUDI

Joint 13th Int. Conf. on Artificial Neural Networks and 10th Int. Conf. on Neural Information Processing (ICANN/ICONIP). Istanbul, Turkey, June 26-29, 2003.

• "Continuous Multi-Band Speech Recognition using Bayesian Networks : a Fast Decoding Algorithm"

Murat DEVIREN and Khalid DAOUDI

Probalistic Graphical Models Workshop (PGM'O2). Cuenca, Spain, Nov 6-8, 2002.

• "Automatic Speech Recognition : the New Millennium"

Khalid DAOUDI

International Conference on Industrial and Engineering Application of Artificial Intelligence and Expert Systems (IEA/AIE). Cairns, Australia, June 17-20th, 2002.

This paper recieved the "Best Paper Award".

• "Continuous Speech Recognition using Structural Learning of Dynamic Bayesian Networks"

Murat DEVIREN and Khalid DAOUDI

European Signal Processing Conference (EUSIPCO). Toulouse, France, September 3-6th, 2002.

• "Continuous Multi-Band Speech Recognition using Bayesian Networks"

Khalid DAOUDI, Dominique FOHR and Christophe ANTOINE

IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Trento, Italy, December 9 -13th, 2001.

• "Structural Learning of Dynamic Bayesian Networks in Speech Recognition"

Murat DEVIREN and Khalid DAOUDI

EUROSPEECH. Alborg, Denemark, September 3-7th, 2001.

• "Modeling Dependency between Regression Classes in MLLR using Multiscale Autoregressive Models"

Christophe CERISARA and Khalid DAOUDI

ISCA Workshop on Adaptation methods for speech recognition. Sophia-Antipolis, France, August 29 - 30th, 2001.

• "A Bayesian Network for Time-Frequency Speech Modeling and Recognition"

Khalid DAOUDI, Dominique FOHR and Christophe ANTOINE

International Conference on Artificial Intelligence and Soft Computing (ASC). Cancun, Mexico, May 21-24th, 2001.

• "A New Approach for Multi-Band Speech Recognition based on Probabilistic Graphical Models"

Khalid DAOUDI, Dominique FOHR and Christophe ANTOINE

International Conference on Spoken Language Processing (ICSLP). Beijing, China, October 16-20th, 2000.

B . II . Reconnaissance automatique du locuteur

• "Combination of SVM and Large Margin GMM modeling for speaker identification"
K. Daoudi, R. Jourani, R. André-Obrecht and D. Aboutajdine
EUSIPCO 2013, Marrakech, Morocco, Sept 9-13, 2013

• "Speaker Identification Using Discriminative Learning of Large Margin GMM"
K. Daoudi, R. Jourani, R. André-Obrecht and D. Aboutajdine
ICONIP 2011. Published in Neural Information Processing. Theory and Algorithms, volume 7063 of Lecture Notes in Computer Science, pp. 300–307, 2011.

• "Fast training of Large Margin diagonal Gaussian mixture models for speaker identification"
R. Jourani, K. Daoudi, R. André-Obrecht and D. Aboutajdine
6th Conference on Speech Technology and Human-Computer Dialogue (SpeD), 2011.

• "Speaker verification using Large Margin GMM discriminative training"
R. Jourani, K. Daoudi, R. André-Obrecht and D. Aboutajdine
In Proc. of International Conference on Multimedia Computing and Systems (ICMCS), pp. 1–5, 2011.

• "Speaker verification using Large Margin GMM discriminative training"
R. Jourani, K. Daoudi, R. André-Obrecht, D. Aboutajdine
International Conference on Multimedia Computing and Systems (ICMCS), Ouarzazate, Morocco. April 7-9 2011.

• "Large Margin Gaussian mixture models for speaker identification"
R. Jourani, K. Daoudi, R. André-Obrecht, D. Aboutajdine
Interspeech 2010, Makuhari, Japan, September 26-30, 2010.

• "A Comparison between Sequence Kernels for SVM Speaker Verification"
Khalid Daoudi and Jérome Louradour
IEEE ICASSP 2009, Taiepi, Taiwan, 19/04/2009-24/04/2009, p. 4241-4244, 2009.

• "State-of-the-art sequence kernels for SVM speaker verification"

Jérome Louradour and Khalid Daoudi

IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2008), Cancún, Mexico, 16/10/2008-19/10/2008, p. 498-503, 2008.

• "Pair-of-Sequences SVM Speaker Verification"

Jérome Louradour and Khalid Daoudi

EUSIPCO 2007, Poznan, Poland, p. 2370-2374, 2007.

• "A Novel Strategy for Speaker Verification based on SVM Classification of Pairs of Speech Sequences"

Khalid Daoudi and Jérôme Louradour

International Symposium on Signal Processing and its Applications (ISSPA). Feb 12-15, 2007. Sharjah, U.A.E.

• "SVM Speaker Verification using an Incomplete Cholesky Decomposition Sequence Kernel"

Jérôme Louradour, Khalid Daoudi and Francis Bach

IEEE Odyssey'2006. June 28-30, 2006. San Juan, Puerto-Rico.

• "Conceiving a new Sequence Kernel and Applying it to SVM Speaker Recognition"

Jérôme LOURADOUR and Khalid DAOUDI

Interspeech, Lisboa, Portugal. Sept 4-8, 2005.

• "SVM Speaker Verification using a new Sequence Kernel"

Jérôme LOURADOUR and Khalid DAOUDI

EUSIPCO, 2005, Antalya, Turkey.

• "Discriminative Power of Transient Frames in Speaker Recognition"

Jérôme LOURADOUR and Khalid DAOUDI

IEEE ICASSP'2005. Philadelphia, USA

• "Segmentation and Relevance Measure for Speaker Verification"

Jérôme LOURADOUR, Régine ANDRE-OBRECHT and Khalid DAOUDI

International Conference on Spoken Language Processing (ICSLP). Jeju Island, South-Korea, Oct 4-8, 2004.

B . III . Analyse non-linéaire de la parole

• Towards disease-specific speech markers for differential diagnosis in Parkinsonism
Biswajit Das, Khalid Daoudi, Jiri Klempir, Jan Rusz
IEEE ICASSP 2019 , May 2019, Brighton, United Kingdom

• Linear classification in speech-based objective differential diagnosis of parkinsonism
Gongfeng Li, Khalid Daoudi, Jiri Klempir, Jan Rusz
IEEE-ICASSP, Apr 2018, Calgary, Canada

• An analysis of psychoacoustically-inspired matching pursuit decompositions of speech signals
Khalid Daoudi, Nicolas Vinuesa
International Conference on Natural Language, Signal and Speech Processing, Dec 2017, Casablanca, Morocco

• Degree of Parkinson's Disease Severity Estimation Based on Speech Signal Processing
Zoltan Galaz, Zdenek Mzourek, Jiri Mekyska, Zdenek Smekal, Tomas Kiska, Irena Rektorova, Juan Rafael Orozco-Arroyave, Khalid Daoudi
IEEE 39th International Conference on Telecommunications and Signal Processing, Jun 2016, Vienna, Austria

• Pitch-based speech perturbation measures using a novel GCI detection algorithm : Application to pathological voice classification
Khalid Daoudi, Ashwini Jaya Kumar
INTERSPEECH'2015 - 16th Annual Conference of International Speech Communication Association, Sep 2015, Dresden, Germany

• Discrimination between pathological voice categories using matching pursuit
Ashwini Jaya Kumar, Khalid Daoudi
IEEE International Work Conference on Bioinspired Intelligence (IWOBI), Jun 2015, San Sebastian, Spain. 2015

• On classification between normal and pathological voices using the MEEI-KayPENTAX database : Issues and consequences

Khalid Daoudi, Blaise Bertrac

INTERSPEECH-2014, Sep 2014, Singapour, Singapore

• "Efficient multipulse approximation of speech excitation using the most singular manifold"

V. Khanagha, K. Daoudi

IINTERSPEECH-2012, Portland, USA, Sept 9-13, 2012

• "Improving SVF with DISTBIC for Phoneme Segmentation"

J. Winebarger, K. Daoudi, H. Yahia.

In Proc. of SPECOM'2011.

• "Reconstruction of Speech Signals from their Unpredictable Points Manifold"

V. Khanagha, H. Yahia, K. Daoudi, O. Pont, A. Turiel

NoLISP 2011 (Non-Linear Speech Processing) Conference. Published in the Springer LNAI, vol. 7015.

• "Improving text-independent phonetic segmentation based on the Microcanonical Multiscale Formalism"

V. Khanagha, K. Daoudi, O. Pont, H. Yahia

IEEE-ICASSP, Prague, Czech republic, May 22-27, 2011.

• "A novel text-independant phonetic segmentation algorithm based on the microcanonical multiscale formalism"

V. Khanagha, K. Daoudi, O. Pont, H. Yahia

Interspeech 2010, Makuhari, Japan, September 26-30, 2010.

This paper was finalist for the best student paper award.

• "Application of the microcanonical multiscale formalism to segmentation of speech signals"

V. Khanagha, K. Daoudi, O. Pont, H. Yahia

European Signal Processing Conference (EUSIPCO 2010), August 23-27, 2010, Aalborg, Denmark.

B . IV . Upwelling

• The contribution and influence of coherent mesoscale eddies off the North-West African Upwelling on the open ocean

Anass El Aouni, Khalid Daoudi, Hussein Yahia, Khalid Minaoui

SIAM Conference on Mathematics of Planet Earth (MPE18), Sep 2018, Philadelphia, USA.

• Coherent Vortex Detection from Particles Trajectories Analysis

Anass El Aouni, Khalid Daoudi, Hussein Yahia, Khalid Minaoui

2018 SIAM Conference on Nonlinear Waves and Coherent Structures, Anaheim, United States. 2018

• Surface Mixing and Biological Activity in The North African Upwelling

Anass El Aouni, Khalid Daoudi, Hussein Yahia, Khalid Minaoui

AGU Ocean Sciences Meeting 2018, Feb 2018, Portland, Oregon, United States.

• An improved method for accurate computation of coastal upwelling index using Sea Surface Temperature Images

Anass El Aouni, Khalid Minaoui, Ayoub Tamim, Khalid Daoudi, Hussein Yahia, Driss Aboutajdine

13th ACS/IEEE International Conference on Computer Systems and Applications, Nov 2016, Agadir, Morocco.

• Detection of Moroccan coastal upwelling using sea surface chlorophyll concentration

Anass El Aouni, Khalid Minaoui, Ayoub Tamim, Khalid Daoudi, Hussein Yahia, Abderrahman Atillah, Driss Aboutajdine

12th ACS/IEEE International Conference on Computer Systems and Applications AICCSA, Nov 2015, Agadir, Morocco

• Detection of Moroccan Coastal Upwelling in SST images using the Expectation-Maximization

Ayoub Tamim, Khalid Minaoui, Khalid Daoudi, Abderrahman Atillah, Driss Aboutajdine

IEEE International Conference on Intelligent Systems and Computer Vision (ISCV),

March 2015, Fez, Morocco

• On Detectability of Moroccan Coastal Upwelling in Sea Surface Temperature Satellite Images

Ayoub Tamim, Khalid Minaoui, Khalid Daoudi, Abderrahman Atillah, Driss Aboutajdine

10th International Symposium on Visual Computing, Dec 2014, Las Vegas, United States.

• Upwelling Detection in SST Images Using Fuzzy Clustering with Adaptive Cluster Merging

Ayoub Tamim, Khalid Minaoui, Khalid Daoudi, Abderrahman Atillah, Hussein Yahia, Driss Aboutajdine

The eighth edition of International Symposium on signal, Image, Video and Communications (ISIVC), Nov 2014, Marrakech, Morocco.

• A Simple Tool for Automatic Extraction of Moroccan Coastal Upwelling from Sea Surface Temperature Images

Ayoub Tamim, Khalid Minaoui, Khalid Daoudi, Abderrahman Atillah, Hussein Yahia, Driss Aboutajdine, Mohammed Faouzi Smiej

IEEE -SITA'14 - 9th International Conference on Intelligent Systems : Theories and Applications,, May 2014, Rabat, Morocco

B . V . Analyse multi-fractale

• Fast and Accurate Texture Recognition with Multilayer Convolution and Multifractal Analysis

H. Badri, K. Daoudi, H. Yahia

In Proc. ECCV, 2014

• "A New Approach for Multifractal Analysis of Turbulence Signals"

Khalid DAOUDI

Fractal'98. Valetta, Malta. October 25-28, 1998.

• "Statistical Analysis and Representation of TCP Traffic"

Khalid DAOUDI and Jacques LEVY VEHEL

Wavelet Applications in Signal and Image Processing (part of SPIE Annual Meeting). San Diego, USA, July 20-24th, 1998.

• "Multifractal Representation of Turbulence Signals : A Wavelet Based Approach"

Khalid DAOUDI

International Wavelets Conference. Tangier, Morocco, April 13-17th, 1998.

• "Weak Self-Similarity in Speech Signals"

Khalid DAOUDI

International Conference on Signal and Image Processing (SIP). New Orleans, USA, December 4-7th, 1997.

• "Generalized IFS for Signal Processing"

Khalid DAOUDI and Jacques LEVY VEHEL

IEEE Digital Signal Processing Workshop (IEEE'DSP). Loen, Norway, September 1-4th, 1996.

• "Speech Modeling Based on Local Regularity Analysis"

Khalid DAOUDI and Jacques LEVY VEHEL

International Conference on Signal and Image Processing (SIP). Las Vegas, USA, November 20-23th, 1995.

B . VI . Divers

• "A Robust Approach for Multivariate Binary Vectors Clustering and Feature Selection"

M. Al Mashrgy, N. Bouguila, K. Daoudi

ICONIP 2011. Published in Neural Information Processing. Theory and Algorithms, volume 7063 of Lecture Notes in Computer Science, 2011.

• "Learning Concepts from Visual Scenes Using a Binary Probabilistic Model"

Nizar Bouguila and Khalid Daoudi

IEEE International Workshop on Multimedia Signal Processing (MMSP 2009), Rio de Janeiro, Brazil, 2009.

• "A Statistical Approach for Binary Vectors Modeling and Clustering"

Nizar Bouguila and Khalid Daoudi

Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2009), Bangkok, Thailand, 2009.

• "Automatic target recognition of aircraft models based on ISAR images"

Mohamed Nabil Saidi, Khalid Daoudi, Ali Khenchaf, Brigitte Hoeltzener, Driss Aboutajdine

IEEE International Geoscience And Remote Sensing Symposium (IGARSS 2009), Cap Town, South Africa, 2009.

• "Cleaning Statistical Language Models"

Reda Jourani, David Langlois, Kamel Smaïli, Khalid Daoudi, Driss Aboutajdine

International Conference on Information Systems and Economic Intelligence (SIIE'2010), Sousse, Tunisia, 2010.

• "Building Arabic textual corpus from the Web"

Reda Jourani, Kamel Smaïli, Driss Aboutajdine, Khalid Daoudi

International Conference on Information Systems and Economic Intelligence (SIIE'2008), Hammamet, Tunisia, 2008.

C . Chapitres de livres

• Iterated Function Systems and some generalizations : Local Regularity Analysis and Multifractal Modeling of Signals

Khalid DAOUDI

In Scaling Laws, Fractals and Wavelets. Patrice Abry, Paulo Gonçalves, Jacques Lévy Véhel (Eds.), ISTE - WILEY, p. 301-332, january 2009.

• Continuous Speech Recognition Using Dynamic Bayesian Networks : A Fast Decoding Algorithm

Murat DEVIREN and Khalid DAOUDI
Chapter in the book "Advances in Bayesian Networks". José A. Gámez, Serafin Moral, Antonio Salmerón (Eds.), Springer, p. 289-308, Vol. 146, Studies in Fuzziness and Soft Computing, 2004.

• Généralisations des systèmes de fonctions itérées : Analyse de la régularité locale et modélisation multifractale des signaux
Khalid DAOUDI
Chapter (in French) in the book "Loi d'échelle, Fractales et Ondelettes". Traité IC2, Hermes Editions, 2002.

D . Autres publications internationales (posters, short papers)

• North-West African Upwelling dynamics from physical and biological satellite observations
Anass El Aouni, Khalid Minaoui, Khalid Daoudi, Hussein Yahia
Poster communication at the 4th GEO Blue Planet Symposium, Jul 2018, Toulouse, France

• Objective discrimination between Progressive Supranuclear Palsy and Multiple System Atrophy using speech analysis
Khalid Daoudi, Nicolas Brodu, Jan Rusz, Jiri Klempir
Short paper at the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE-EMBC'16), Aug 2016, Orlando, United States

• "Dynamic Bayesian Networks for Robust Automatic Speech Recognition"
Khalid DAOUDI
Poster presentation at the International Meeting on Bayesian Statistics (Valencia 7). Tenerife, Spain, June 2 - 6th, 2002.

E . Conférences nationales avec comité de lecture

• Une nouvelle approche non-linéaire pour la segmentation phonétique

Vahid Khanagha, Oriol Pont, Khalid Daoudi, Hussein Yahia

GRETSI 2011 - XXIIIe Colloque GRETSI, Sep 2011, Bordeaux, France

• Apprentissage discriminant des GMM à grande marge pour la vérification automatique du locuteur

Reda Jourani, Khalid Daoudi, Régine André-Obrecht, Driss Aboutajdine

GRETSI 2011 - XXIIIe Colloque GRETSI, Sep 2011, Bordeaux, France

• Une nouvelle architecture de compensation du bruit pour la reconnaissance robuste de la parole

Khalid Daoudi, Murat Deviren

XXVes Journées d'Etudes sur la Parole - JEP-TALN-RECITAL 2004, Fès, Maroc.

• Une nouvelle approche de modélisation du langage par des réseaux Bayésiens dynamiques

Khalid Daoudi, Murat Deviren

XXVes Journées d'Etudes sur la Parole - JEP-TALN-RECITAL 2004, Fès, Maroc.

• Apprentissage de structures de réseaux bayésiens dynamiques pour la reconnaissance de la parole

Murat Deviren, Khalid Daoudi

XXIVèmes Journées d'Études sur la Parole - JEP'2002, Jun 2002, Nancy, France, pp.293-296

• Réseaux Bayésiens Dynamiques pour la Reconnaissance Multi-Bandes de la Parole

Murat Deviren, Khalid Daoudi

XXIVèmes Journées d'Études sur la Parole - JEP'2002, Jun 2002, Nancy, France, pp.293-296