UNIVERSITÉ DE SHERBROOKE
Faculté de génie
Département de génie électrique et de génie informatique

IMPACT PERCEPTUEL D'UNE MISE À ZÉRO DES SEGMENTS PLOSIFS DE PAROLE

Mémoire de maîtrise
Spécialité : génie électrique

Vincent SANTINI

Jury : Roch LEFEBVRE (directeur)
Philippe GOURNAY (co-directeur)
Philippe MABILLEAU
Paul CHARRETTE (rapporteur)

Sherbrooke (Québec) Canada Juillet 2016

RÉSUMÉ

En traitement du signal audio, les plosives sont des sons de parole très importants au regard de l'intelligibilité et de la qualité. Les plosives sont cependant difficiles à modéliser à l'aide des techniques usuelles (prédiction linéaire et codage par transformée), à cause de leur dynamique propre importante et à cause de leur nature non prédictible.

Cette étude présente un exemple de système complet capable de détecter, segmenter, et altérer les plosives dans un flux de parole. Ce système est utilisé afin de vérifier la validité de l'hypothèse suivante : La phase d'*éclatement* (de *burst*) des plosives peut être mise à zéro, de façon perceptuellement équivalente.

L'impact sur la qualité subjective de cette transformation est évalué sur une banque de phrases enregistrées. Les résultats de cette altération hautement destructive des signaux tendent à montrer que l'impact perceptuel est mineur. Les implications de ces résultats pour le codage de la parole sont abordées.

Mots-clés : Plosive, transitoire, éclatement, fermeture, segmentation, modélisation

REMERCIEMENTS

Ce projet de maitrise a vu le jour grâce à une initiative de Roch Lefebvre, vice doyen à la faculté de génie électrique et informatique.

Un grand merci à Philippe Gournay, qui a toujours été disponible durant ces années de maîtrise.

LEXIQUE

- **Burst** : Voir éclatement.
- **Coarticulation** : Régime transitoire entre deux phonèmes causée par l'inertie du conduit vocal. Présumé non audible, mais primordial à l'intelligibilité.
- **Consonne affriquée (affricate)** : Consonne composée d'une phase occlusive suivie d'un relâchement fricatif. (Exemple : /j/ en anglais prononcé comme /d/+/j/)
- **Consonne battue (flap)** : Similaire à une occlusive, mais sans pression d'air lors de la production
- **Consonne occlusive (stop)** : Consonne faisant intervenir un blocage complet de l'écoulement d'air au niveau de la bouche, suivi d'un relâchement soudain de celui-ci.
- **Consonne sonnante** : Consonne ne faisant pas intervenir d'écoulement turbulent
- **Cordes vocale** : Membranes muqueuses du larynx dont la vibration le voisement.
- **Éclatement** : C'est le moment où la plosive est jouée, où l'écoulement de l'air est relâché.
- **Fermeture (closure)** : C'est la période initiale de la plosive durant laquelle l'écoulement de l'air est bloqué.
- **Fricative** : Consonne produite par resserrement de la bouche, du pharynx ou de la glotte sans qu'il y ait fermeture complète de ceux-ci.
- **Hauteur** (pour la voix) : pitch en anglais. Sensation auditive liée à la vitesse de vibration des cordes vocales (son grave / son aigu) – Fréquence fondamentale, en Hz
- **Intensité** : Sensation auditive liée au débit d'air (son fort / faible) – Amplitude en dB
- **Phonation** : Production des phones ou des sons.
- **Phonème** : Plus petite unité discrète isolable par segmentation dans une langue parlée. Chaque phonème est identifié par un symbole. (Exemple : /p/ dans **p**ère)
- **Phones** : Différentes réalisations d'un phonème.
- **Phonétique** : Etude des sons utilisés dans la communication verbale.
- **Plosive** : Consonne occlusive
- **Sonnante** : Son de parole produite avec un écoulement d'air continu et non turbulent
- **Timbre** : Sensation auditive liée à la qualité du son (son sombre / son clair). Nombre et intensité relative des composantes aigus et graves dans le spectre. Donne la couleur à la voix, critère d'identification du locuteur.
- **Voice-onset time (VOT)** : Paramètre des plosives, délai entre le relâchement et le début du voisement.
- **Voisement** : Un son est dit voisé si sa production nécessite une vibration des cordes vocales.

ACRONYMES ET ABREVIATIONS

- **ACR :** Absolute Category Rating = Test audio subjectif par note absolue
- **CBT :** Closure Burst Transition = Instant de transition fermeture-relâchement
- **CCR :** Comparison Category Rating = Méthode d'évaluation par catégories de comparaison
- **CMOS :** Comparison Mean Opinion Score = Note moyenne d'appréciation par comparaison
- **DCT :** Discrete Cosine Transform = Transformée en cosinus discrète
- **DWT :** Discrete Wavelet Transform = Transformée en ondelettes discrète
- **eaQHM :** Extended adaptative Quasi-Harmonic model
- **ESPRIT :** Estimation of Signal Parameters via Rotational Invariant Techniques = Algorithme d'estimation spectrale
- **FChT :** Fan-Chirp Transform
- **GCI :** Glottal Closure Instant = Instants d'excitation significatifs du conduit vocal
- **HNR :** Harmonic-to-Noise Ratio = Degré de périodicité
- **LPC :** Linear Predictive Coding = Codage par prédiction linéaire
- **MOS :** Mean Opinion Score = Note moyenne d'appréciation
- **MUSIC :** MUltiple SIgnal Classification = autre algorithme d'estimation spectrale
- **NMSE** : Normalized Mean Square Error
- **PFE :** Position of First Extrema = Position dans la trame du premier extrema
- **PI :** Plosion Index = Mesure temporelle de détection (article)
- **PLE** : Position of Last Extrema = Position dans la trame du dernier extrema
- **QHM :** Quasi Harmonic Model
- **RSB** : Rapport signal à bruit (SNR en anglais)
- **SRER** : Signal to Reconstruction Error Ratio
- **SVM** : Support Vector Machine
- **TIMIT :** Corpus phonétique (Entre autre créé par TI + MIT)
- **VOT** : Voice-onset time
- **ZCR (ZC)** : Zero Crossing Rate = Taux de passage par zéro

NOTATIONS

- **/p/ :** représente un phonème
- **[r] :** représente un allomorphe

Exemple : [r] correspond au /r/ roulé et [ʁ] correspond au /r/ non roulé : ce sont des allophones du phonème /r/

TABLE DES MATIÈRES

RÉSUMÉ i
REMERCIEMENTS iii
LEXIQUE v
ACRONYMES ET ABREVIATIONS vii
NOTATIONS vii
TABLE DES MATIÈRES ix
LISTE DES TABLEAUX xiii
LISTE DES FIGURES xv
CHAPITRE 1 : INTRODUCTION 1
1.1 Mise en contexte et problématique 1
1.2 Définition du projet de recherche 2
1.3 Objectifs du projet de recherche 3
1.4 Méthodologie 4
1.4.1 Matériel 4
1.4.2 Choix du modèle 5
1.5 Structure du document 7
CHAPITRE 2 : ÉTAT DE L'ART 9
2.1 Caractérisations des sons humains et des plosives 9
2.1.1 Principaux articulateurs et production de la voix 9
2.1.2 Articulation des plosives 10
2.1.3 Acoustique des plosives 11
2.1.4 Plosives voisées et non-voisées 12
2.1.5 Détails sur l'éclatement des plosives 13
2.1.6 Quelques statistiques sur les plosives 13
2.2 Détection des plosives 15
2.2.1 Indice de plosion 15
2.2.2 Autres méthodes 16
2.3 Segmentation des plosives 17
2.3.1 Mesure de la durée de l'éclatement grâce à l'indice de plosion 17
2.3.2 Segmentation de l'aspiration grâce au taux de passage par zéro (ZCR) 17
2.3.3 Pente spectrale 18
2.3.4 Transitions entre classes phonétiques 20

2.4 Classification des plosives 21
2.4.1 Classification par caractéristiques acoustiques et phonétiques 21
2.4.2 Détection et classification basée sur un découpage phonémique 22
2.4.3 Détection des plosives aspirées (Indo-Aryen) 23
2.5 Modélisations des plosives 24
2.5.1 Modèles sinusoïdaux 24
2.5.2 Modèles sinusoïdaux atténués 24
2.5.3 Modèles sinusoïdaux adaptatifs 25
2.5.4 Approximation par transformée en ondelettes discrète (DWT) 25
2.6 Codage et psycho-acoustique 27
2.6.1 Grands principes 27
2.6.2 Compression psycho-acoustique 27
2.7 Plosives et perception auditive 28
2.7.1 Impact du bruit précédant les plosives 28
2.7.2 Impact de la durée du silence et de l'amplitude 28
2.7.3 Manipulation de la bande de fréquence des plosives 28
2.8 Tests perceptuel 29
2.8.1 Tests objectifs 29
2.8.2 Tests subjectifs 31
CHAPITRE 3 : RELEVANCE OF THE BURST PHASE OF STOP CONSONANTS 33
3.1 Abstract 36
3.2 Introduction 37
3.3 Characterization of stop consonants 38
3.4 Overview of the alteration system 39
3.5 Detection of the closure-burst transition 39
3.5.1 Original CBT detection algorithm using the plosion index 40
3.5.2 Modified CBT detection algorithm 41
3.6 Burst segmentation of stop consonants 41
3.7 Muting of the burst phase of stop consonants 42
3.8 Test Material 44
3.8.1 The TIMIT Acoustic-Phonetic Continuous Speech Corpus 44
3.8.2 The test sub-corpus 45
3.9 Results 46
3.9.1 Objective evaluation 46

3.9.2 Subjective evaluation 47
3.9.3 Perspectives for speech coding 48
3.10 Conclusion 50
3.10.1 Summary 50
3.10.2 Possible future work 50
3.11 Acknowledgment 50
CHAPITRE 4 : CONCLUSION 51
4.1 Sommaire 51
4.2 Contributions 52
4.3 Travaux futurs 52
BIBLIOGRAPHIE 53
ANNEXES 57
Annexe A : Procédure de test CCR et implémentation 57
Directives du test 57
Interface du programme de test subjectif 58
Structure des fichiers test 58
Structure des fichiers résultats 59
Annexe B : Détails sur les résultats 60
Résultats individuels des auditeurs 60

LISTE DES TABLEAUX

Tableau 2.1 Totaux des types de sons dans TIMIT en lien avec les plosives 13
Table 3.1 Sub-corpus test items distribution 45

LISTE DES FIGURES

Figure 2.1 Schéma de l'appareil respiratoire humain 9
Figure 2.2 Schéma des lieux d'articulation 10
Figure 2.3 Phases acoustiques des plosives 11
Figure 2.4 Schématisation des différents VOT 12
Figure 2.5 Durées minimales et maximales de la phase de fermeture (en ms) 14
Figure 2.6 Durées minimales et maximales de la phase de relâchement (en ms) 14
Figure 2.7 Détection grâce à l'indice de plosion 16
Figure 2.8 Segmentation de l'éclatement grâce à l'indice de plosion 17
Figure 2.9 Segmentation grâce au ZCR 18
Figure 2.10 Transitions entre classes, segmentation réussie (/p/ /u/) 20
Figure 2.11 Transitions entre classes, segmentation de l'aspiration échouée (/t/ /a/) 20
Figure 3.1 Phase segmentation of a stop consonant 38
Figure 3.2 The proposed alteration system 39
Figure 3.3 Detection of a stop consonant release 40
Figure 3.4 CBT modification 41
Figure 3.5 Burst segmentation of a stop consonant 42
Figure 3.6 Muting mask 43
Figure 3.7 Illustration of the muting of the burst phase of a stop consonant 44
Figure 3.8 Distribution of the muted stop consonants within the sub-corpus 45
Figure 3.9 Mean duration of the burst phase of each stop consonant 46
Figure 3.10 CMOS results for the sub-corpus test items (with stop consonants muting) 47
Figure 3.11 CMOS results for the control pairs 47

CHAPITRE 1 : INTRODUCTION

1.1 Mise en contexte et problématique

Le traitement du signal est un domaine scientifique vaste qui est utilisé notamment dans les supports de média ou de communication. Le traitement du son est l'une de ses branches les plus importantes, dont les applications sont entre autres la communication téléphonique, et les opérations de codage et décodage des fichiers audio.

Les objectifs en traitement du signal audio sont multiples : amélioration de la qualité, amélioration de l'intelligibilité, compression des données, extraction de caractéristiques pour l'analyse et la reconnaissance, … Le projet de cette maîtrise se situe dans ces horizons. En compression du signal audio par exemple, afin de trouver le meilleur équilibre entre la qualité du son et de la quantité de données nécessaires lors d'une communication, on distingue généralement les signaux de parole des autres types de sons (bruits, musique, …).

Cette maîtrise s'inscrit dans le cadre d'un projet de modélisation des différents signaux de parole du GRPA (Groupe de Recherche sur la Parole et l'Audio) de l'Université de Sherbrooke sous la direction de Roch Lefebvre, et sous la codirection de Philippe Gournay. Le sujet du projet de recherche concerne le domaine du traitement de la parole.

En traitement de la parole, on classe en trois grandes catégories les sons produits par l'homme :
- ➔ Les sons voisés, qui font intervenir une vibration des cordes vocales (exemple : voyelle /u/)
- ➔ Les fricatives, qui ne présentent pas cette vibration (exemple : /f/)
- ➔ Les sons transitoires, qui sont caractérisés par une grande énergie en un court intervalle de temps (exemples : coarticulations /ao/, **plosives** /p/)

Les segments voisés peuvent être efficacement codés à l'aide d'une transformée sinusoïdale, ou en combinant une excitation périodique et un filtre à prédiction linéaire (LPC). Les segments fricatifs sont bien modélisés par des processus générateurs de bruits associés à une mise en forme spectrale et un filtre LPC. Enfin, il existe également des approches hybrides, comme dans le codeur CELP, où aucune décision sur la nature du son n'est prise. Des outils permettent de coder les aspects périodiques et stochastiques du signal.

Cependant, les codeurs utilisés dans le cadre d'une communication téléphonique ne sont en général pas efficaces pour traiter les **sons transitoires**. Certains codeurs possèdent des modes propres pour traiter les sons transitoires (VMR, et certains G.7xx tel G.722), mais utilisent des méthodes générales en augmentant la résolution temporelle, et non pas une modélisation pertinente adaptée à la nature de ces sons.

Le but de cette maîtrise est donc de parfaire notre connaissance d'un certain type de sons transitoires : les plosives (/p/, /t /, /k/, /d/, /b/, /g/) dans l'optique d'améliorer la qualité de leur reconstruction après codage et décodage, et également de minimiser la quantité de données pour les représenter. La finalité du projet de recherche est de proposer un mode de traitement spécifique pour ces sons, qui pourra ensuite être implémenté dans un codeur de parole.

1.2 Définition du projet de recherche

Au travers d'un réseau cellulaire, des modélisations complexes sont utilisées afin de fournir une qualité sonore satisfaisante. Suivant le type de réseau de communication impliqué, et suivant la stabilité de la transmission, un débit binaire maximal est imposé. Il faut donc choisir la modélisation qui présente le meilleur compromis qualité/débit, tout en pouvant s'adapter aux contraintes techniques du réseau.

Certain sons produits par l'homme sont plus délicats à modéliser : les transitoires, dont les plosives forment un sous ensemble. Ces sons sont fortement dynamiques et possèdent une nature imprédictible. En effet, l'enregistrement successif de plusieurs plosives articulées de la même manière produira une forme d'onde singulièrement différente, alors qu'elle sera perçue de la même manière par l'appareil auditif humain.

Le but du projet de recherche est donc de proposer une alternative à la modélisation des plosives basée sur leur aspect perceptuel qui n'altère pas leur qualité subjective – donc sans forcément s'intéresser à toutes les formes d'ondes qui peuvent les représenter.

Question de recherche :

« Quel est l'impact perceptuel d'une mise à zéro des trames transitoires des plosives dans un signal de parole propre ? »

1.3 Objectifs du projet de recherche

L'objectif général est le suivant :

Proposer et évaluer un modèle perceptuel des sons plosifs dans lequel les phases d'éclatement (burst) sont représentées par des segments à zéro. Sur le plan de la perception, l'application de ce modèle (ou de cette transformation) ne devra pas altérer la perception des sons plosifs. Sur le plan de la compression, l'application de ce modèle devra permettre une forte compression.

Les objectifs spécifiques sont les suivants :

1. Proposer une méthode de détection et une méthode de segmentation des sons plosifs, permettant ainsi de rendre la transformation (l'application du modèle) automatique.
2. Réaliser une évaluation perceptuelle permettant de mesurer l'impact d'une telle transformation
3. Envisager l'intégration d'une telle transformation au sein du codage de la parole

1.4 Méthodologie

Cette partie expose les outils et les différentes étapes qui ont menés à définir puis réaliser les objectifs exposés dans la partie précédente.

1.4.1 Matériel

Afin de conduire ce projet de recherche, plusieurs outils étaient nécessaires :

- Une banque d'enregistrements audio de phrases, possédant un étiquetage au niveau phonémique (donc une segmentation des plosives), permettant d'effectuer des transformations automatisées
- Un code structuré en plusieurs modules permettant d'automatiser les différentes opérations sur ces phrases
- Une norme de test perceptuel subjectif, et son implémentation logicielle afin de réaliser des tests formels.

Les trois sous parties qui suivent détaillent les raisons des choix effectués pour ces différents outils.

1.4.1.1 Banque de phrases

Pour la banque de phrases, l'attention c'est très rapidement portée sur le corpus TIMIT, contenant plusieurs milliers de phrases différentes prononcées par près de 630 locuteurs hommes et femmes de différentes régions des États-Unis. Le premier avantage de ce corpus est son utilisation répandue dans le milieu du traitement de la parole, on retrouve plusieurs études statistiques sur cette banque. Le second avantage est la documentation en ligne faisant état d'une segmentation à plusieurs niveaux : phrase, mot, et phonème, tout en proposant de découper les plosives en trois parties : la fermeture, le relâchement, et l'aspiration. Un tel découpage permettant de réaliser des tests informels préliminaires sans besoin de se soucier d'une méthode de détection et de segmentation automatique, ou d'avoir besoin de réaliser l'opération à la main. Ce choix s'avérera judicieux concernant la qualité du corpus, mais révèlera quelques faiblesses au niveau du découpage phonémique (voir 3.8.1).

1.4.1.2 Structure du code

Concernant la structure de code, le choix s'est immédiatement porté sur l'environnement Matlab, offrant des outils d'analyse et de présentation très poussés. Le code produit s'articule autour de plusieurs module, et s'est tout d'abord concentré sur l'exploitation de la banque de phrase TIMIT, et de ses métadonnées (étiquettes, segments, type de locuteur, …). Ensuite, les tests informels décrits ci-après (1.4.2) ont été intégrés dans un module indépendant. Enfin, les étapes de détection et de segmentation ont elles aussi été rajoutées comme des modules indépendants.

Cet outil a donc servi à exploiter les données, leur appliquer les transformations, produire diverses figures présentes dans ce mémoire, et enfin à créer et diffuser le test perceptuel choisi.

1.4.1.3 Test perceptuel

Afin d'évaluer l'impact de la transformation réalisée sur les plosives, un test perceptuel audio est nécessaire. Comme détaillé ci-après dans la section 2.8, il existe de nombreux tests, principalement répartis en deux catégories : les tests objectifs, et les tests subjectifs. Les tests objectifs consistent en un algorithme qui évalue le signal suivant divers critères, dont certains en lien avec la perception humaine. Les tests subjectifs consistent à analyser les votes produits par des auditeurs selon des règles établies.

Dans le cadre du projet de recherche, la transformation appliquée n'affecte qu'une infime partie d'un signal de parole usuel (environ 2% du signal utile de parole, comme présente en section 3.9.3). Le choix c'est donc porté sur un test subjectif, durant lequel des auditeurs pourraient comparer le signal avant et après transformation, sachant au préalable que la différence se ferait sur les segments plosifs. Le test CCR (Comparison Category Rating) respectant ce cahier des charges, il a été choisi.

Au sein du GRPA, le logiciel de test audio subjectif usuellement utilisé ne permettait pas d'effectuer un tel test, et aucun code source n'a été trouvé pour remplir cette fonction. Le codage du logiciel capable de faire passer le test aux auditeurs, et de dépouiller les résultats a donc dû être effectué dans le cadre du projet de recherche. Ce logiciel a été entièrement codé sous Matlab, ceci présentant l'avantage de s'intégrer facilement au reste du projet. Les détails des directives du test et de son déroulement sont fournis en Annexe A : Procédure de test CCR.

1.4.2 Choix du modèle

Au tout début du projet de recherche, une recherche préliminaire a été effectuée sur une modélisation perceptuellement transparente des segments plosifs de parole. Plusieurs transformations ont ainsi été envisagées et comparées à l'aide de tests informels, avant d'aboutir au projet de recherche présenté dans ce mémoire sur la mise à zéro des segments plosifs de parole.

Dans un premier temps, en prenant du recul par rapport aux divers outils disponibles en traitement du signal, une idée a été explorée : malgré certains invariants permettant d'identifier les différentes plosives (voir 2.4), il apparait que la prononciation d'une plosive peut générer une multitude de formes d'ondes d'attaques significativement différentes. Une première hypothèse a ainsi été proposée : « L'identité d'un locuteur n'est pas présente dans les segments plosifs de parole qu'il produit » Autrement dit, est-il possible de remplacer chacun des segments plosifs dans un signal de parole par une forme d'onde générique, sans en altérer la perception ? Des tests informels ont ainsi été réalisés, consistant à substituer dans un signal de parole chaque occurrence de plosive par une plosive générique (en utilisant un dictionnaire contenant une plosive générique pour chacune des six plosives possibles). Ces tests consistaient à remplacer la totalité d'une plosive par son homologue générique (c'est-à-dire les phases d'éclatement et les phases d'aspirations). Il est rapidement apparu à l'aide de tests informels que cette hypothèse était invalidée principalement à cause du segment d'aspiration, qui contenait trop d'information du locuteur pour être simplement remplacé.

Le projet de recherche étant à la base ciblé sur les sons transitoires, une seconde hypothèse plus ciblée que la première fut explorée : « L'identité d'un locuteur n'est pas présente dans les parties

transitoires (les *éclatements*) des segments plosifs de parole qu'il produit ». Les tests informels réalisés pour cette hypothèse étaient identiques à ceux de l'hypothèse précédente, mais consistaient uniquement à remplacer les *éclatements* de chaque plosive dans un signal de parole par un *éclatement* générique (en utilisant un dictionnaire contenant un *éclatement* pour chacune des six plosives possibles obtenu en les sélectionnant à la main sur des enregistrements de la banque de sons TIMIT). Les résultats furent encourageants, mais des artefacts étaient fréquemment produits à cause de la mise au bon niveau énergétique du segment d'éclatement générique, pour correspondre au segment remplacé. Les problèmes rencontrés à haut niveau énergétique ont ainsi conduit à baisser significativement ce niveau, avant de mener à la troisième hypothèse.

Tirant parti des inconvénients de la précédente hypothèse, une troisième hypothèse fut explorée : « Il est possible de remplacer les segments d'éclatement des plosives par du bruit à faible énergie sans dégrader la qualité subjective de la parole ». Le principe consistait à remplacer les segments d'éclatement des plosives d'un signal de parole par différents types de bruits à faible énergie. Des tests informels ont étés réalisés pour des bruits blancs ou colorés, et les résultats furent encore meilleurs que ceux de l'hypothèse précédente. Cependant, certains auditeurs arrivaient à percevoir le profil du bruit utilisé pour le remplacement. D'autres tests informels ont été conduits à des niveaux énergétiques encore plus faible, et donnèrent de meilleurs résultats, qui menèrent à explorer une dernière hypothèse.

Au regard des meilleurs résultats obtenus à très faible niveau énergétique, une quatrième hypothèse fut explorée : « Il est possible de remplacer les segments d'éclatement des plosives par du silence sans dégrader la qualité subjective de la parole ». Le principe consistait à remplacer les segments d'éclatement des plosives d'un signal de parole par des zéros. Des tests informels ont également été menés, et leurs résultats surpassant ceux de toutes les autres hypothèses explorées conduisirent à réaliser un test subjectif formel avec des auditeurs experts. Ce sont le développement et les résultats de cette hypothèse qui sont présentés dans ce mémoire.

1.5 Structure du document

Afin de répondre à la question de recherche, ce mémoire par article est divisé en quatre chapitres. Le premier chapitre présente une introduction au projet de recherche, et définit les objectifs et la méthodologie du projet.

Le second chapitre présente une revue de littérature en lien avec le projet de recherche. Les principaux sujets abordés concernent les différents sons présents dans la parole et l'évaluation de leur qualité, ainsi que de nombreux outils de modélisation et d'analyse statistique et perceptuelle en lien avec les plosives.

Le troisième chapitre présente en détails une réponse à la question de recherche, sous la forme d'un article de recherche soumis et accepté à la conférence MMSP 2016 (IEEE Workshop on Multimedia Signal Processing). On y retrouve une présentation détaillée du système choisi pour automatiser une transformation sur les plosives, et les résultats de l'évaluation perceptuelle d'une telle transformation.

Enfin, le quatrième chapitre présente les conclusions du projet de recherche, met en valeur les contributions apportées, et propose des pistes de travaux futurs.

CHAPITRE 2 : ÉTAT DE L'ART

2.1 Caractérisations des sons humains et des plosives

2.1.1 Principaux articulateurs et production de la voix

La production de la voix résulte de l'action coordonnée de plusieurs acteurs, mais elle peut être résumée en deux phases : la production d'un souffle, et l'altération dynamique de ce souffle. La production de la voix est assurée par les poumons qui produisent un souffle d'air, et surtout par le larynx qui module ce souffle et produit donc une vibration (voir Figure 2.1). De nombreux muscles et organes permettent au locuteur de contrôler la hauteur, l'intensité, la qualité, et le timbre de la voix. L'articulation des sons est quant à elle effectuée par plusieurs organes tels que la langue, la mâchoire inférieure, les lèvres, et le palais. Ces organes influent sur la résonnance et la modulation de l'air, et peuvent produire des sons additionnels pour les consonnes [1].

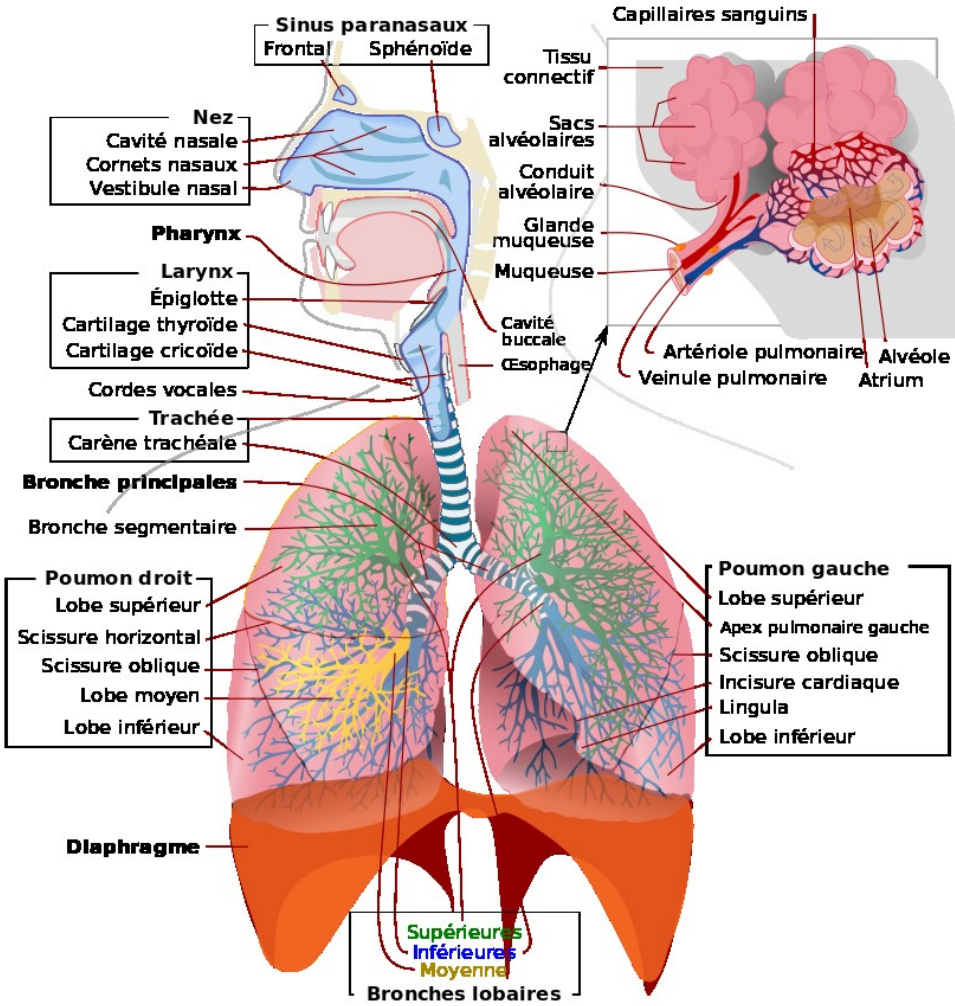


Figure 2.1 Schéma de l'appareil respiratoire humain
(Source : LadyofHats traduction by Berru - Wikipédia)

2.1.2 Articulation des plosives

La production d'une plosive s'effectue en trois temps. Ce processus commence par le blocage de l'écoulement d'air dans le conduit vocal, ce qui génère une accumulation de pression dans la bouche derrière le point de blocage, puis se termine par un relâchement soudain de celle-ci [2] [3]. Cette production peut ainsi varier selon trois critères principaux, conduisant aux différentes plosives possibles. Tout d'abord, une vibration des cordes vocales peut intervenir durant l'articulation. On distingue ainsi les plosives dites voisées qui présentent cette vibration durant l'articulation (/b/, /d/ et /g/), des plosives dites non-voisées, qui ne présentent pas cette vibration (/p/, /t/ et /k/). De plus amples détails sont fournis en section 2.1.4. Ensuite, le blocage peut être réalisé par différents articulateurs spécifiques : les lèvres, la pointe de la langue, et le dos de la langue. Enfin, on distingue également le lieu d'articulation des plosives :

- Bilabiale (n°1 et 2 sur Figure 2.2) : au niveau des lèvres (/b/ et /p/).
- Alvéolaire (n°4 sur Figure 2.2) : au niveau des alvéoles des dents de la mâchoire supérieure (/t/ et /d/).
- Vélaire (n°8 sur Figure 2.2) : au niveau de l'arrière du palais (/k/ et /g/).

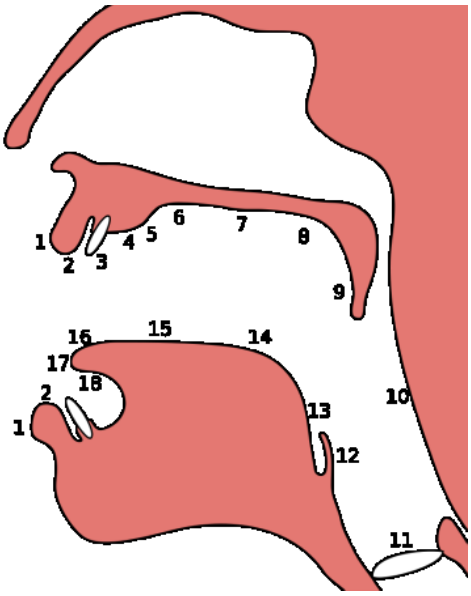


Figure 2.2 Schéma des lieux d'articulation
Licence : Sagittal Section w/ Points of Articulation. Sagittal section image based on Minifie et al. (1973); articulation places are from Catford (1977). Author: Ishwar; svg by Rohieb. WikiCmns; CC 2.5 Generic.

Dans ses travaux sur la modélisation des plosives, Stevens [2] [3] a réalisé des captures par rayons X successives de la zone de production de la voix humaine de manière à pouvoir être visionnées comme un film. Des séquences VCV ont été enregistrées (telles que /a//k//a/), afin de mettre en lumière les différents articulateurs qui entrent en jeux lors de leur production.

2.1.3 Acoustique des plosives

La production d'une plosive génère de nombreux évènements (ou phases) acoustiques. Dans la littérature, ces évènements sont parfois considérés indépendamment, parfois regroupés. Ces phases peuvent varier en durée et même être absentes suivant la langue. Voici une analyse de ces évènements découpés en 5 phases acoustiques (voir Figure 2.3) :

- Occlusion : Blocage du conduit vocal et donc installation d'une période de silence, avec un éventuel bruit de fond et/ou de voisement sourd pour les plosives voisées
- Éclatement : Violent relâchement du flot d'air caractérisé par un pic d'énergie, qui dure en général moins de 10 ms
- Frication turbulente : Résulte de l'écoulement turbulent de l'air lors de l'éclatement et du déclin de ce régime transitoire
- Aspiration : Après l'écoulement turbulent, l'air circule au fur et à mesure que le conduit vocal se débloque, provoquant un bruit d'aspiration.
- Coarticulation de transition : Cette phase n'est pas forcément considérée comme partie intégrante de la plosive, et est surtout remarquable si le prochain phonème est une voyelle.

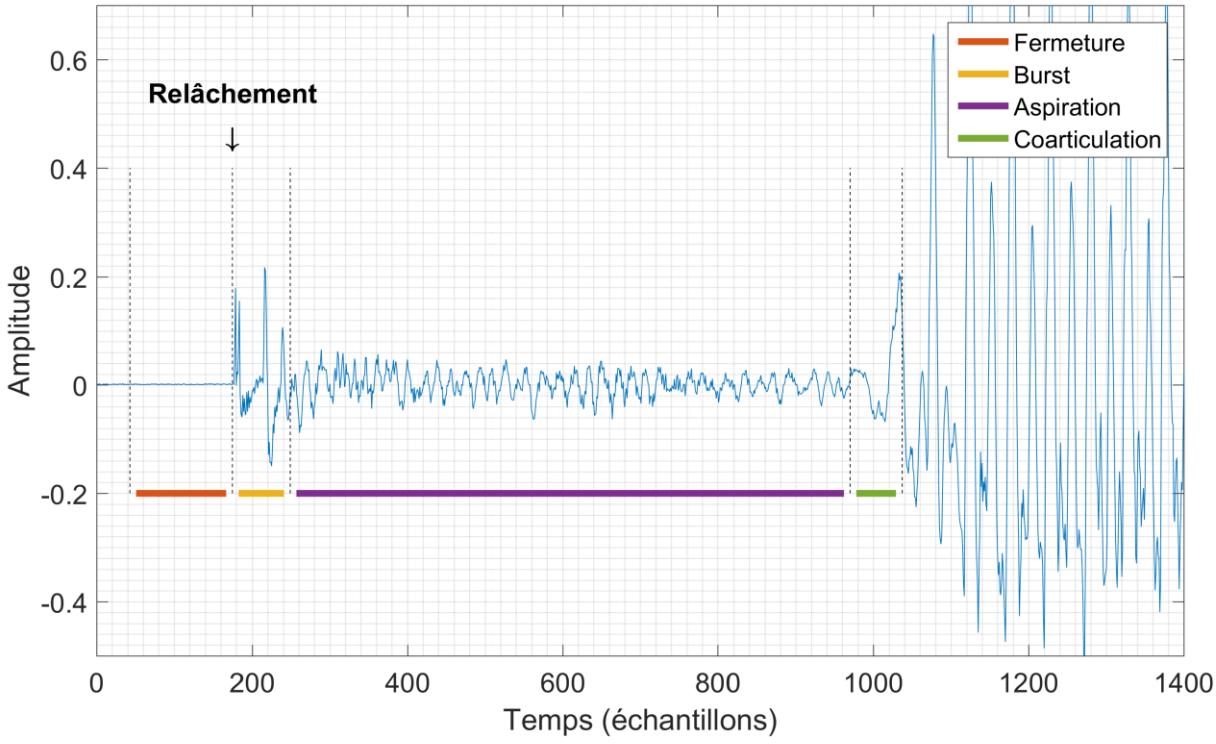


Figure 2.3 Phases acoustiques des plosives

2.1.4 Plosives voisées et non-voisées

Une plosive peut-être voisée ou non-voisée. En français, les plosives /p/, /t/, /k/ sont non-voisées, et les plosives /b/, /d/, /g/ sont voisées. Afin de caractériser cette différence, une mesure temporelle est utilisée : le délai d'établissement du voisement, appelé Voice-onset time (VOT) en anglais. Il s'agit de la durée au bout de laquelle la vibration des cordes vocales survient, par rapport au moment où le blocage du conduit vocal a commencé à céder. Cette durée peut ainsi être (voir Figure 2.4) :

- **Négative** : c'est le cas pour les plosives voisées qui présentent déjà une vibration des cordes vocales pendant la phase de fermeture,
- **Positive** : c'est le cas pour les plosives non-voisées, la vibration des cordes vocales survient peu de temps après l'éclatement.
- **Proche de zéro** : la vibration survient très rapidement après le relâchement, c'est le cas pour certaines plosives non aspirées (cela se produit souvent pour le /t/)

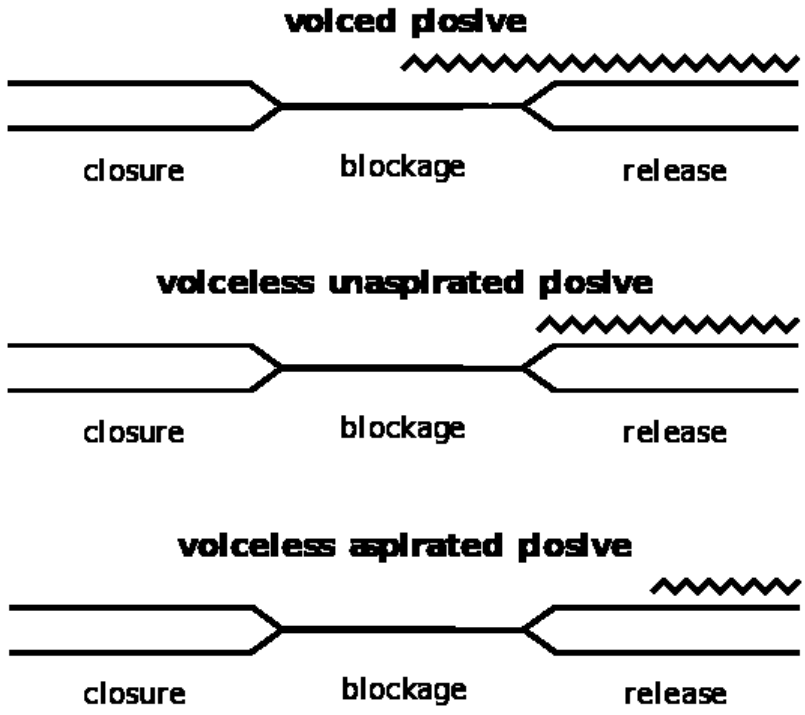


Figure 2.4 Schématisation des différents VOT
(Source : Wikipédia, libre de droits)

Ces différences résident dans la coordination des évènements supra-laryngés et laryngés. Les valeurs de VOT peuvent caractériser certaines régions articulatoires, mais sont spécifiques à chaque langue, et peuvent varier selon le contexte. Ainsi les plosives voisées à VOT négatif sont très présentes dans la langue Française.

2.1.5 Détails sur l'éclatement des plosives

Le VOT et la durée de la phase dite d'éclatement peuvent varier selon le lieu d'articulation (contrairement à la phase d'aspiration) [4]. Certains indices exposés dans [4] tendent à prouver que ceci est dû à la masse des articulateurs, et pas seulement aux langages, étant donné que ces variations se retrouvent dans plus d'une dizaine de langages.

D'autre part, suivant le lieu d'articulation, la longueur du couloir d'étranglement varie. L'article [4] montre que cette longueur a un impact sur la durée de l'éclatement, due à une évolution de la pression plus lente pour les vélaires et labiales.

L'article [4] montre également que les propriétés des parois du conduit vocal sont utiles à la modélisation des plosives, au moins pour la langue anglaise. Il prouve que l'ajout de ces considérations dans la modélisation de l'articulation au moment du relâchement est nécessaire pour obtenir un éclatement suffisamment puissant et étiré sur le temps. Enfin, il semblerait que ces propriétés jouent également un rôle dans la variation du VOT (mais on ne connait pas le facteur principal).

2.1.6 Quelques statistiques sur les plosives

TIMIT est une banque de phrases étiquetée en anglais largement utilisée dans le milieu du traitement de la parole, particulièrement pour tester l'impact de codeurs à usage téléphonique. L'étude [5] fournit des statistiques intéressantes sur les plosives contenues dans TIMIT. Tout d'abord cette étude renseigne la quantité de plosives et affriquées étiquetées sur la base (24 414 plosives et 2055 affriquées), ce qui permet de faire des calculs de ratio. Le Tableau 2.1 fournit les quantité de sons présentées dans [5]. On déplore cependant la non-distinction du contexte de ces plosives : le niveau de détail de contexte proposé par TIMIT comme le type de phrase ou l'identité du locuteur, n'est pas exploité.

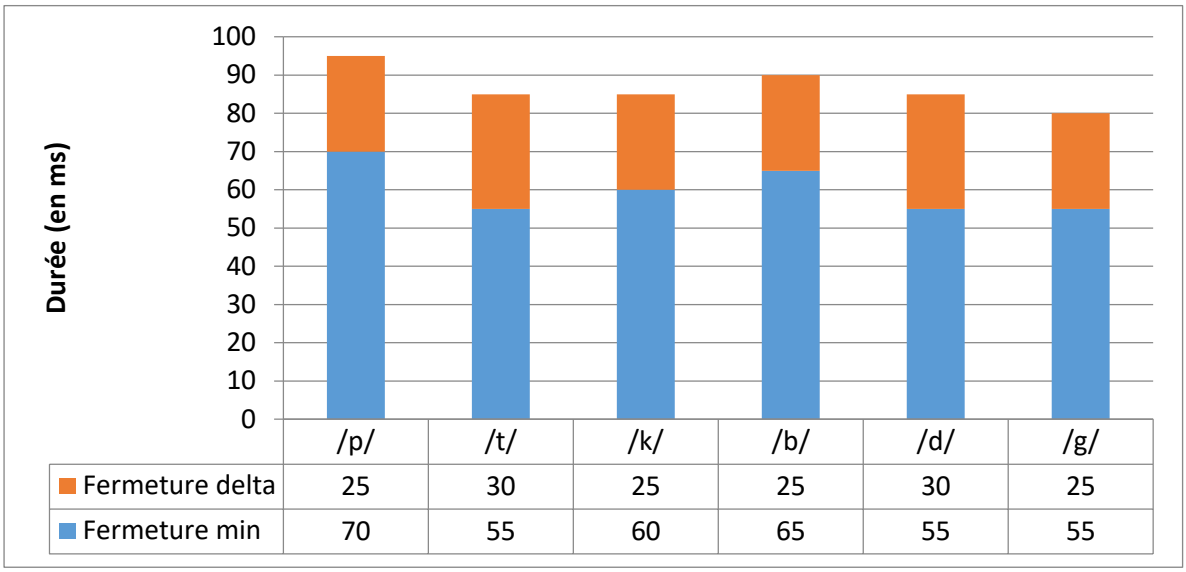
Tableau 2.1 Totaux des types de sons dans TIMIT en lien avec les plosives

Stops	Oral stop closures	24 414
	Nasal stop closures	18 101
	Affricates	2 055
Flaps	Oral flaps	3 649
	Nasal flaps	1 331
	Glottal stops	4 834

Tout d'abord, l'étude compare le ratio du nombre de fermeture de plosives étiquetées par TIMIT par rapport au nombre de relâchements. Ceci pourrait être intéressant, cependant l'article ne mentionne pas que généralement l'étiquetage d'un seul de ces deux évènements correspond en fait à une erreur d'étiquetage, ou plus rarement à un défaut de prononciation.

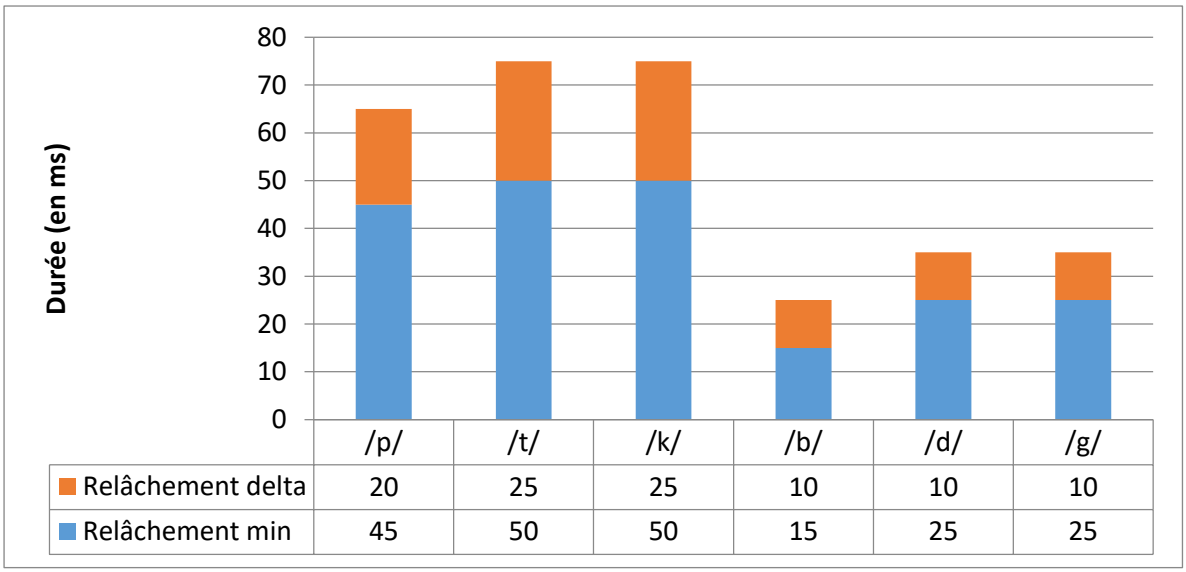
Enfin, l'étude expose les durées moyennes des phases des plosives segmentées et étiquetées dans TIMIT. Cette segmentation est faite en deux parties : la fermeture, et le relâchement. La partie relâchement, qui contient au moins l'éclatement et une partie de l'aspiration de la plosive, peut contenir jusqu'à la coarticulation vers le prochain phonème voire même une petite partie

de celui-ci. Ces données sont donc également intéressantes, mais il faut garder à l'esprit l'impact des erreurs de segmentation de TIMIT sur ces chiffres.
Les figures suivantes présentent les durées minimales et maximales observées pour les phases de silence et d'éclatement des plosives obtenues dans [5] (voir Figure 2.5 et Figure 2.6).



	/p/	/t/	/k/	/b/	/d/	/g/
Fermeture delta	25	30	25	25	30	25
Fermeture min	70	55	60	65	55	55

Figure 2.5 Durées minimales et maximales de la phase de fermeture (en ms)



	/p/	/t/	/k/	/b/	/d/	/g/
Relâchement delta	20	25	25	10	10	10
Relâchement min	45	50	50	15	25	25

Figure 2.6 Durées minimales et maximales de la phase de relâchement (en ms)

2.2 Détection des plosives

L'altération des plosives étudiée dans ce mémoire nécessite une méthode de détection afin de pouvoir être utilisée automatiquement. La technique utilisée dans le projet de recherche est dans un premier temps détaillée. Puis, d'autres méthodes et aspects liés à la détection des plosives sont abordés.

2.2.1 Indice de plosion

Cette technique est détaillée ici, et est également abordée en anglais dans l'article situé au cœur du mémoire (voir 3.5).

L'article [6] propose une méthode de détection de l'instant de transition entre la phase fermeture et la phase d'éclatement des plosives. Cette méthode s'appuie sur une mesure temporelle appelée le "plosion index", que l'on pourrait traduire en français par "Indice de plosion", ainsi que sur la MNCC, le maxima normalisé de la valeur de corrélation croisée à pitch synchrone. Un segment de parole à valeur élevée de MNCC correspond à un segment voisé, pour lequel les cordes vocales vibrent. A l'inverse, une valeur faible de MNCC correspond à un segment non-voisé. Ces deux paramètres sont évaluées suivant un découpage de la parole par « époques glottales » (grâce à une méthode proposée par les auteurs dans une autre publication [7]), une mesure qui a du sens en traitement de la parole car elle fournit une information sur la réponse fréquentielle du conduit vocal.

Ces paramètres sont ensuite exploitées par un système expert (à base de règles) qui statue sur la nature plosive d'un segment de parole énergétique, et peut donner une indication supplémentaire sur son aspect voisé ou non-voisé. Enfin l'article propose une évaluation des performances de cette méthode sur plusieurs banques de sons étiquetés (en anglais et en indien), dont TIMIT.

L'indice de plosion a été élaboré de façon à mettre en évidence la fluctuation de ratio d'amplitude moyenne de signal par rapport aux valeurs précédentes du signal. Ces considérations sont faites sur l'enveloppe de Hilbert du signal, pour s'affranchir de l'impact de la phase. Le principe est de comparer à un seuil le ratio entre l'amplitude du signal en un point, par rapport aux m2 échantillons précédents, espacés de m1 échantillons par rapport au point considéré (m2>m1) (voir Figure 2.7). On peut ainsi détecter les grandes variations d'énergie propres aux éclatements des plosives, en fixant un seuil de ratio obtenu suite à des essais expérimentaux. Les durées m1 et m2 ont elles aussi été obtenues expérimentalement, ces valeurs restent donc empiriques et sujettes à amélioration selon les auteurs.

Enfin, un filtrage est appliqué au préalable, ce qui permet une détection fiable des plosives voisées, et une élimination de la plupart des attaques voisées qui ne sont pas des plosives.
Concernant le design du filtre, l'article évoque simplement un filtrage passe haut avec fréquence de coupure à 400 Hz.

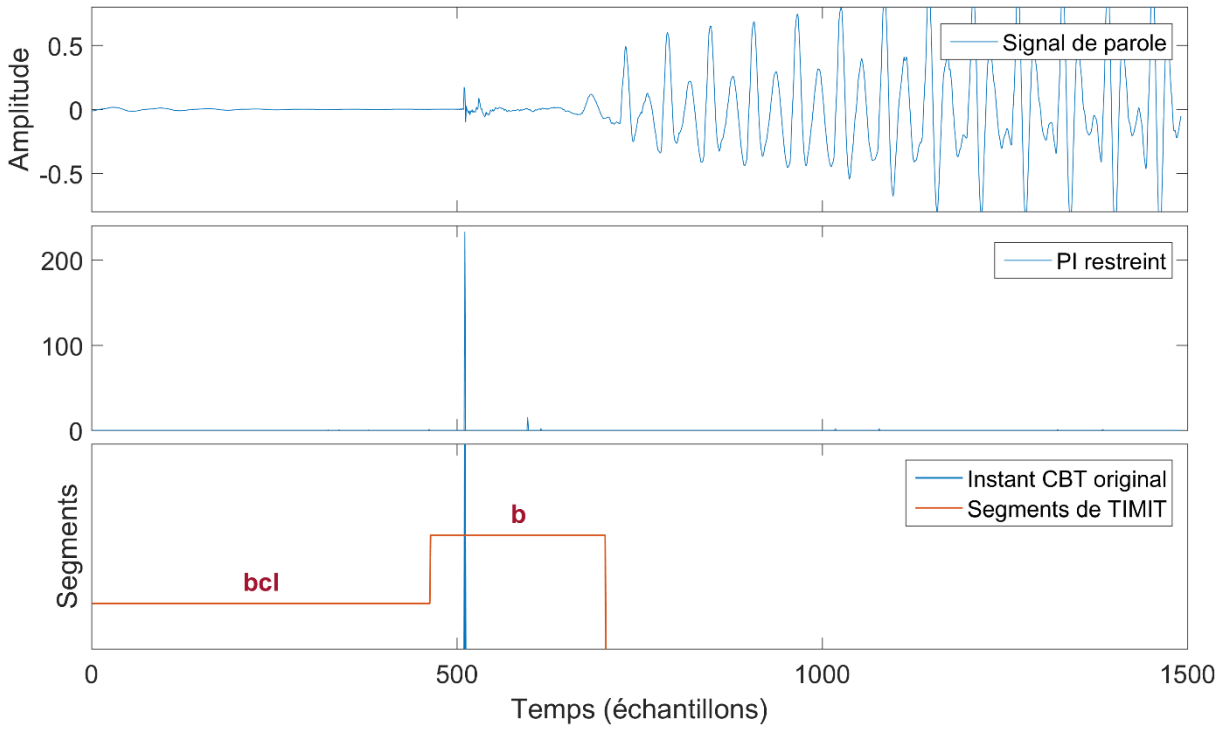


Figure 2.7 Détection grâce à l'indice de plosion
Dans le 3ème sous graphique, l'escalier représente la segmentation de TIMIT, et le pic vertical l'instant détecté

2.2.2 Autres méthodes

Il existe d'autres types d'approches pour la détection des plosives. Certaines sont ciblées sur les caractéristiques des plosives, d'autres sont des méthodes générales qui ont été entrainées pour ces types de sons.

Une équipe de chercheurs français [8] a élaboré un principe de détection par transformée en ondelette (ondelette mère de Morlet). La technique consiste à comparer la corrélation entre le module de la transformée d'une impulsion, et cette même transformée appliquée au signal de parole. Chaque maxima de corrélation est associé à un *éclatement*, un prétraitement s'étant dans un premier temps assuré que les caractéristiques du candidat correspondent bien à celles d'une occlusion voisée ou non-voisée. On a donc une détection du type d'occlusion, et la localisation temporelle de l'éclatement.

D'autres méthodes se basent sur la reconnaissance de motifs, tel que l'étude [9] basée sur l'utilisation du SVM (Support Vector Machine). Ce détecteur peut être utilisé sur les plosives, mais est utilisable pour détecter un autre motif. Il y a deux façons de l'utiliser si on postule que la durée de l'éclatement est comprise entre 10 et 50 ms. Soit on utilise un modèle statistique pour chaque durée possible, soit on utilise un modèle unique correspondant à la durée maximale. L'utilisation basée sur une durée maximale de 50 ms fournit ainsi de meilleurs résultats qu'un classificateur à base de HMM (Hidden Markov Models). Les performances obtenues sont : 1,2% de plosives non détectées et 2,0% de trames fausses positives (par rapport au nombre de frames n'étant pas des plosives. Ces résultats sont obtenus sur le corpus de test de TIMIT.

2.3 Segmentation des plosives

2.3.1 Mesure de la durée de l'éclatement grâce à l'indice de plosion

Dans ce projet de recherche, c'est cette méthode qui est utilisée, elle est détaillée dans la section 3.6. L'indice de plosion est considéré comme un indicateur de changement brutal de régime. Lors du processus de détection, un traitement est appliqué à l'indice de plosion pour restreindre ses valeurs aux points hautement transitoires. La version non-restreinte quant à elle contient donc des faux positifs, mais également des informations sur le déroulement des phases d'éclatement des plosives. Ainsi, une version non-restreinte de cet indicateur est comparée à partir de l'instant de transition fermeture-éclatement détecté. Lorsque la version non-restreinte de l'indice de plosion passe sous un certain seuil par rapport à la valeur lors de la transition (25%), on considère que le régime hautement transitoire, la partie la plus énergétique de l'éclatement, est terminé.

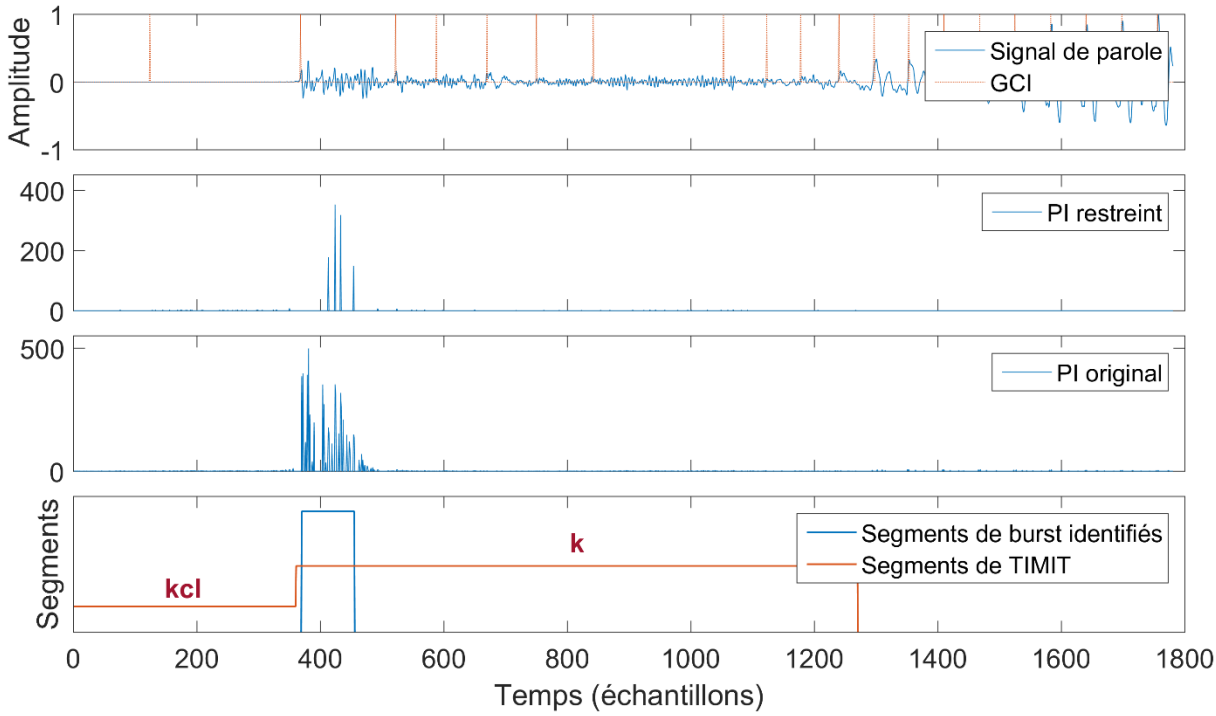


Figure 2.8 Segmentation de l'éclatement grâce à l'indice de plosion

2.3.2 Segmentation de l'aspiration grâce au taux de passage par zéro (ZCR)

Le taux de passage par zéro (Zero-crossing rate ZCR en anglais) est usuellement utilisé en traitement du signal pour distinguer des segments voisés et non-voisés de parole. Ce taux sera généralement plus élevé sur des segments non-voisés de parole. L'avantage de cette méthode est que la mesure fonctionne directement dans le domaine temporel. L'idée est de comparer à un seuil l'évolution du taux de passage par zéro du signal après la phase d'éclatement afin de déterminer la durée de sa phase d'aspiration, si elle est présente.

Tout d'abord, la présence de l'aspiration est déterminée en fonction de la valeur moyenne du ZCR (mesuré sur une fenêtre glissante) du signal pendant la seconde moitié de l'éclatement. Le seuil de présence est fixé à 0,2 (basé sur les observations). Si la valeur moyenne de ZCR dépasse ce seuil, on considère qu'il y a une phase d'aspiration.

Ensuite, la segmentation de cette phase d'aspiration est obtenue en analysant les évolutions de la MNCC (définie en 2.2.1) et du ZCR après la fin de la phase d'éclatement à chaque époque (on travaille sur des signaux non bruités). On considère que la phase d'aspiration s'achève lorsque :

- La moyenne de la MNCC dépasse le seuil de 0,6 (qui représente une présence de voisement) durant les 3 époques qui suivent
- La moyenne du ZCR franchis le seuil de 30% de sa valeur durant la phase d'éclatement, ou repasse en dessous du seuil de 0,2

Cette méthode a pour désavantage d'être sensible aux basses fréquences. Le choix de la fenêtre peut être difficile car elle doit permettre une évolution claire de la valeur de ZCR au voisinage de la transition pour tous les cas de figure.

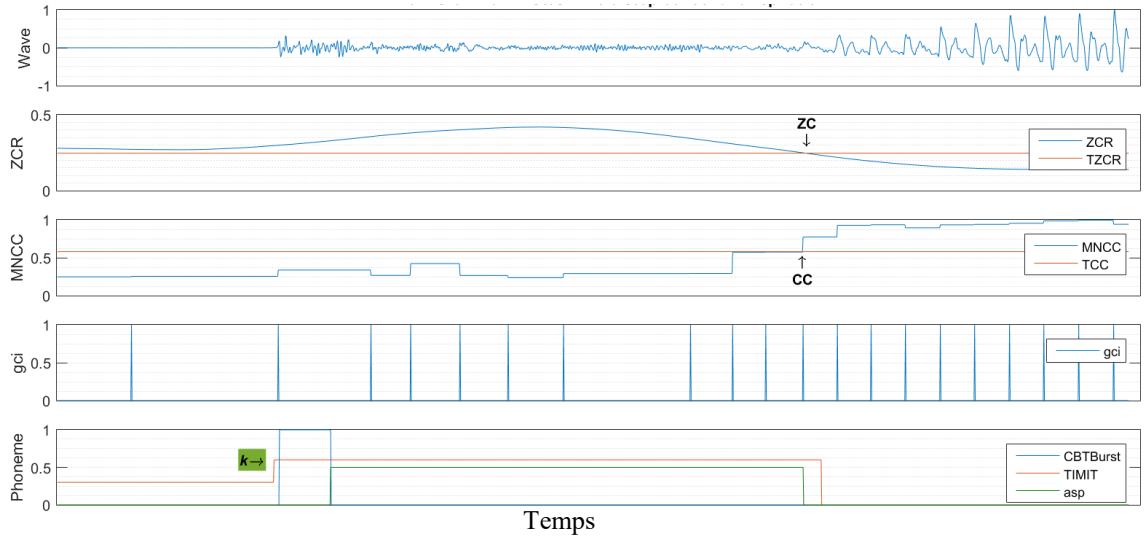


Figure 2.9 Segmentation grâce au ZCR

Sous graphique 1 : Signal de parole ; Sous graphique 2 : ZCR fenêtré, et seuil de 0,2 ; Sous graphique 3 : MNCC fenêtrée en escalier, et seuil de 0,6 ; Sous graphique 4 : époques glottales ; Sous graphique 5 : De gauche à droite, étiquette de TIMIT, segmentation obtenue de l'éclatement, segmentation obtenue de l'aspiration

2.3.3 Pente spectrale

L'étude de la variation de la pente spectrale au cours du temps donne une bonne information concernant les transitions phonémique, mais son utilisation pour distinguer les différentes phases des plosives est difficile à mettre en œuvre.

Le principe est de calculer la densité spectrale de puissance pour un segment de taille définie, d'effectuer une régression linéaire sur cette densité, et de récupérer le coefficient de la pente.

La taille du segment par rapport à la résolution de la fenêtre utilisée pour calculer le spectre pose problème pour détecter les changements de phase dans une plosive suivant l'échantillonnage utilisé, et la précision obtenue et souvent moindre par rapport à d'autres méthodes.

Une piste pourrait être de réduire l'intervalle de fréquence subissant la régression, ou d'appliquer une transformation préalable sur le signal.

2.3.4 Transitions entre classes phonétiques

La méthode présentée en [10] est utilisable dans le cadre plus vaste de la segmentation phonétique. Le principe est d'étudier le signal par trames de 40 ms (valeur choisie pour détecter un éventuel pitch de 20 ms max qui correspondrait à 50 Hz). Ces trames sont ensuite filtrées par un filtre passe bande entre 70 Hz et 500 Hz (voir Figure 2.10 et Figure 2.11). Ensuite, un algorithme recherche les positions par rapport au milieu de la trame de ses maxima les plus éloignés. Suivant la distance au centre de la trame du premier extrema (PFE) et du dernier (PLE), on peut déduire si la trame :

- Est homogène (pas de transition)
- Présente une forte/faible transition haute -> basse amplitude
- Présente une forte/faible transition basse -> haute amplitude

Cette méthode est combinée en amont avec une détection Silence/Bruit, et résulte enfin en un classificateur à 5 classes : Haute amplitude, Faible amplitude, Silence, Transition Haut bas, Transition bas haut. Les résultats sont assez souvent intéressants pour détecter avec précision l'instant de début d'un *éclatement*, mais ne sont pas fiables pour détecter la transition *éclatement* / aspiration ou la fin d'une plosive très courte ou au contenu spectral trop varié.

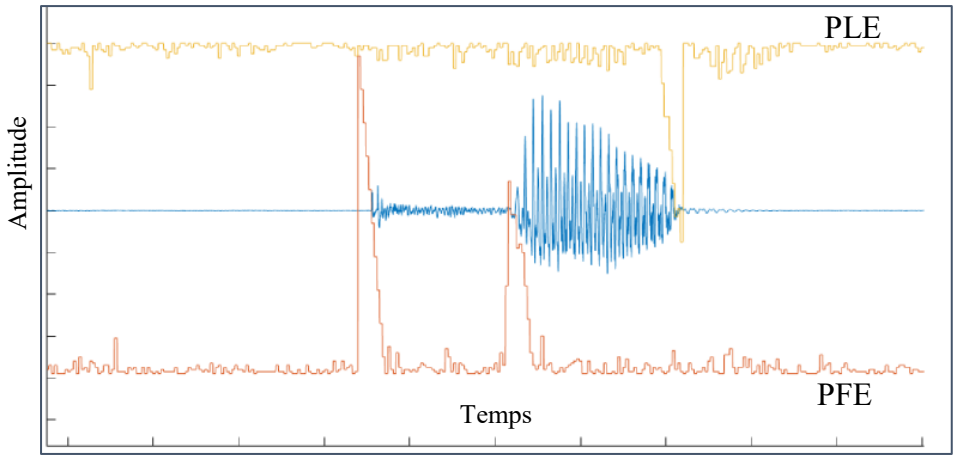


Figure 2.10 Transitions entre classes, segmentation réussie (/p/ /u/)
En jaune le PLE, en orange le PFE, en bleu le signal de parole

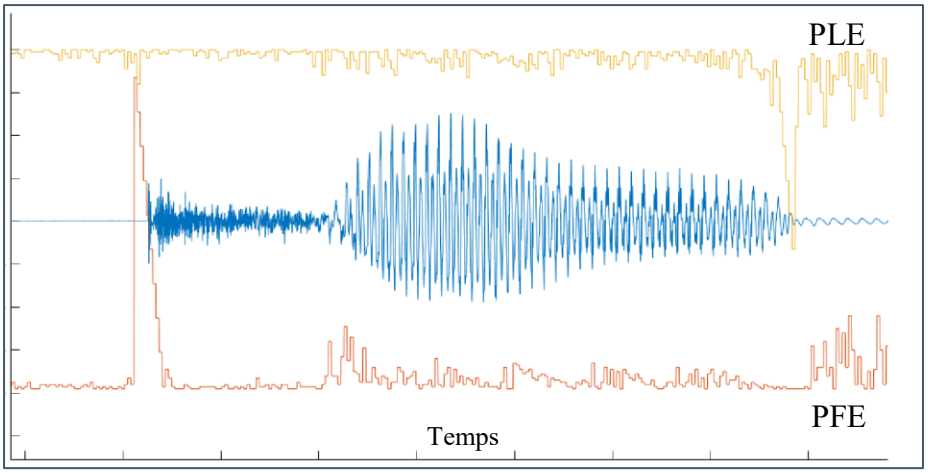


Figure 2.11 Transitions entre classes, segmentation de l'aspiration échouée (/t/ /a/)

2.4 Classification des plosives

Une classification des plosives est une information intéressante dans le cas d'un traitement ciblé par lieu d'articulation ou par phonème. Certaines méthodes s'appuient sur un découpage phonémique de la parole, d'autres sur une segmentation préliminaire précise des plosives, et enfin certaines se contentent de l'instant de transition fermeture-éclatement pour extraire les caractéristiques nécessaires à la classification.

La plupart des techniques s'appuient sur des caractéristiques acoustiques et phonétiques. La méthode présentée en [11] utilise des réseaux de neurones entrainés et testés sur le découpage phonétique de TIMIT, et obtient environ 90% réussite. Une autre méthode [12] estime les formants du modèle LPC grâce à MUSIC et à ESPRIT puis utilise les six premiers formants et leurs propriétés comme caractéristiques. La recherche de caractéristiques est faite sur des segments de 35 ms.

Les sous-parties suivantes présentent en détail trois méthodes utilisant respectivement des caractéristiques acoustiques, des segmentations variables, et une application aux plosives dites aspirées indo aryennes.

2.4.1 Classification par caractéristiques acoustiques et phonétiques

La méthode décrite dans [13] est une combinaison de système expert et d'analyse statistique. Elle utilise les caractéristiques suivantes :

- Transitions des formants et spectre de l'éclatement : l'importance de ces caractéristiques est controversé, elles sont importantes pour trouver le lieu d'articulation, mais ne seraient pas significatives pour la décision sur le voisement.
- Amplitude de l'éclatement : Une étude décrite dans [14] a prouvé que l'amplitude était plus faible pour un lieu d'articulation labial que pour un lieu d'articulation vélaire ou alvéolaire. Ceci serait d'autant plus vrai pour les plosives non-voisées. Une caractéristique intéressante, mais qui sert surtout à conforter l'information fournie par la première caractéristique présentée.
- Durée et voisement : Voice-onset Time (VOT) moyen plus long pour les non-voisées. Le pré-voisement quant à lui serait une caractéristique suffisante mais non nécessaire à la détection.

L'extraction de ces caractéristiques se base sur une segmentation par phonèmes. Ensuite un processus de décision intervient pour déterminer s'il s'agit d'une plosive, d'une fricative ou d'une consonne sonnante. Ensuite concernant les plosives, un processus détermine s'il y a du voisement et détermine le lieu d'articulation. Ces informations permettent de décider si une plosive est présente.

Pour détecter le voisement, trois caractéristiques sont utilisées : le voisement durant la fermeture (pré voisement), le VOT, et la durée de la fermeture. Les résultats présentés dans [13] montrent une précision de 96% sur un corpus de 1200 plosives. Ces résultats montrent également que la

durée de fermeture joue un rôle indirect très important (le pourcentage tombe à 90% lorsque cette donnée n'est pas utilisée).

Pour déterminer le lieu d'articulation, une séparation préalable des consonnes battues est effectuée (94% de réussite), puis une collection de caractéristiques permet de statuer sur le lieu d'articulation :

- La fréquence de l'éclatement
- Le second formant de la voyelle suivante
- Le maximum normalisé de la pente spectrale
- La proéminence fréquentielle de l'éclatement
- Les transitions des formants avant et après l'occlusion
- Une détection du voisement (similaire à celle déjà faite)

La fréquence de l'éclatement est la caractéristique la plus importante pour la détection du lieu d'articulation. Cependant cette caractéristique dépend beaucoup du contexte, c'est pourquoi elle n'est pas la seule utilisée.

Appliqué à la base TIMIT, cette méthode obtient un score de 96% pour la décision du voisement des plosives, de 90% pour le lieu articulation, et de 86% pour la classification individuelle finale des plosives.

2.4.2 Détection et classification basée sur un découpage phonémique

L'article [15] aborde le problème de la détection des plosives sous plusieurs angles. Les caractéristiques extraites du signal sont entre autres l'énergie dans certaines bandes importantes, l'énergie des enveloppes d'autres bandes, et la fréquence et la largeur de bande des quatre premiers formants. Ces caractéristiques sont ensuite utilisées pour classer les plosives, et plusieurs configurations de classes sont testées :

- Deux classes : Plosives et non-plosives (ce qui revient à une détection simple)
- Trois classes : Plosives voisées, plosives non-voisées, et non-plosives
- Sept classes : Une pour chaque plosive, et les non-plosives

Ce processus est appliqué de deux manières différentes sur TIMIT, premièrement en considérant des segments de 10 ms, puis en utilisant la segmentation par phonèmes de TIMIT sans considérer les étiquettes. Cette technique n'est pas très performante pour les plosives voisées, et repose sur la segmentation de TIMIT, ce qui fait que la segmentation phonétique effectuée en amont est indépendante mais impacte directement les performances. Les résultats indiquent de meilleures performances de classification lorsque les segments sont de durée fixe pour un nombre de caractéristiques observées faibles, mais la classification basée sur le découpage phonémique de TIMIT est améliorée si on augmente le nombre de caractéristiques (à environ 20) et présente une matrice de confusion plus équilibrée.

2.4.3 Détection des plosives aspirées (Indo-Aryen)

Les plosives aspirées sont très présentes dans les langues Indiques/Indo-Aryennes. Pour des applications de reconnaissance de parole, il est donc important de pouvoir les distinguer par rapport aux plosives non aspirées. L'article [16] vise donc la détection d'aspiration dans les plosives de la langue 'Marathi', qu'elles soient voisées ou non, par le biais de différentes caractéristiques acoustiques (On rappelle que le VOT est une caractéristique importante de cette distinction). Les caractéristiques utilisées sont :

- Le Voice-onset time (VOT)
- Caractéristiques fréquentielles : les deux premiers formants et leurs pentes spectrales
- L'index de synchronisation : représente la présence d'une composante haute fréquence liée à l'aspiration
- Les puissances et pentes de certaines bandes de fréquence : des indicateurs de tendance sifflante ou brutale de la voix
- RSB (SNR) : Analyse cepstrale du taux d'enrouement (ou de vieillissement) Harmonic-to-Noise-Ratio (HNR)

2.5 Modélisations des plosives

2.5.1 Modèles sinusoïdaux

$$\hat{x}(n) = \sum_{k=1}^{K} a_k(n) \cos(n\omega_k(n) + \varphi_k(n)) \tag{2.1}$$

La modélisation sinusoïdale consiste à projeter un signal dans une base de fonctions sinusoïdales (représentée par l'équation (2.1). Utilisée directement sur un signal de parole, le traitement appliqué est le même pour les sons périodiques et apériodiques. Pour un signal périodique contenant du voisement, une bonne représentation peut être obtenue avec un nombre de coefficients limités. En effet, ce modèle est capable de représenter à la fois le contenu périodique et quasi-périodique d'un signal.

Cependant, pour les sons non périodiques comme les plosives, deux principaux problèmes se posent. Tout d'abord, un grand nombre de coefficients doit être utilisé du fait de leur nature non harmonique. Ensuite, la résolution temporelle nécessaire pour capturer la nature de ses sons est plus élevée que celle nécessaire pour les sons périodiques. Il peut ainsi arriver que la taille de la fenêtre utilisée pour cette transformée soit plus large que certains sons transitoires, ceci pouvant également générer des problèmes de préécho.

Cette technique ce heurte donc à divers problèmes en manipulant les sons transitoires et ne présente pas de lien avec la perception des plosives.

2.5.2 Modèles sinusoïdaux atténués

Les modèles sinusoïdaux atténués sont proches du modèle sinusoïdal classique, mais au lieu d'utiliser une base de sinusoïdes, il repose sur une base de sinusoïdes possédant un facteur d'atténuation exponentielle. On retrouve dans cette catégorie plusieurs modèles présentés dans les parties suivantes. Ces modèles permettent de modéliser les attaques des sons transitoires de façon plus efficiente que les modèles sinusoïdaux classiques [17].

2.5.2.1 EDS (Exponentially Damped Sinusoids)

Le modèle EDS est présenté dans l'article [17], et il s'appuie sur une base de fonctions définie dans l'équation (2.2).

$$\begin{aligned} \hat{x}(n) &= \sum_{k=1}^{K} a_k e^{\gamma_k n} \cos(n\omega_k + \varphi_k) \\ &= \sum_{k=1}^{2K} r_k \phi_k^n, \quad 0 \le n < N \end{aligned} \tag{2.2}$$

Le paramètre r_k symbolise les phases et amplitudes initiales, le paramètre γ_k représente l'atténuation, et le paramètre ϕ_k désigne la fréquence et l'atténuation.

Une version préliminaire de ce modèle [17] pouvait présenter une instabilité aux bornes du segment considéré dans le cas d'un paramètre $|\phi_k| > 1$ correspondant à une amplification exponentielle. Ainsi, une distinction est faite dans le modèle pour les cas $|\phi_k| < 1$ et $|\phi_k| > 1$. Dans le premier cas, c'est une analyse du passé vers le présent, et dans le second cas l'analyse est effectuée dans le sens inverse. La distinction de ces deux cas implique ainsi une stabilité par rapport au facteur d'atténuation.

2.5.2.2 DDS (Damped & Delayed Sinusoids)

Comparé au modèle EDS, le modèle DDS [18] [19] comporte un paramètre supplémentaire permettant de contrôler le délai. Il procure donc une meilleure modélisation des signaux transitoires les plus puissants, et réduit les effets parasites de préécho rencontrés avec le modèle EDS. Enfin le délai permet de mieux modéliser l'attaque des sons plosifs. Cependant l'article [18] présente des mesures de performances en termes de RSB (rapport signal à bruit), et l'article [19] en termes de Normalized Mean Square Error (NMSE), l'aspect perceptuel n'est donc pas mis en avant. De plus, il a été remarqué que plus la modélisation DDS commence loin de l'attaque de la plosive, moins bonnes sont les performances.

2.5.2.3 PDDS (Partial Damped & Delayed Sinusoids)

Ce modèle est une évolution du DDS, la différence étant qu'il utilise un seul délai pour modéliser un ensemble de sinusoïdes.

2.5.3 Modèles sinusoïdaux adaptatifs

Les modèles sinusoïdaux adaptatifs sont des modèles capables de s'adapter aux caractéristiques locales d'un signal de parole. Le développement du modèle « Extended adaptative Quasi-Harmonic model (eaQHM) » [20] est inspiré par des recherches récentes sur les transformées non-stationnaires telles que la Fan-Chirp Transform (FChT) [21], [22]. Les paramètres du modèle sont ajustés itérativement.

L'article [23] présente une application de ce modèle à la représentation des sons non-voisés, incluant donc les plosives. La méthode présentée donne en moyenne une erreur de signal à reconstruction (Signal to Reconstruction Error Ratio SRER) inférieure de 93% par rapport au modèle sinusoïdal classique. La qualité a également été vérifiée à l'aide de tests perceptuels subjectifs de type ACR.

2.5.4 Approximation par transformée en ondelettes discrète (DWT)

L'article [24] présente un processus d'approximation appliqué aux coefficients d'une transformée en ondelettes discrète (DWT) utilisées sur les segments transitoires. Plusieurs techniques sont ainsi combinées de différentes manières et une évaluation de la qualité subjective résultant de ces méthodes d'approximation est fournie.

Dans un premier temps, les coefficients de la DWT sont calculés, et un seuillage est appliqué afin de ne conserver que la partie significative de la transitoire.

Ensuite, il y a des combinaisons d'application de :

- deux DCT aux coefficients obtenus avec la transformée en ondelette. La première sur les coefficients du bas de l'échelle (c'est l'approximation), et une autre sur les coefficients du haut de l'échelle (ce sont les détails).
- une LPC qui sert à prédire de nouveaux coefficients de DCT à partir des coefficients DCT de fréquences moindres.

Les résultats subjectifs obtenus sur 12 auditeurs pour les différentes méthodes, bien qu'assez proches dans l'ensemble, sont légèrement meilleurs pour la combinaison DWT + DCT + LPC.

2.6 Codage et psycho-acoustique

2.6.1 Grands principes

Les méthodes de codage de la parole et de l'audio se classent en plusieurs catégories. Elles peuvent être :

- Sans compression : le codage par impulsions (Linear pulse code modulation LPCM) [25], utilisé par exemple dans le format WAV
- Avec compression
    - Sans pertes : FLAC (Free Lossless Audio Codec) [26], ALAC (Apple Lossless Audio Codec) [27], …
    - Avec pertes : Ogg Vorbis [28], MP3 [29], Advanced Audio Coding (AAC)[30], …

Les codeurs qui effectuent une compression sans pertes ont pour stratégie d'exploiter les redondances statistiques des signaux. Les codeurs qui effectuent de la compression avec pertes peuvent cibler musique, parole, ou les deux. Ils ont pour objectif de compresser tout en optimisant le rendu perceptuel du signal reconstitué.

2.6.2 Compression psycho-acoustique

Le système auditif humain présente des caractéristiques qui permettent d'appliquer aux signaux audio des simplifications et des améliorations perceptuelles. On retiendra les principales caractéristiques suivantes :

- Réponse fréquentielle de l'appareil auditif : Suivant la fréquence, on n'obtient pas la même réponse en amplitude
- Bandes critiques : Chaque fréquence perçue s'inscrit dans une bande de fréquence appelée bande critique. Ces bandes s'élargissent lorsque la fréquence augmente.
- Masquage fréquentiel : Un son prépondérant dans une certaine bande de fréquence peut masquer ceux d'une bande voisine ou de la même bande.
- Masquage temporel : Un son peut avoir un effet masquant sur ceux qui le suivent, et également sur ceux qui le précèdent dans certains cas.

Ces propriétés sont utilisées dans la plupart des codecs audio utilisant une compression avec perte, comme MP3, et permettent une importante compression.

2.7 Plosives et perception auditive

2.7.1 Impact du bruit précédant les plosives

L'article [31] s'intéresse à l'impact sur l'intelligibilité d'un bruit additif précédant les plosives voisées, par rapport à l'impact d'un bruit coïncident. Différents spectres de bruits déphasés d'une durée de 0 ms (signifiant que le bruit coïncide) à 800 ms sont testés.

L'étude montre que l'intelligibilité des plosives voisées avec du bruit coïncidant précisément avec la parole est plus faible que si le bruit est continu (donc présent également avant la plosive et ininterrompu). Un corollaire de cette propriété est que l'augmentation de la durée du bruit précédant une plosive voisée équivaudrait à baisser le niveau de bruit effectif.

Après avoir conduit des tests utilisant différents contenus spectraux de bruit, une variation de l'adaptation suivant la fréquence est mise en lumière. L'étude conclut donc que ceci est dû à un processus d'adaptation de l'appareil auditif.

2.7.2 Impact de la durée du silence et de l'amplitude

L'article [32] présente des expériences de perception liées à des modifications de durée de silence et d'amplitude relative de l'éclatement. L'étude se résume à la réalisation et à l'interprétation des expériences décrites ci-après :

Expérience 1 : Cette expérience démontre que les durées de la fermeture et de l'éclatement jouent un rôle dans la distinction des mots « say » et « stay ».

Expérience 2 : Cette expérience démontre qu'il existe une corrélation entre la durée du silence et l'amplitude de l'éclatement pour les différentes plosives.

Expérience 3 & 4 : Ces expériences démontrent de façon inattendue que l'amplitude absolue (et non relative) de l'éclatement est perceptuellement significative.

Expérience 5 : Cette expérience démontre que la performance d'étiquetage des *éclatements* (pour plusieurs méthodes) est au moins égale à la performance de détection des *éclatements* pour les auditeurs.

Expérience 6 & 7 : Ces expériences mettent également en lumière une corrélation entre la durée de fermeture et l'amplitude de l'éclatement pour les labiales (avec les mots slit / split, slash / splash).

La conclusion sous-jacente de ces expériences serait que l'amplification des *éclatements* n'a aucun effet, à l'inverse de l'atténuation qui en aurait beaucoup.

2.7.3 Manipulation de la bande de fréquence des plosives

L'article [33] mets en lumière des caractéristiques temps-fréquences très importantes pour l'intelligibilité des plosives. En effet, il démontre qu'il existe une dépendance entre l'application d'amplifications sur certaines bandes de fréquences sur une plosive entière (*éclatement* et

aspiration) et le taux de reconnaissances à l'aide de tests perceptuels. On peut cependant noter que chaque bande de fréquence à modifier est sélectionnée manuellement, impliquant une difficulté de reproductibilité.

Cette étude contribue très largement à la compréhension de la perception auditive des plosives. Par rapport au projet de recherche décrit dans ce mémoire, cette étude propose une approche différente des plosives. Dans ce mémoire, le projet de recherche s'intéresse à une transformation appliquée sur la partie d'éclatement uniquement, alors que l'étude [33] applique des transformations groupées sur deux parties fondamentalement différentes des plosives, à savoir l'éclatement et l'aspiration.

2.8 Tests perceptuel

Afin d'évaluer l'impact perceptuel d'une transformation affectant un type de sons ou une transformation appliquée à toute la parole, un test perceptuel est nécessaire. Il peut s'agir d'un test objectif, dont les résultats sont basés soit sur une comparaison statistique, soit sur l'impact sur les caractéristiques perceptuelles humaines, ou subjectif, dont les résultats sont obtenus par le biais de scores attribués par des auditeurs experts ou non.

Il est donc important de connaître les différents outils disponibles dans ce domaine pour quantifier l'impact d'une modélisation en lien avec la perception. La présentation des différentes évaluations dans les parties qui suivent se limite aux techniques standardisées par l'ITU, car le but est d'utiliser une méthode clairement définie et reproductible. Plusieurs publications récentes proposent d'autres outils pour l'évaluation perceptuelle de la qualité.

2.8.1 Tests objectifs

Dans la catégorie des tests objectifs, on distingue deux sous-ensembles de tests. Les évaluations comparatives évaluent le signal considéré par rapport à une référence. Les évaluations non-comparatives n'ont besoin que du signal considéré, et confrontent le signal à des modélisations de l'appareil auditif ou vocal humain.

2.8.1.1 Évaluations comparatives

2.8.1.1.1 PSQM (Perceptual Speech Quality Measure)

La méthode PSQM correspond à la recommandation ITU-T P.861 de 1996. C'est une méthode d'évaluation comparative de deux signaux basée sur la perception de la qualité de la parole. Le principe est que le signal de parole subit une transformation qui le transpose dans un domaine perceptuel, qui permet ainsi de comparer avec d'autres signaux au moyen d'une distance. Ce calcul est effectué en utilisant les bandes critiques, et plus particulièrement l'échelle de Bark (composée de 24 bandes critiques). Enfin, cette méthode se base sur la réponse en fréquence d'une transmission téléphonique, et prend en compte le rajout d'un bruit ambiant [34], classiquement pris en compte dans ce genre de test.

2.8.1.1.2 PEAQ (Perceptual Evaluation of Audio Quality)

La méthode PEAQ correspond à la recommandation ITU-R BS.1387 de 1994, dont la dernière mise à jour date de 2001. Dans cette méthode, les propriétés perceptuelles de l'appareil auditif humain sont modélisées, en utilisant surtout les propriétés du masquage fréquentiel. Le principe consiste à appliquer un modèle perceptuel sur le signal de référence et sur le signal cible, puis à comparer de leur score. Les résultats tendent statistiquement vers ceux d'un test ACR (décrit en section 2.8.2.1).

2.8.1.1.3 PESQ (Perceptual Evaluation of Speech Quality)

La méthode PESQ décrite dans [35] correspond à la recommandation ITU-T P.862 de 2001. C'est une méthode de mesure de qualité de la parole lors d'une conversation téléphonique à bande réduite. PESQ utilise de vrais échantillons de parole enregistrés dans son processus. Cette méthode est principalement utile dans les cas où on a un délai variable au cours du temps dans un signal de parole.

2.8.1.1.4 POLQA (Perceptual Objective Listening Quality Assessment)

La méthode POLQA décrite dans [36] correspond à la recommandation ITU-T Rec. P.863 de 2011 (mise à jour en 2014). Appliqué à la parole et successeur du PESQ, c'est le test objectif pour la parole le plus aboutit actuellement, ses résultats sont généralement bien corrélés avec ceux obtenus avec un test subjectif de type ACR (méthode par catégories absolues). POLQA utilise de vrais échantillons de parole, tout comme PESQ. Enfin, cette méthode prend en charge les signaux de parole dans une plage de fréquence étendue : téléphone (300 à 3400 Hz), et voix HD (50 à 14000 Hz).

2.8.1.2 Évaluations non comparatives

2.8.1.2.1 Mesure de la qualité mono extrémité P.563

La méthode P.563 décrite dans [37] correspond à la recommandation ITU-T P.563, la dernière mise à jour date de 2004. Elle est utilisée par exemple pour évaluer la dégradation de la qualité due à la transmission de l'information par réseau téléphonique (tout particulièrement pour les cas à bande étroite). Son processus d'évaluation tient compte à la fois d'une modélisation du conduit vocal humain, et d'une modélisation perceptuelle des sons perçus comme anormaux à l'oreille. Enfin, ses résultats tendent vers de ceux d'un test ACR.

2.8.1.2.2 ANIQUE (Auditory Non intrusive quality estimation)

La méthode ANIQUE correspond à la recommandation ITU-T SQ12 de 2001. Le processus d'évaluation de cette méthode se base sur l'étude de l'enveloppe temporelle du signal, et sur une modélisation du système auditif humain aux niveaux périphériques et centraux.
Selon une étude qui compare les performances de P.563 et d'ANIQUE à celles d'un véritable test ACR, ANIQUE surpasse P.563. Sur certaines bases de données, ANIQUE présente autour de 85% de similitude à ACR, où PESQ en présente 93%.

2.8.2 Tests subjectifs

2.8.2.1 ACR (Absolute Category Rating)

La méthode ACR décrite dans [38] correspond à la recommandation ITU-T P.800 de 1996, la dernière mise à jour date de 1998. C'est une méthode d'évaluation de la qualité perçue que l'on utilise soit pour des conversations téléphonique ou IP, soit pour mesurer la distorsion causée par un codec, soit pour mesurer l'effet d'une perte de données dans un contexte particulier.

Cette méthode utilise une échelle de notation ACR à 5 niveaux (5 : Excellent, 4 : Bon, 3 : Convenable, 2 : Médiocre, 1 : Mauvais). Les évaluateurs doivent donner un score à l'échantillon écouté (sans références). À l'issue d'un tel test, le score obtenu est appelé score MOS (Mean Opinion Score).

2.8.2.2 DCR (Degradation Category Rating)

La méthode DCR correspond à la recommandation ITU-T P.800, Annexe D. Les évaluateurs comparent un échantillon audio à un échantillon référence, et doivent évaluer le degré de dégradation par rapport à l'original.

Cette méthode utilise une échelle de notation DCR à 5 niveaux (5 : Pas de dégradation audible, 4 : dégradation audible mais non dérangeante, 3 : dégradation légèrement audible, 2 : dégradation dérangeante, 1 : dégradation très dérangeante).

Compte tenu de la présence d'une référence, on considère que les résultats d'un tel test sont plus précis que pour un test ACR, cependant la durée du test est doublée. À l'issue d'un tel test, le score obtenu est appelé score DMOS (Degradation Mean Opinion Score).

2.8.2.3 CCR (Comparison Category Rating)

La méthode CCR correspond à la recommandation ITU-T P.800, Annexe E. C'est le même principe que pour DCR, à la différence près que les échantillons référence et cible sont présentés dans un ordre aléatoire inconnu de l'évaluateur.

Cette méthode utilise une échelle de notation CCR à 7 niveaux de -3 à +3 (+3 : Bien meilleur, +2 : Meilleur, +1 : Légèrement Meilleur, 0 : Semblable, -1 : Légèrement pire, -2 : Pire, -3 : Bien Pire). Le signe sera inversé à postériori suivant l'ordre de diffusion. À l'issue d'un tel test, le score obtenu est appelé score CMOS (Comparison Mean Opinion Score).

Une évaluation de type CCR présente l'avantage de pouvoir donner un meilleur score à la cible qu'à la référence. Ceci le rend donc particulièrement adapté dans les situations où l'on teste des processus d'amélioration de la qualité de la parole.

2.8.2.4 PC (Pair Comparison)

La méthode PC est utilisée pour évaluer la supériorité de la qualité de plusieurs échantillons entres eux. L'évaluateur doit simplement désigner le meilleur échantillon à chaque combinaison (BA, AC, …). La mise en œuvre de cette méthode est simple, mais le temps nécessaire augmente grandement suivant le nombre de systèmes à évaluer.

2.8.2.5 MUSHRA (Multiple Stimuli with Hidden Reference and Anchor)

La méthode MUSHRA correspond à la recommandation ITU-R BS.1534-1, de 2001-2003. L'évaluateur écoute la référence, puis plusieurs autres échantillons, incluant une référence cachée et plusieurs ancres. Les ancres sont des sortes de «références dégradées » dont le score se situe autour de 20. Elles servent à donner un intervalle de notation à l'évaluateur.

Dans le test MUSHRA, l'échelle de notation va de 0 à 100 : on peut finement départager plusieurs échantillons.

L'avantage par rapport au test ACR, c'est qu'un nombre inférieur de participants suffit pour obtenir des résultats statistiquement significatifs. Cependant, alors qu'un test ACR peut être réalisé avec des auditeurs naïfs, un test MUSHRA doit être effectué par des auditeurs experts.

CHAPITRE 3 : RELEVANCE OF THE BURST PHASE OF STOP CONSONANTS

ÉTUDE DE LA PERTINENCE PERCEPTUELLE DE LA PHASE D'ÉCLATEMENT DES PLOSIVES, ET IMPLICATIONS POUR LE CODAGE DE LA PAROLE

A STUDY OF THE PERCEPTUAL RELEVANCE OF THE BURST PHASE OF STOP CONSONANTS WITH IMPLICATIONS IN SPEECH CODING

Vincent Santini, Philippe Gournay, Roch Lefebvre

AVANT-PROPOS

Auteurs et affiliation :

V. Santini : étudiant à la maîtrise, Université de Sherbrooke, Faculté de génie, Département de génie électrique et de génie informatique.

R. Lefebvre : Professeur titulaire, Directeur du Groupe de Recherche sur la Parole et l'Audio, Ingénieur électrique (membre de l'OIQ), Université de Sherbrooke, Faculté de génie, Département de génie électrique et de génie informatique.

P. Gournay : Professeur associé, Ingénieur électrique, Université de Sherbrooke, Faculté de génie, Département de génie électrique et de génie informatique.

Date de soumission : 13 mai 2016

Date d'acceptation : 29 juin 2016

État d'acceptation : Accepté, en attente de retouches finales avant présentation et publication

Revue : 2016 IEEE Workshop on Multimedia Signal Processing (MMSP 2016)

Titre français : Étude de la pertinence perceptuelle de la phase d'éclatement des plosives, et implications pour le codage de la parole.

Contribution au document : Cet article contribue au mémoire en élaborant une technique d'évaluation de l'impact de la mise à zéro de la phase d'éclatement des plosives, en décrivant en détail la méthodologie et l'analyse des résultats.

Résumé français : Les plosives sont des sons importants dans un signal de parole. Ils contribuent grandement à l'intelligibilité, et ont un impact important sur la qualité subjective. Cependant, de par leur nature dynamique et imprédictible, ils sont difficile à coder en utilisant des techniques conventionnelles telles que la prédiction linéaire, et le codage par transformée. Cet article présente un système capable de détecter, segmenter, et modifier les plosives dans un flux de parole. Ce système est utilisé pour évaluer l'hypothèse suivante : mettre à zéro la phase d'éclatement des plosives a un impact négligeable sur la qualité subjective de la parole. Cet impact est évalué sur une banque de signaux de paroles. Les résultats montrent que cette altération drastique a en réalité peu d'impact perceptuel. Les débouchés pour le codage de la parole sont également abordés.

Note : Le contenu de cet article diffère légèrement de celui qui a été soumis.

3.1 Abstract

Stop consonants are an important constituent of the speech signal. They contribute significantly to its intelligibility and subjective quality. However, because of their dynamic and unpredictable nature, they tend to be difficult to encode using conventional approaches such as linear predictive coding and transform coding. This paper presents a system to detect, segment, and modify stop consonants in a speech signal. This system is then used to assess the following hypothesis: Muting the burst phase of stop consonants has a negligible impact on the subjective quality of speech. The muting operation is implemented and its impact on subjective quality is evaluated on a database of speech signals. The results show that this apparently drastic alteration has in reality very little perceptual impact. The implications for speech coding are then discussed.

Keywords— Stop consonant; plosive; affricate; transient; muting; burst; closure; segmentation; detection; TIMIT; CMOS

3.2 Introduction

Human speech sounds are typically classified into three categories: voiced sounds, fricatives, and transient sounds. Simple models that form the basis for efficient coding modes exist for the first two categories of sounds. Voiced segments are efficiently coded using sinusoidal coding [39], or by passing a periodic excitation into a linear prediction (LPC) filter [40]. Noise generators and spectral shaping by an LPC filter are often used for fricatives segments. There are also hybrid approaches, such as the CELP [41] coding model, in which no decision is taken about the category of sound, but that combine an LPC filter with a pitch and a fixed codebook to encode periodic and stochastic components of a signal.

This paper takes a closer look at the third category of sounds, transient sounds, and especially at stop consonants. These sounds are particularly complex, dynamic and unpredictable. There are no simple models to represent them, therefore they are generally considered as hard to encode efficiently. Indeed, their unpredictable nature makes them rather unsuitable for coding using linear prediction, and their dynamic nature makes them susceptible to pre-echoes in transform coding [42], especially at lower bitrates.

Several models for stop consonants have already been proposed, including transform models based on Damped and Delayed Sinusoids (DDS). These models achieve good transient waveform modeling without pre-echo, and can be used in more general sinusoids + noise + transient models [18]. A specific frame-based approximation with wavelet transforms applied to transients has also been proposed [24]. The interest of these models for coding, especially at low bitrates, is however limited because their aim is to reproduce the waveform of these transient segments. This paper is based on the assumption that perceptual properties of stop consonants can be used to design non-waveform matching models that would be more efficient for coding. The perceptual properties of plosives have already been the subject of investigations. For example, several relevant time-frequency perceptual features in stop consonants have been highlighted in [33]. These features correspond to the whole plosive, including the burst and aspiration phases, which represents a duration of around 60 ms. It was also shown in [33] that these features are highly correlated with the recognition score of consonant vowel sequences.
The goal of the present paper is to improve the understanding of the perceptual aspect of these transient sounds in order to develop new coding strategies. To this end, an experiment is conducted in which the burst phase of stop consonants is altered, to evaluate the perceptual relevance of this particular phase.

The paper is organized as follows. Stop consonants are characterized in section 3.3. An overview of the proposed system to mute the burst of stop consonants is given in section 3.4. This system is composed of three subsystems. In a first step, stop consonants are detected (section 3.5). Then, the burst phase of these stop consonants is segmented (section 3.6). Finally, the alteration (muting of the transient segments of stop consonants) is applied to the speech signal (section 3.7). The material used for the experiment is presented in section 3.8. Listening test results that show that this apparently drastic alteration has in reality a rather limited perceptual impact are given and discussed in section 3.9. The possible implications for speech coding are also discussed in section 3.9.3, and finally conclusions are drawn in section 0.

3.3 Characterization of stop consonants

Stop consonants are an important component of human speech, especially for intelligibility. From an acoustical point of view, they are complex and variable sounds. Their production mechanism differs widely from the production of other more stationary speech sounds, such as vowels and fricatives [43], [44]. As indicated in [2], they are “produced by forming a closure in the vocal tract, building up pressure in the mouth behind this closure, and releasing the closure”. The acoustical nature of the sound events resulting from this production mechanism is very specific and five distinct phases can be observed. First, the closure phase is composed of either acoustical silence or background noise (and may contain some periodicity for voiced plosives). The burst phase consists in a sudden impulse; it is followed first by a short turbulent noise phase, then by an aspiration noise phase. Then again, depending on the context, there is a coarticulation phase to the next phone. These phases are highly variable in duration and intensity; depending on the speaking rate, the degree of articulation, and the accent of the speaker, some phases can even be missing or imperceptible. A good example is the aspiration phase for voiced stop consonants such as /b/.

In the literature, some phases are often grouped together. In this paper, a segmentation into only four phases is considered: closure, burst (which regroups both the burst and the turbulent noise phases mentioned above), aspiration, and coarticulation (Figure 3.1). This study focuses on the burst phase because it is the most dynamic and unpredictable part of stop consonants.

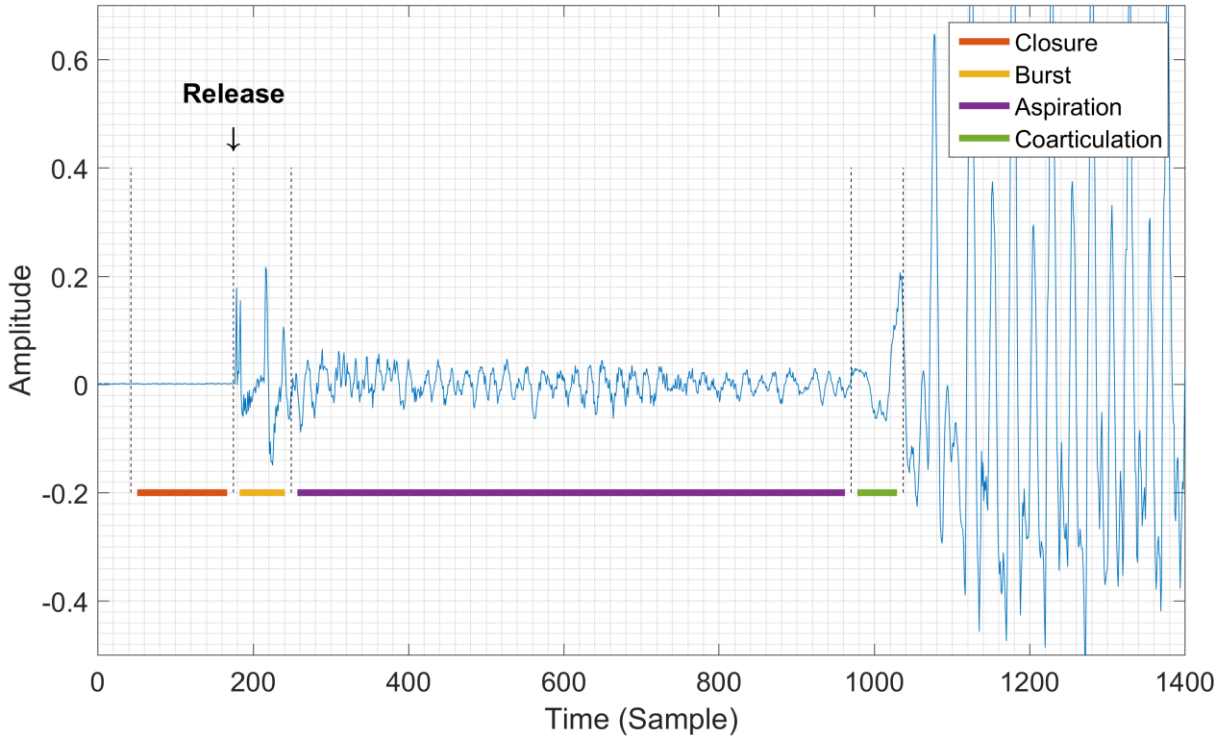


Figure 3.1 Phase segmentation of a stop consonant
By definition, the closure phase is the silence preceding the release; it is not represented here in its entirety for clarity.

3.4 Overview of the alteration system

The purpose of the proposed system is to automate signal alterations specifically targeted to stop consonants. As shown in Figure 3.2, this system is composed of four processing blocks which are presented briefly in this section, and then described in more details in sections 3.5 to 3.7. The first processing block (pre-processing) consists in applying a 50 Hz low pass filter (to remove the DC component). The second processing block aims at determining the closure-burst transitions (CBT) which marks the beginning of the significant part of each stop consonant. A modified version of the CBT detection algorithm from [6] is used. The third processing block (burst segmentation) aims at determining the beginning and end of the transient segment of each detected stop consonant. The fourth and last processing block consists in altering the detected and segmented burst phases of stop consonants. It applies a muting operation: the burst segment is replaced by an all zeros segment of the same duration. The rest of the signal is kept unchanged. The following three sections give algorithmic details on the techniques that were selected and implemented in each processing block of the system. These techniques were chosen keeping in mind the goal of this study which is to evaluate the impact of the muting alteration.

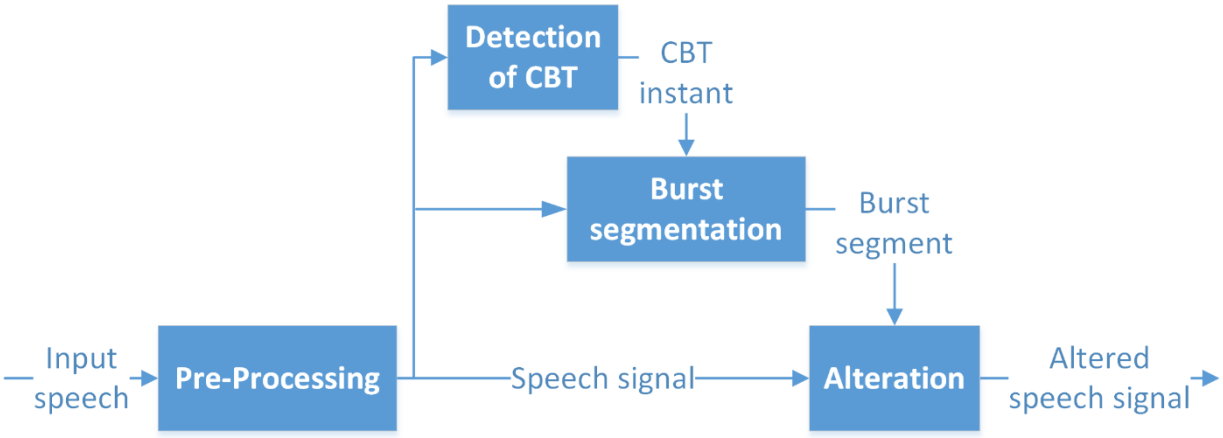


Figure 3.2 The proposed alteration system
This diagram shows the succession of all sub-systems required to implement the desired alteration.

3.5 Detection of the closure-burst transition

Several detection algorithms for stop consonants can be found in the literature [9], [15], [45]–[47]. The system presented in this paper uses the CBT as an indicator of the presence of a plosive. The CBT marks the beginning of the burst phase of a stop consonant, and is located using the plosion index (PI) [6].

The main advantage of this approach is that the PI is a temporal feature which can provide a good temporal resolution. However, our implementation of the original method was found to sometimes provide a delayed decision. This method was nonetheless chosen for three reasons: Firstly, because a detailed algorithmic description is available in [6]. Secondly, because it has been tested on several corpora, including TIMIT which is used in our experiments. Finally, its performances are among the best of all methods, with an equal error rate (EER) of only around 7.8% on the TIMIT corpus. Thus, the original algorithm will be described briefly first, and then the modification introduced to minimize the delay will be presented.

3.5.1 Original CBT detection algorithm using the plosion index

The CBT detection algorithm presented in [6] uses the PI which characterizes a local abrupt increase in energy. The PI is defined in [6], as "the ratio of the peak amplitude in the transient to the average of absolute values over an appropriate interval excluding the immediate neighborhood of the peak". The PI is given by the following formula and is called "raw plosion index" or "raw PI" in the rest of this paper:

$$\mathrm{PI}(n0, m1, m2) = |s(n0)| \,/\, \mathrm{savg}(m1, m2) \tag{3.1}$$

where n0 is the time instant considered, s(n) is the sampled signal, m1 is the immediate neighborhood duration to exclude, and m2 is the non-immediate neighborhood duration to consider (after excluding m1), and where savg(m1,m2) is the mean absolute value of s(n) in the interval [n0-(m1+1) ; n0-(m1+m2)].

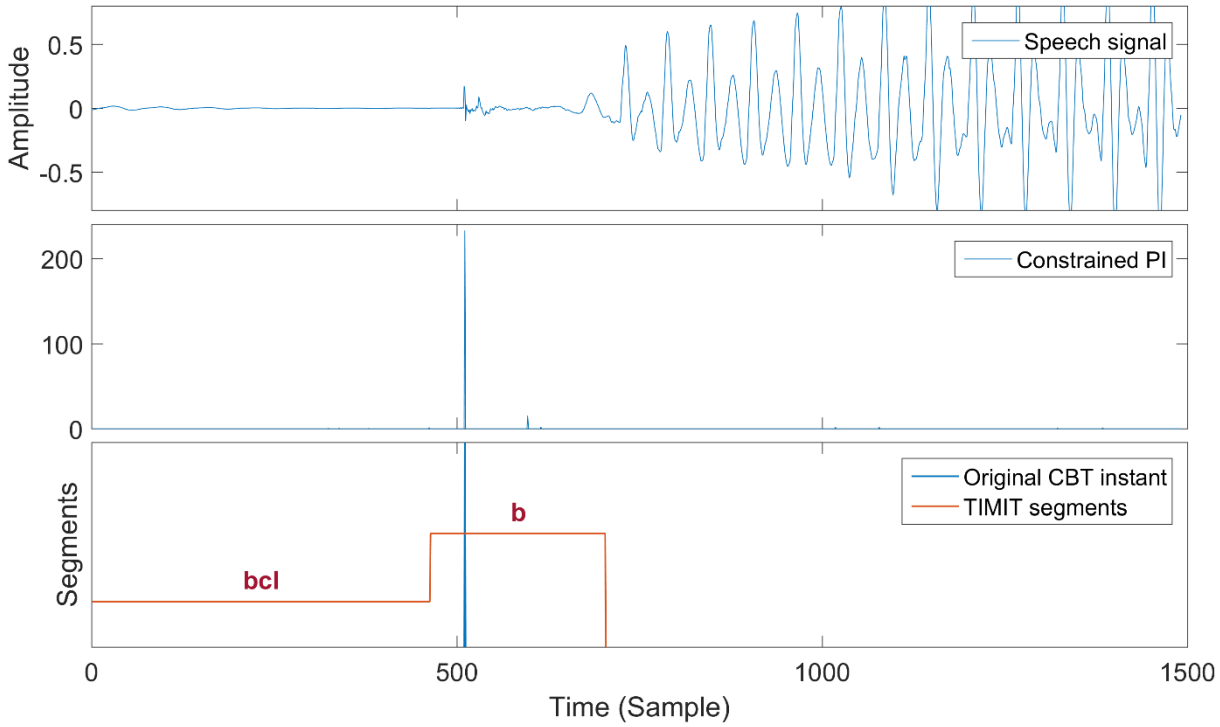


Figure 3.3 Detection of a stop consonant release
Showing TIMIT labels (in orange) and the output of our implementation of the CBT algorithm (in blue). The segment labeled "bcl" is the closure phase of the stop, and the segment labeled "b" is the release (in red).

To reduce the occurrence of false positives, the PI is calculated only at local maxima of the Hilbert envelope and at zero-crossing (ZC) instants. In this paper, the PI calculated with these two conditions is called "constrained plosion index", or "constrained PI". Finally, a decision system based on the PI value and on a neighborhood analysis (see details in [6]) is used to determine whether the time instant corresponds to a CBT or not. This process also eliminates most false positives, especially during coarticulations and voiced onsets. The behavior of the detection algorithm is illustrated in Figure 3.3, where a segment of speech containing a stop consonant is shown with the corresponding PI and detection instants.

3.5.2 Modified CBT detection algorithm

In the previous section, a process used in the original CBT algorithm in order to reduce the occurrence of false positives was mentioned. In many cases where using the constrained PI defined in the previous section results in a late detection of the beginning of the burst, it can be seen that the corresponding raw PI actually exhibits high values around the actual beginning of the burst (Figure 3.4). Late detections can have a devastating impact on quality, which would prevent us from concluding on the perceptual relevance of the burst phase of plosives. Therefore, to avoid late detections, the CBT instant detected using the constrained PI is modified using the raw PI, thus maintaining the low false positive advantage provided by the constrained PI.

To use the information contained in the raw PI, the three glottal epochs (segments of signal between moments of significant excitation) that precede the CBT instant detected using the original algorithm are analyzed. The glottal epoch instants (GCIs) are obtained using the method described in [7]. The modified CBT instant is the earliest instant for which the raw PI is at least 25% of the raw PI at the original CBT instant. This process is illustrated in Figure 3.4.

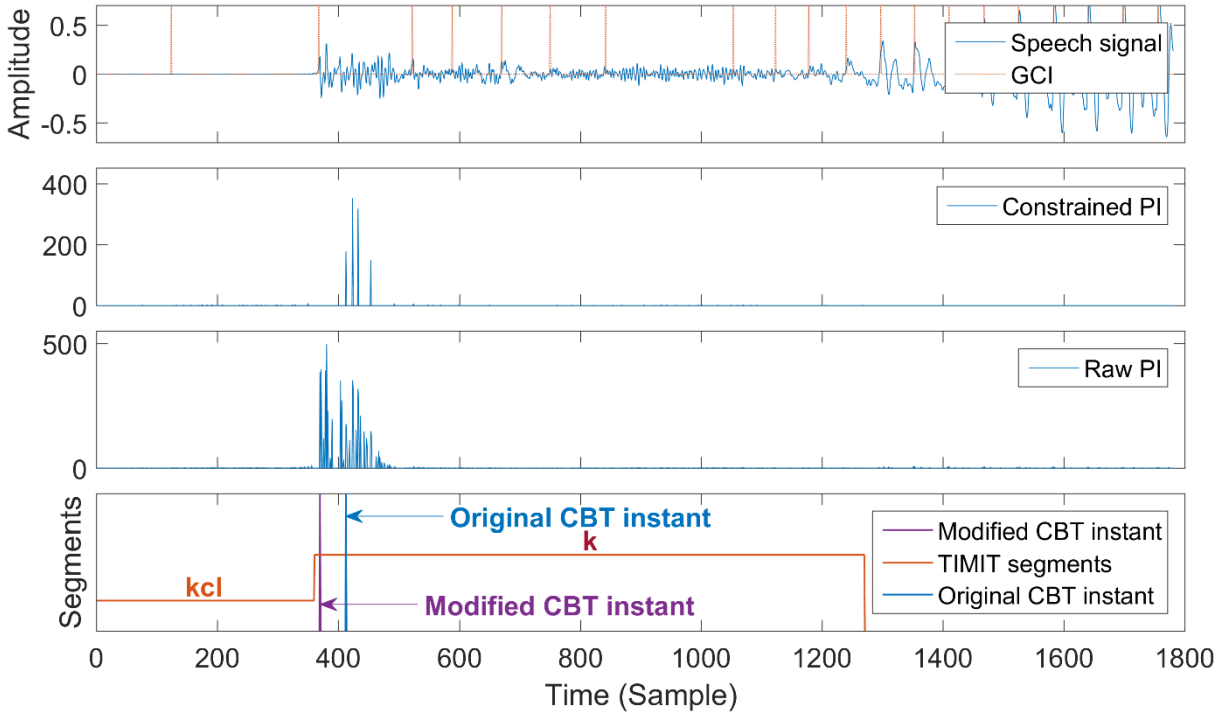


Figure 3.4 CBT modification
This is an example where our implementation of the original CBT algorithm results in a late detection (indicated in blue). The modified instant is represented in purple, and the TMIT labels are indicated in orange for reference.

3.6 Burst segmentation of stop consonants

The algorithm described in section 3.5 provides the beginning of the burst of stop consonants. The alteration process also requires the end of the burst. The identification of that endpoint relies on the evolution of the raw PI after the CBT. The PI is not only a meaningful feature to detect the CBT; it is also a good indicator of transient regions of a signal, which is characteristic of the

entire burst phase. Thus it is used to monitor the decay of the burst, before the aspiration phase begins.

The end of the burst is searched for between the modified CBT instant and the beginning of the next GCI. The last sample in this interval for which the raw PI is more than 25% of the PI at the CBT instant is identified as the end of the burst. Considering the way the PI is calculated, the end of the burst also corresponds to a ZC instant. This process is illustrated in Figure 3.5.

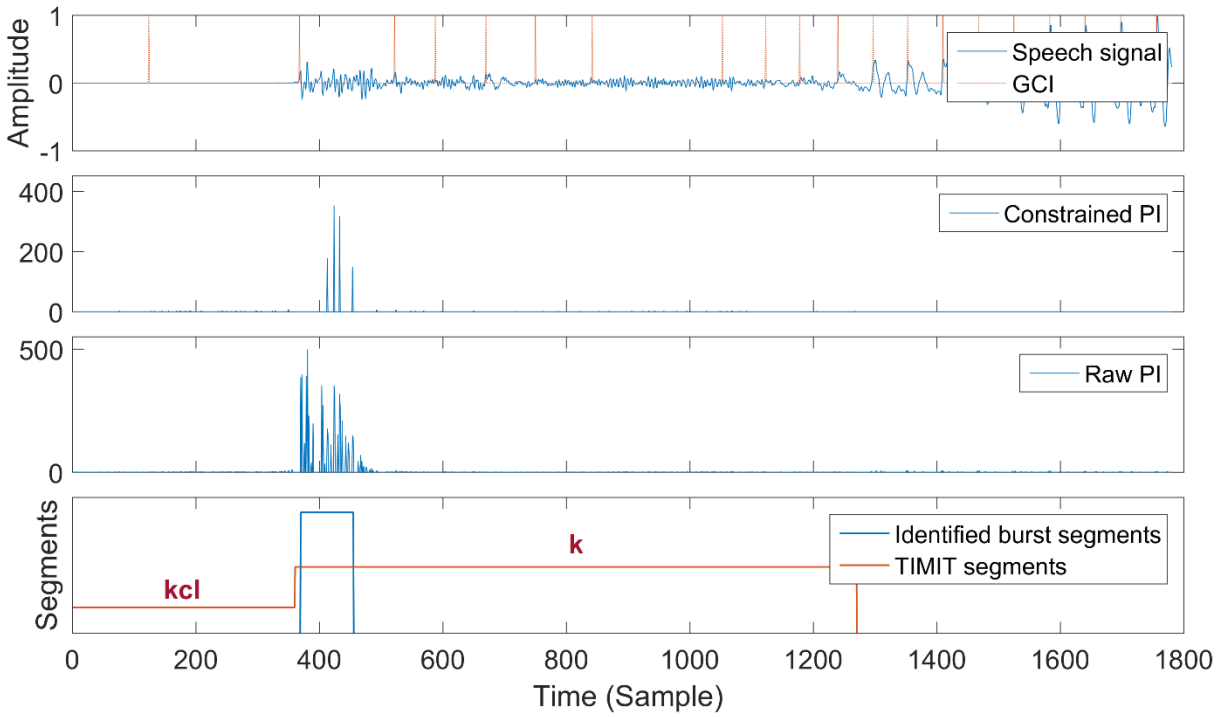


Figure 3.5 Burst segmentation of a stop consonant
This shows the constrained PI used for the detection, and the raw PI used for the segmentation.

3.7 Muting of the burst phase of stop consonants

The goal of this paper is to evaluate the perceptual relevance of the burst phase of stop consonants by altering the corresponding audio signal. To this end, several informal experiments using various alterations of these burst phase have been conducted. As explained hereafter, these preliminary experiments led to the muting alteration.

A first alteration method consisted in replacing the burst segment of each stop consonant in a sentence by a burst segment taken from a dictionary of bursts. The dictionary contained one entry for each category of stop consonant. The burst was adjusted to match the energy of the original signal. This approach did not produce good results, most of the time the alteration was noticeable.

Then, a second alteration method consisted in using a single generic replacement for all stop consonants. The burst segments were replaced by low-level noise segments. This second method

showed better results, but the presence of white or colored noise was sometimes noticeable by the listener, except at a very low-level.

These two approaches proved nonetheless to be perceptually less subjectively transparent than a simple muting, according to further informal tests conducted. Therefore it has been decided to conduct a more formal experiment using only the muting alteration.

The segmentation process described in section 3.6 always produces detected and adequately labeled segments that begin and end at ZC instants. Each of these segments is simply zeroed out. A fade-out triangular window whose length is equal to 5% of the duration of the segment is applied before the beginning of the segment. Similarly, a fade-in triangular window of the same length is applied after the end of the segment. This muting process is illustrated in Figure 3.6. A typical example of muted stop consonant in a speech signal is also given in Figure 3.7.

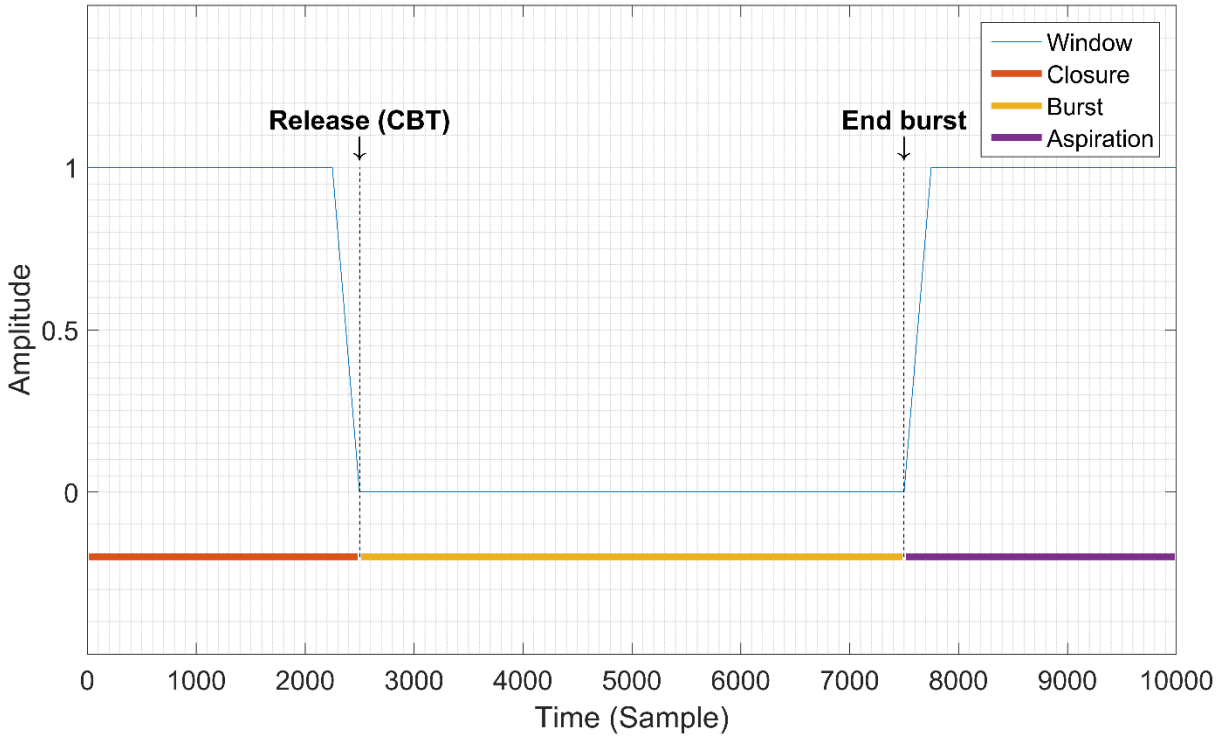


Figure 3.6 Muting mask
This shows an example of the mask used to mute the burst phase of a stop consonant. The length of the fade-in and fade-out triangular windows is equal to 5% of the duration of the muted segment.

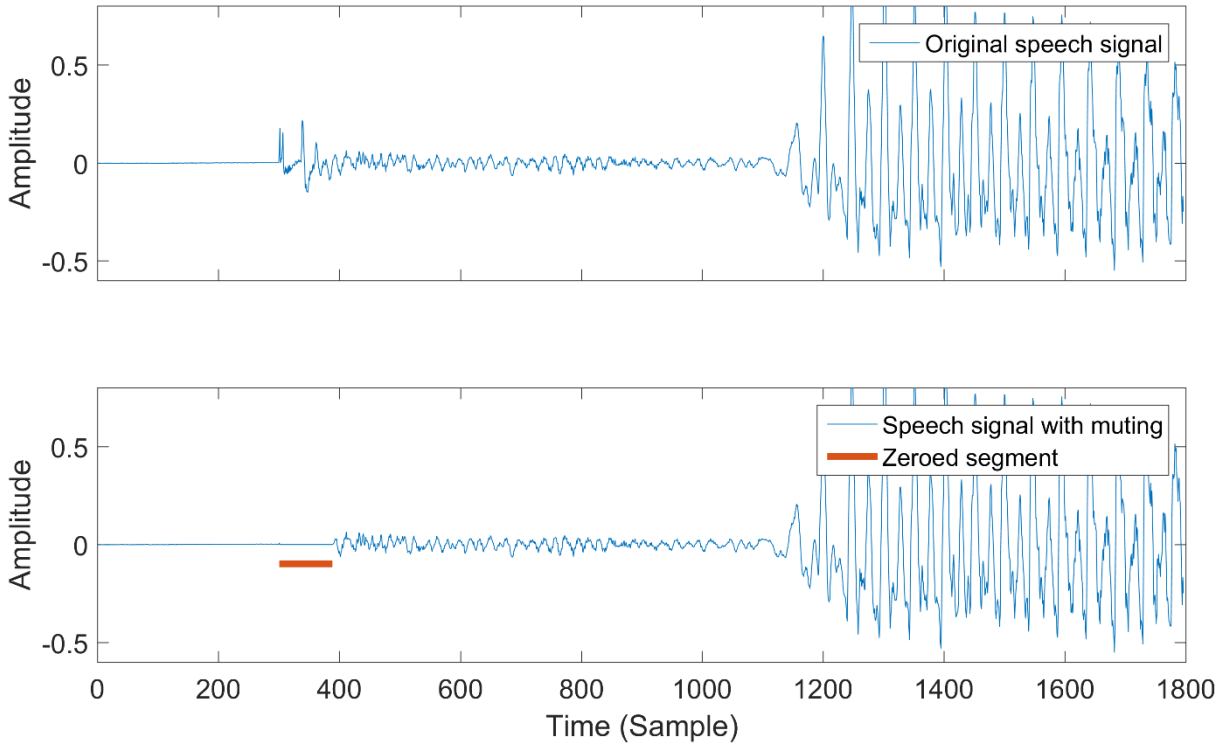


Figure 3.7 Illustration of the muting of the burst phase of a stop consonant

3.8 Test Material

3.8.1 The TIMIT Acoustic-Phonetic Continuous Speech Corpus

The proposed system is tested on the TIMIT corpus [48]. This corpus is composed of phonetically-rich sentences produced by male and female American speakers with various American English dialects. This corpus includes a time-aligned phonetic transcription. Although the phonetic labeling has been hand-verified and is assumed to be correct, the time alignment is frequently too imprecise to be used to assess the perceptual relevance of any category of signal. In TIMIT, stop consonants and affricates are segmented into two parts: a closure part, and a "stop release" part. A close inspection of that segmentation has revealed that the segmentation of the release part is not always perfect. It normally begins at the CBT instant, but there generally is a fluctuating offset of a few milliseconds, which is not negligible considering the average duration of the burst phase of stop consonants. The segmentation of the release part normally ends after the plosive's aspiration phase, but sometimes it ends after the coarticulation to the next phoneme has begun.

Thus, since the proposed muting system requires a precise segmentation of the stop consonant burst phase, the detection and segmentation algorithms presented in sections 3.5 to 3.6 were used instead of using TIMIT labels. Nevertheless, TIMIT labels are used after the detection process to avoid false positives (which rate is around 15% with our implementation of the presented method), since these false positives would prevent us from evaluating the perceptual relevance of the alteration of the bursts.

3.8.2 The test sub-corpus

The selection of items that composes the test sub-corpus used in the objective and subjective evaluations of section 3.9 is given in Table 3.1. The duration of the test items is not constant, but it is typically around 3 seconds.

Table 3.1 Sub-corpus test items distribution

Sentence	*SA1*	*SA2*	*SX/SI*	
%	25%	25%	50%	
Total	20	20	40	80

This table shows the proportion of sentences used from the TIMIT corpus, which were randomly chosen among several speakers. SA1 and SA2 are dialect sentences. SX are phonetically-diverse sentences. SI are phonetically-compact sentences.

The distribution of stop consonants in the entire TIMIT corpus compared to what is altered in the test sub-corpus is shown in Figure 3.8.

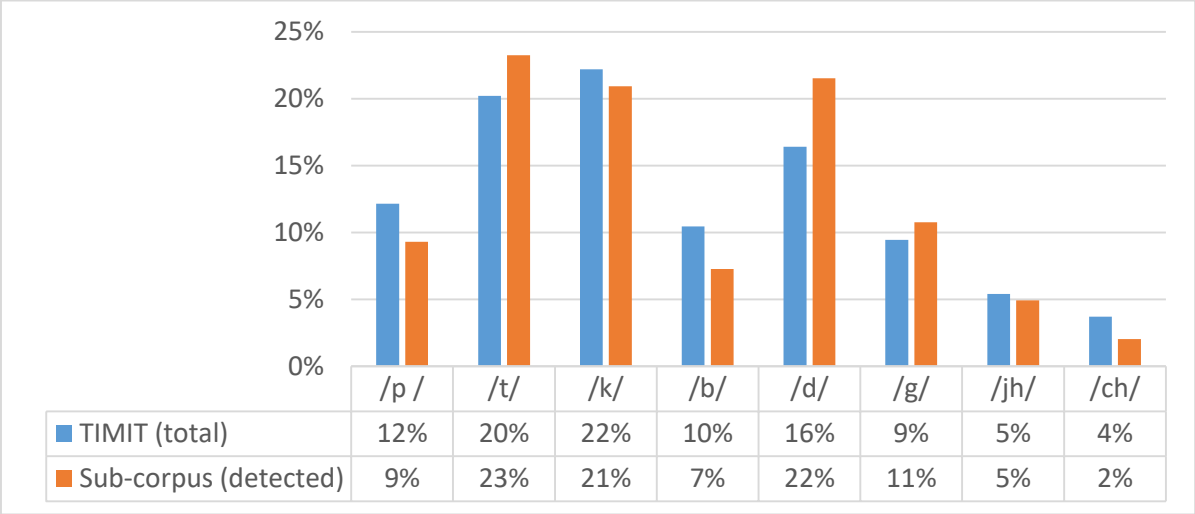
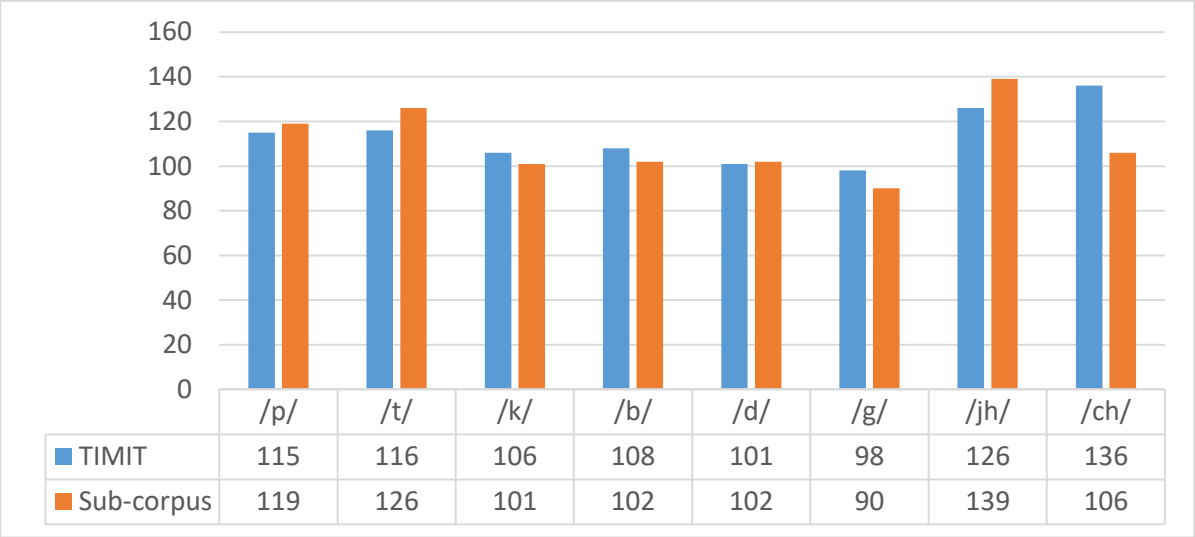


Figure 3.8 Distribution of the muted stop consonants within the sub-corpus
This figure also shows the distribution of stop consonants in the entire TIMIT corpus (where approx. 15% of stop consonants are not detected).

3.9 Results

3.9.1 Objective evaluation

Figure 3.9 shows the mean duration of the burst phase of stop consonants and affricates, estimated using the system described in sections 3.5 to 3.6. This shows that on the TIMIT corpus, the muting alteration affects approximately 107 samples of audio signal per stop consonant, and around 130 samples per affricate (respectively 6.7 ms and 8.1 ms).



	/p/	/t/	/k/	/b/	/d/	/g/	/jh/	/ch/
TIMIT	115	116	106	108	101	98	126	136
Sub-corpus	119	126	101	102	102	90	139	106

Figure 3.9 Mean duration of the burst phase of each stop consonant
This figure represents the mean burst phase duration for each muted stop consonant and affricate (measured on TIMIT and on the test sub-corpus). Duration unit is in samples, with a sampling frequency of 16 kHz.

3.9.2 Subjective evaluation

The perceptual impact of the muting of stop consonants is evaluated using a subjective Comparison Category Rating (CCR) test, according to the ITU-T Recommendation P.800 [38]. The evaluation uses a Comparison Category Rating (CCR) scale with 7 levels from -3 to +3. The test contains 80 test sentence pairs (original speech waveform and speech waveform with muted plosive burst phases), and 20 control sentence pairs (original speech waveform repeated twice). The test items used for the control pairs are drawn randomly from the test sub-corpus. The CCR test was conducted in a quiet room with Beyerdynamic DT770 headphones. Twelve non-native English expert listeners participated. All test items were normalized, in order to have the same global sound level during the subjective audio test.

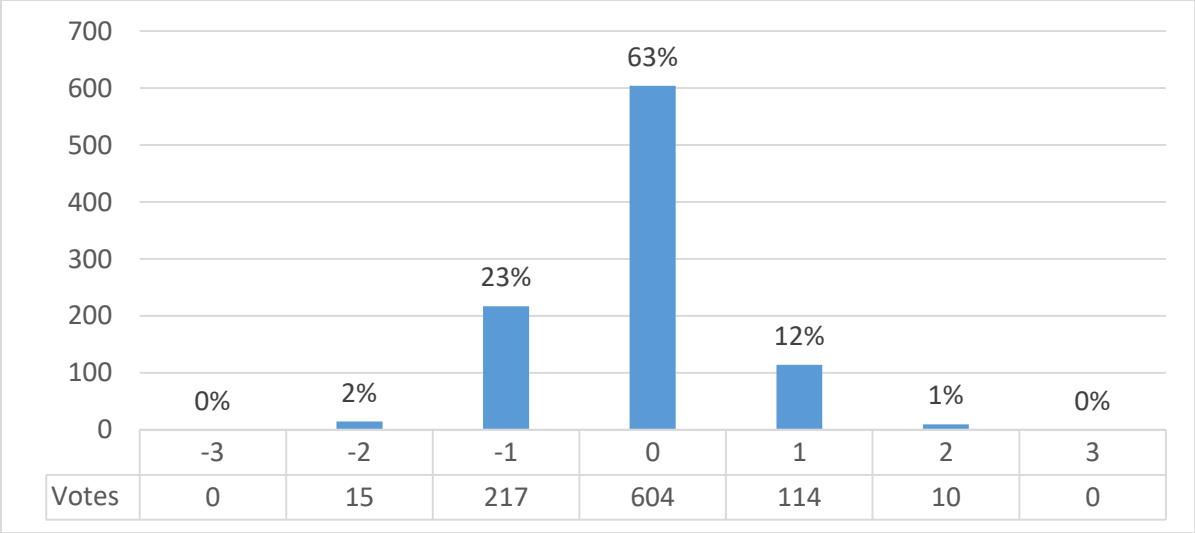


Figure 3.10 CMOS results for the sub-corpus test items (with stop consonants muting)

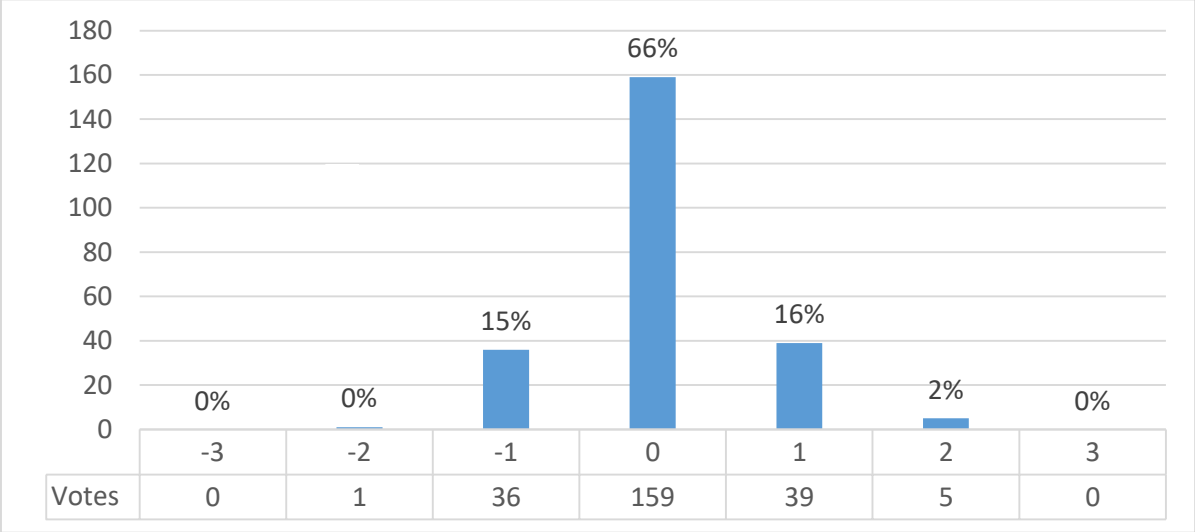


Figure 3.11 CMOS results for the control pairs

The results presented in Figure 3.10 show that the alteration was perceived negatively for 25% of the test items. Note that all listeners reported the degradation to be very tricky to notice. The CMOS value is -0.12 for the entire sub-corpus. To analyze the significance of this result, a one-sample t-test is used. The null hypothesis is that the CMOS value is 0 (H0: CMOS = 0). In this

situation the altered stimulus would be perceptually equivalent to the original one. The alternate hypothesis is that the CMOS value is less than 0 (H1: CMOS < 0). The t-test is conducted at a 5% significance level and fails, with a confidence interval of [-0.15 -0.08].

The conclusion of the t-test is that the degradation introduced by the muting operation is very small (-0.12 in average on a scale from -3 to +3) though statistically significant to the 5% significant level.

3.9.3 Perspectives for speech coding

Transient sounds are known to be particularly difficult to encode. Several speech and audio coders need to adapt their coding strategy to maintain a good subjective quality when a transient is detected. For example, OPUS [49] and CELT [50] switch to a different time-frequency configuration, and the VMR-WB codec [51] uses a special coding mode that operates at the maximum bitrate. This adaptation to transient sounds requires additional complexity and/or an increased bitrate.

Furthermore, in some telecommunications systems, bitrate adaptation or mode switching is subject to technical constraints that can result in a delayed adaptation or an increased encoding delay. This is the case for the AMR and AMR-WB codecs in the GSM system, as explained in [52]. Thus, preprocessing the speech signal to mute transient segments would limit the need for such adaptation, and potentially save bitrate, complexity and delay.

Some informal listening tests have been conducted using speech signals that were first altered (muting of the burst phase of all stop consonants), and then coded and decoded using the AMR-WB codec (at different modes) and the OPUS codec (at several bitrates, mono, wideband, with a frame size of 20 ms). The subjective quality of the coded and decoded altered speech signals was found to be equivalent to the quality of the coded and decoded original speech signals. This indicates that the muting alteration, though radical, is perfectly compatible with speech coding operations.

There would be other ways of using the fact that muting the burst phase of stop consonants has a very limited impact on the subjective quality of a speech signal in the field of speech coding. Rather than avoiding mode switching, one could develop a new, very simple mode to code transient segments. This new mode would consist in muting, rather than really coding, these segments. To assess the potential of the proposed approach in terms of bitrate saving, the ratio between the cumulated duration of segmented burst phases and the duration of active speech for the TIMIT database has been computed. This ratio is equal to 1.3% for stop consonants and 0.58% for affricates. Hence, assuming a fixed-rate codec, with a variable framing scheme that perfectly adapts to muted segments, and supposing that no information needs to be transmitted to indicate muted segments, a maximum of 1.88% of the total coding rate could be saved by not coding these segments.

Furthermore, the proposed detection and segmentation algorithm applies to both recorded speech files and real-time speech flows, and does not require any training. For an application to real-time speech flows, the required look-ahead delay is equal to one glottal epoch [7]. In a

speech coder, this would result in an additional algorithmic delay of up to 6 ms, if this look-ahead is not already available.

Finally, the algorithmic complexity of the proposed detection, segmentation and muting system is quite reasonable and compatible with a real time application on a typical microprocessor.

3.10 Conclusion

3.10.1 Summary

This paper has presented a system to detect and segment the burst phase of stop consonants (plosives and affricates). This system relies on the plosion index (PI) proposed in 2014 for detection of the CBT instant, but introduces a modification to avoid late detection. The PI is also used in a new method to segment the burst phase of stop consonants.

In a second part, this detection and segmentation system has been used to evaluate the perceptual relevance of the burst phase of stop consonants. Specifically, a CCR listening test has been conducted to evaluate the impact of muting that burst phase. With a total of twelve expert listeners, and a CMOS score of -0.12, this experiment has shown that these segments can actually be drastically altered at a very limited cost in terms of subjective quality.

Because of their dynamic and unpredictable nature, stop consonants are generally regarded as difficult to encode. The fact that the burst phase of these speech sounds has very little perceptual relevance has interesting implications in the field of speech coding, especially regarding special coding modes and bitrates used for transient sound in several audio codecs. Moreover, on the TIMIT corpus, the burst phase represents respectively around 1.3% and 0.58% of the duration of active speech signals for plosives and affricates (using the proposed segmentation method), which suggests that up to 1.88% of the total coding rate could be saved by simply muting the segmented burst phases rather than coding them (assuming a fixed-rate codec, with a variable framing scheme that perfectly adapts to muted segments and no information being transmitted to signal muted segments).

3.10.2 Possible future work

It has been shown that muting the burst phase of stop consonants has a very limited impact on the subjective quality of a speech signal. Although quality and intelligibility are to some extent correlated, it would be interesting to also evaluate the impact of this operation on intelligibility, for example using a diagnostic rhyme test (DRT) [53].

Moreover, the impact of false positives has not been evaluated in this paper. The impact of the muting alteration on other phones than stop consonants is certainly highly destructive. Finally, the impact of background noise deserves to be studied. Background noise is likely to have negative consequences on the performance of the proposed detection and segmentation method. It is also incompatible with the muting alteration, since the background noise needs to be preserved. Solutions will have to be found to cope with noisy acoustic environments.

3.11 Acknowledgment

The authors wish to thank their lab coworkers from the Université de Sherbrooke for participating in the listening test, as well as T. V. Ananthapadmanabha, A. P. Prathosh and K.V. Vijay Girish from the Indian Institute of Science (Bangalore, India) for providing source code and help understanding their work.

CHAPITRE 4 : CONCLUSION

4.1 Sommaire

Ce mémoire avait pour but d'apporter une plus grande compréhension d'un certain type de sons produits par l'homme : les consonnes plosives et affriquées. L'objectif principal de ce projet de recherche était de proposer et d'évaluer une transformation des sons plosifs consistant à mettre à réduire au silence les phases d'éclatement. Trois objectifs secondaires découlaient de cet objectif principal.

Le premier objectif secondaire consistait à proposer une méthode de détection et de segmentation, afin de rendre la transformation automatique. Ces étapes de détection et de segmentation sont basées sur l'utilisation de l'index de plosion introduit en 2014 [6], et sont présentées dans les parties 3.5 à 3.6. Une modification est apportée cette méthode de détection, afin d'éviter une détection retardée, et la méthode de segmentation présentée est originale.
L'atteinte du premier l'objectif secondaire a nécessité la mise au point d'un système capable d'appliquer de façon automatique la transformation à un signal de parole. Ce système est structuré en plusieurs modules (détection, segmentation, altération) et est présenté en section 3.4. La transformation appliquée consiste à réduire à zéro les segments des phases d'éclatement des plosives, qui présente des caractéristiques de signaux transitoires. Les détails de cette transformation sont présentés en section 3.7.

Le deuxième objectif secondaire consistait à choisir et à réaliser une évaluation perceptuelle afin de mesurer l'impact de la transformation proposée à travers l'objectif principal. À cette fin, un sous ensemble du corpus de phrases TIMIT (présenté en section 3.8) a été sélectionné, et un test perceptuel subjectif de type CCR a été réalisé sur un ensemble de douze auditeurs experts. Le test consistait à attribuer un score comparatif à des paires de phrases, la transformation ayant été appliquée aléatoirement à l'une des deux phrases. Les résultats de ce test d'écoute sont présentés en section 3.9, et révèlent que les segments d'éclatement des plosives peuvent effectivement être drastiquement altérés sans grande conséquences sur la qualité subjective de la parole.

Enfin, le troisième et dernier objectif secondaire consistait à envisager l'intégration d'une telle transformation au sein d'un codeur de parole. En effet, certains codeurs dont VMR-WB et OPUS utilisent des modes spéciaux au voisinage des sons transitoires, afin de bien les modéliser. La transformation proposée pourrait permettre de s'affranchir de tels modes couteux en débits binaires, ou délicats à maintenir suivant le réseau utilisé. Des tests informels présentés en section 3.9.3 indiquent qu'une telle intégration ne génère pas de problèmes par exemple pour les codeurs AMR-WB et OPUS. De plus, les durées des segments d'éclatement dans TIMIT représentant environ 2% du signal de parole utile, il est possible selon certaines conditions d'employer cette transformation sans avoir besoin de coder l'information. Les circonstances permettant cette économie seraient un taux de codage constant, et une taille de trame variable pouvant s'adapter parfaitement aux segments à réduire à zéro.

Les conclusions de ces recherches indiquent qu'une mise à zéro des segments plosifs de parole affecte peu la qualité subjective de la parole. Un tel résultat pourrait se révéler utile en traitement

de la parole, comme par exemple en l'intégrant dans un codeur à des fins de compression, mais également en synthèse vocale.

4.2 Contributions

Chacune des contributions à la recherche de ce mémoire se retrouvent dans l'article de recherche situé au cœur de ce mémoire. Ces contributions abordent différentes facettes de l'étude des plosives.

Tout d'abord, la recherche d'une méthode de détection des transitions fermeture-éclatement pour les plosives a mené à réutiliser les travaux de [6]. La mise en application de ces travaux a cependant donné lieu à un retard dans la détection. Une contribution à l'amélioration cette implémentation a ainsi été proposée en section 3.5.2.

Ensuite, compte tenu du manque de littérature concernant la segmentation des phases d'éclatement des plosives, à partir de leur instant de transition fermeture-éclatement, une nouvelle méthode de segmentation de ces phases a été proposée en section 3.6. Cette nouvelle méthode a ainsi pu être utilisée aux fins de ce mémoire, et a également permis de produire des statistiques sur les durées de ces phases sur la banque de phrases TIMIT [48], en section 3.9.1. Enfin, ce mémoire apporte une réponse à la question de recherche « Quel est l'impact perceptuel d'une mise à zéro des trames transitoires des plosives dans un signal de parole propre ? ». Les résultats de l'évaluation subjective présentés dans l'article en section 3.9.2, et ils suggèrent qu'une telle transformation affecte peu la qualité de la parole. Des pistes d'intégrations de cette transformation à des codeurs de parole connus sont également fournies en section 3.9.3.

4.3 Travaux futurs

Les résultats présentés dans l'article de recherche montrent que la mise à zéro de la phase d'éclatement des plosives a un impact limité sur la qualité subjective d'un signal de parole. Cependant, bien que la qualité et l'intelligibilité soient corrélés jusqu'à un certain point, il pourrait être intéressant d'évaluer l'impact sur l'intelligibilité de cette transformation. Cette évaluation pourrait s'effectuer à l'aide d'un test de rime [53].

Par ailleurs, l'impact de l'application de la transformation sur des faux positifs n'a pas été étudié. En effet, l'impact de la mise à zéro sur d'autres types de sons d'un signal de parole est certainement fortement négatif.

D'autre part, une étude pourrait être réalisée sur l'impact de la présence d'un bruit de fond sur les segments affectés par la transformation. Le fait que le bruit de fond doive être conservé pour préserver la cohérence laisse penser que la transformation proposée serait incompatible avec celui-ci.

Enfin, il pourrait être pertinent de réaliser de nouvelles expériences perceptuelles autour des plosives, avec un nombre conséquent d'auditeurs expert, et la mise en œuvre d'un test de type MUSHRA.

BIBLIOGRAPHIE

[1] K. Honda, « Physiological processes of speech production », dans *Springer Handbook of Speech Processing*, Springer, 2008, p. 7–26.

[2] K. N. Stevens, « Models for production and acoustics of stop consonants », *SST 1992 Proc.*, 1992.

[3] A. Suchato, « Classification of stop consonant place of articulation », Ph.D, Massachusetts Institute of Technology, 2004.

[4] H. M. Hanson et K. N. Stevens, « Modeling stop-consonant releases for synthesis », *J. Acoust. Soc. Am.*, vol. 107, p. 2907, juill. 2011.

[5] D. Byrd, « 54,000 American stops », *UCLA Work. Pap. Phon.*, vol. 83, p. 97–116, 1993.

[6] T. V. Ananthapadmanabha, A. P. Prathosh, et A. G. Ramakrishnan, « Detection of the closure-burst transitions of stops and affricates in continuous speech using the plosion index », *J. Acoust. Soc. Am.*, vol. 135, nº 1, p. 460-471, janv. 2014.

[7] A. P. Prathosh, T. V. Ananthapadmanabha, et A. G. Ramakrishnan, « Epoch Extraction Based on Integrated Linear Prediction Residual Using Plosion Index », *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, nº 12, p. 2471-2480, déc. 2013.

[8] F. Malbos, M. Baudry, et S. Montresor, « Detection of stop consonants with the wavelet transform », dans *Time-Frequency and Time-Scale Analysis, 1994., Proceedings of the IEEE-SP International Symposium on*, 1994, p. 612–615.

[9] J. Keshet, D. Chazan, et B.-Z. Bobrovsky, « Plosive spotting with margin classifiers », dans *INTERSPEECH*, 2001, p. 1637–1640.

[10] T. V. Ananthapadmanabha, K. V. Girish, et A. G. Ramakrishnan, « Detection of transitions between broad phonetic classes in a speech signal », *ArXiv Prepr. ArXiv14110370*, 2014.

[11] A. Esposito, C. E. Ezin, et M. Ceccarelli, « Preprocessing and neural classification of english stop consonants [b, d, g, p, t, k] », dans *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, 1996, vol. 2, p. 1249–1252.

[12] Y. Zheng, M. Hasegawa-Johnson, et S. Borys, « Stop consonant classification by dynamic formant trajectory. », dans *INTERSPEECH*, 2004, p. 396-399.

[13] A. M. A. Ali, J. Van der Spiegel, et P. Mueller, « Acoustic-phonetic features for the automatic classification of stop consonants », *Speech Audio Process. IEEE Trans. On*, vol. 9, nº 8, p. 833–841, 2001.

[14] T. J. Edwards, « Multiple features analysis of intervocalic English plosives », *J. Acoust. Soc. Am.*, vol. 69, nº 2, p. 535-547, févr. 1981.

[15] A. Madsack, G. Dogil, S. Uhlich, Y. Zeng, et B. Yang, « Phone-based Plosive Detection », University of Stuttgart, Tech. rep., 2009.

[16] V. Patil et P. Rao, « Acoustic features for detection of aspirated stops », dans *Communications (NCC), 2011 National Conference on*, 2011, p. 1–5.

[17] J. Nieuwenhuijse, R. Heusens, et E. F. Deprettere, « Robust exponential modeling of audio signals », dans *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, 1998*, 1998, vol. 6, p. 3581-3584 vol.6.

[18] R. Boyer et K. Abed-Meraim, « Audio transients modeling by damped & delayed sinusoids (DDS) », dans *Proc. of ICASSP*, 2002, vol. 2, p. II-1729-II-1732.

[19] R. Boyer et K. Abed-Meraim, « Damped and delayed sinusoidal model for transient signals », *IEEE Trans. Signal Process.*, vol. 53, nº 5, p. 1720-1730, 2005.

[20] G. P. Kafentzis, Y. Pantazis, O. Rosec, et Y. Stylianou, « An extension of the adaptive Quasi-Harmonic Model », dans *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, p. 4605-4608.
[21] M. Kepesi, « Adaptive Chirp-based time-frequency analysis of speech signals », *Speech Commun.*, vol. 48 num. 5, p. 474-492, 2006.
[22] L. Weruaga et M. Képesi, « The fan-chirp transform for non-stationary harmonic signals », *Signal Process.*, vol. 87, nº 6, p. 1504-1522, juin 2007.
[23] G. P. Kafentzis et Y. Stylianou, « HIGH-RESOLUTION SINUSOIDAL MODELING OF UNVOICED SPEECH ».
[24] F. X. Nsabimana et U. Zolzer, « Transient encoding of audio signals using dyadic approximations », dans *Proc. 10th International Conference on Digital Audio Effects (DAFx-07), Bordeaux, France*, 2007.
[25] « Perfectionnements aux systèmes de signalisation », FR833929, 18-juin-1937.
[26] Xiph.Org Foundation, « FLAC - Free Lossless Audio Codec », 20-juill-2001. [En ligne]. Disponible à: https://xiph.org/flac/index.html.
[27] Apple Inc., « ALAC - Apple Lossless Audio Codec », 28-avr-2004. [En ligne]. Disponible à: http://alac.macosforge.org/.
[28] Xiph.Org Foundation, « Ogg Vorbis », 08-mai-2000. [En ligne]. Disponible à: https://xiph.org/vorbis/.
[29] « ISO/IEC 11172-3:1993 - Information technology -- Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s -- Part 3: Audio », *ISO*. [En ligne]. Disponible à: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=22412.
[30] « ISO/IEC 13818-7:1997 - Information technology -- Generic coding of moving pictures and associated audio information -- Part 7: Advanced Audio Coding (AAC) », *ISO*. [En ligne]. Disponible à: http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?csnumber=25040.
[31] T. Cervera et W. A. Ainsworth, « Effects of preceding noise on the perception of voiced plosives », *Acta Acust. United Acust.*, vol. 91, nº 1, p. 132–144, 2005.
[32] B. H. Repp, « Closure Duration and Release Burst Amplitude Cues To Stop Consonant Manner and Place of Articulation », *Lang. Speech*, vol. 27, nº 3, p. 245-254, juill. 1984.
[33] A. Kapoor et J. B. Allen, « Perceptual effects of plosive feature modification », *J. Acoust. Soc. Am.*, vol. 131, nº 1, p. 478-491, janv. 2012.
[34] S. Voran, « Estimation of speech intelligibility and quality », dans *Handbook of Signal Processing in Acoustics*, Springer, 2008, p. 483–520.
[35] P. ITU-T RECOMMENDATION, « Subjective video quality assessment methods for multimedia applications », 1999.
[36] I. T. S. Sector, « ITU-T Recommendation Z. 120 », *Message Seq. Charts MSC96*, 1996.
[37] L. Malfait, J. Berger, et M. Kastner, « P.563—The ITU-T Standard for Single-Ended Speech Quality Assessment », *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, nº 6, p. 1924-1934, nov. 2006.
[38] International Telecommunication Union, « ITU-T Recommendation P.800. Methods for subjective determination of transmission quality ». mars-1996.

[39] R. McAulay et T. F. Quatieri, « Speech analysis/Synthesis based on a sinusoidal representation », *IEEE Trans. Acoust. Speech Signal Process.*, vol. 34, nº 4, p. 744-754, août 1986.
[40] B. S. Atal et S. L. Hanauer, « Speech Analysis and Synthesis by Linear Prediction of the Speech Wave », *J. Acoust. Soc. Am.*, vol. 50, nº 2B, p. 637-655, août 1971.
[41] M. R. Schroeder et B. S. Atal, « Code-excited linear prediction (CELP): High-quality speech at very low bit rates », dans *Proc. of ICASSP*, 1985, vol. 10, p. 937–940.
[42] K. K. Iwai, « Pre-echo detection & reduction », Thesis, Massachusetts Institute of Technology, 1994.
[43] K. L. Poort, « Stop consonant production : an articulation and acoustic study », Thesis, Massachusetts Institute of Technology, 1995.
[44] P. J. Jackson, « Characterisation of plosive, fricative and aspiration components in speech production », Thesis, University of Southampton, 2000.
[45] P. Niyogi et M. M. Sondhi, « Detecting stop consonants in continuous speech », *J. Acoust. Soc. Am.*, vol. 111, nº 2, p. 1063-1076, févr. 2002.
[46] R. Dokku et R. Martin, « Detection of stop consonants in continuous noisy speech based on an extrapolation technique », dans *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, 2012, p. 2338-2342.
[47] G. Hu et D. Wang, « Separation of stop consonants », dans *Proc. of ICASSP*, 2003, vol. 2, p. 749-752.
[48] John Garofolo, Lori Lamel, William Fisher, Jonathan Fiscus, David Pallett, Nancy Dahlgren, et Victor Zue, « TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1 ». Philadelphia: Linguistic Data Consortium, 1993.
[49] J.-M. Valin, K. Vos, et T. Terriberry, « Definition of the Opus Audio Codec », RFC6716, sept. 2012.
[50] J.-M. Valin, T. B. Terriberry, C. Montgomery, et G. Maxwell, « A high-quality speech and audio codec with less than 10-ms delay », *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, nº 1, p. 58–67, 2010.
[51] M. Jelinek, R. Salami, S. Ahmadi, B. Bessette, P. Gournay, et C. Laflamme, « On the architecture of the CDMA2000® variable-rate multimode wideband (VMR-WB) speech coding standard », dans *Proc. of ICASSP*, 2004, vol. 1, p. 281-284.
[52] J. Sjoberg, M. Westerlund, A. Lakaniemi, et Q. Xie, « RTP Payload Format and File Storage Format for the Adaptive Multi-Rate (AMR) and Adaptive Multi-Rate Wideband (AMR-WB) Audio Codecs », RFC4867, avr. 2007.
[53] S. L. Greenspan, R. W. Bennett, et A. K. Syrdal, « An evaluation of the diagnostic rhyme test », *Int. J. Speech Technol.*, vol. 2, nº 3, p. 201-214, sept. 1998.

ANNEXES

Annexe A : Procédure de test CCR et implémentation

Directives du test

DIRECTIVES POUR LES AUDITEURS

« Évaluation d'une altération plosives »

Dans la présente expérience, vous écouterez des paires d'échantillons de parole qui ont subi ou non une altération de leurs segments plosifs. Vous écouterez ces échantillons par l'intermédiaire du casque audio qui se trouve en face de vous.

Vous entendrez une phrase, suivie d'une pause, puis une autre phrase. Vous évaluerez la qualité de la seconde phrase par rapport à celle de la première.

Vous devez écouter attentivement chaque paire d'échantillons. Vous pouvez réécouter les échantillons à l'aide des boutons à cet effet. Quand vous avez analysé les échantillons, donnez votre opinion sur la qualité du second échantillon par rapport à celle du premier au moyen de l'échelle suivante:

La qualité du second échantillon par rapport à celle du premier est la suivante:

3 Bien meilleure

2 Meilleure

1 Légèrement meilleure

0 A peu près équivalente

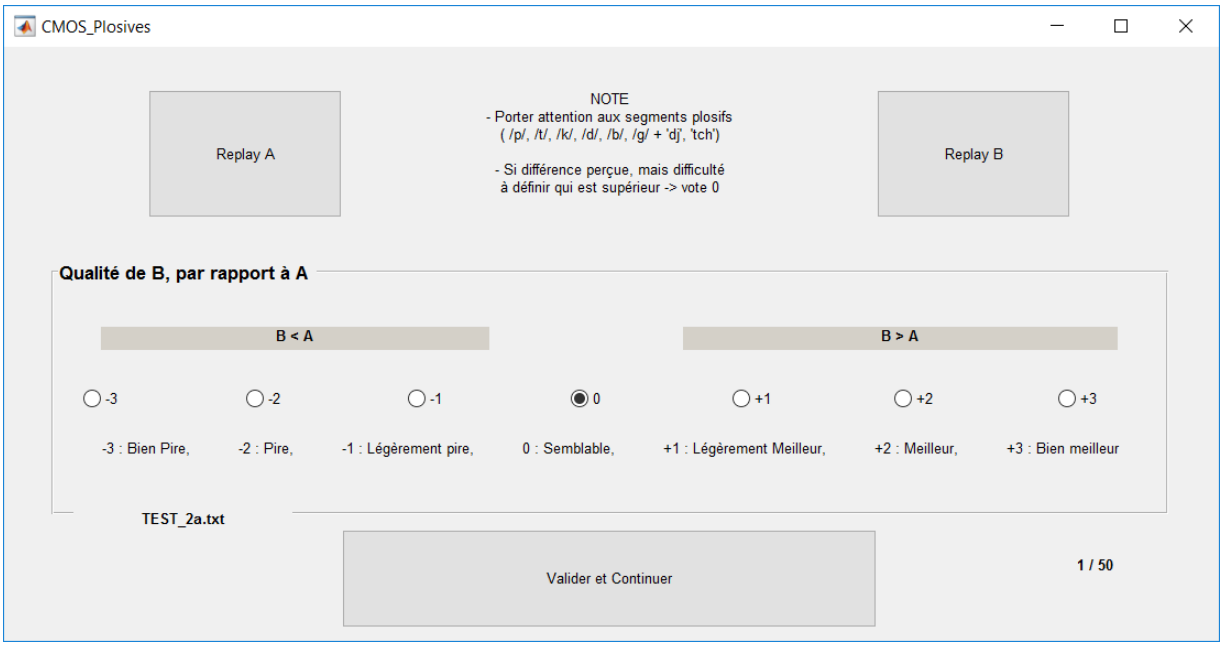
−1 Légèrement moins bonne

−2 Moins bonne

−3 Beaucoup moins bonne

Vous disposerez de tout votre temps pour enregistrer votre réponse en sélectionnant la valeur correspondant à votre choix, puis en appuyant sur le bouton de validation. Il y aura ensuite une brève pause avant la présentation de la prochaine paire de phrases.

Interface du programme de test subjectif



Structure des fichiers test

```
Test  name : TEST_1z_1019_a
Audio path : ..\wav1019\

Path_Reference Path_Test

DR5_MAHH0_SA1.wav DR5_MAHH0_SA1_1z_20151019150923242.wav
DR6_MJDH0_SA2.wav DR6_MJDH0_SA2_1z_20151019150935370.wav
DR5_FASW0_SI1550.wav DR5_FASW0_SI1550_1z_20151019150908812.wav
DR2_FJRE0_SI1116.wav DR2_FJRE0_SI1116_1z_20151019150919783.wav
DR5_MAHH0_SI1924.wav DR5_MAHH0_SI1924_1z_20151019150852383.wav
DR6_FLNH0_SI584.wav DR6_FLNH0_SI584.wav
DR4_MDRM0_SA1.wav DR4_MDRM0_SA1_1z_20151019150926234.wav
DR5_MAHH0_SI1924.wav DR5_MAHH0_SI1924.wav
DR6_MJDH0_SA1.wav DR6_MJDH0_SA1_1z_20151019150920988.wav
DR4_FEDW0_SI1653.wav DR4_FEDW0_SI1653_1z_20151019150914524.wav
DR4_MBNS0_SA2.wav DR4_MBNS0_SA2_1z_20151019150938641.wav
DR2_MDLD0_SI2173.wav DR2_MDLD0_SI2173_1z_20151019150901190.wav
DR6_FLNH0_SI1214.wav DR6_FLNH0_SI1214.wav
DR6_MPAM1_SA1.wav DR6_MPAM1_SA1_1z_20151019150922309.wav
DR5_MCMB0_SA2.wav DR5_MCMB0_SA2_1z_20151019150938100.wav
DR2_MDBB0_SI1825.wav DR2_MDBB0_SI1825_1z_20151019150900110.wav
DR6_MJDH0_SA2.wav DR6_MJDH0_SA2.wav
```

Structure des fichiers résultats

```
Test  name : TEST_1z_1019_a
Audio path : ..\wav1019\

Path_Reference Path_Test Score

DR5_MAHH0_SA1.wav DR5_MAHH0_SA1_1z_20151019150923242.wav -1
DR6_MJDH0_SA2.wav DR6_MJDH0_SA2_1z_20151019150935370.wav 0
DR5_FASW0_SI1550.wav DR5_FASW0_SI1550_1z_20151019150908812.wav -1
DR2_FJRE0_SI1116.wav DR2_FJRE0_SI1116_1z_20151019150919783.wav 0
DR5_MAHH0_SI1924.wav DR5_MAHH0_SI1924_1z_20151019150852383.wav -1
DR6_FLNH0_SI584.wav DR6_FLNH0_SI584.wav 1
DR4_MDRM0_SA1.wav DR4_MDRM0_SA1_1z_20151019150926234.wav 0
DR5_MAHH0_SI1924.wav DR5_MAHH0_SI1924.wav 0
DR6_MJDH0_SA1.wav DR6_MJDH0_SA1_1z_20151019150920988.wav 1
DR4_FEDW0_SI1653.wav DR4_FEDW0_SI1653_1z_20151019150914524.wav 1
DR4_MBNS0_SA2.wav DR4_MBNS0_SA2_1z_20151019150938641.wav -2
DR2_MDLD0_SI2173.wav DR2_MDLD0_SI2173_1z_20151019150901190.wav 2
DR6_FLNH0_SI1214.wav DR6_FLNH0_SI1214.wav 2
DR6_MPAM1_SA1.wav DR6_MPAM1_SA1_1z_20151019150922309.wav 0
DR5_MCMB0_SA2.wav DR5_MCMB0_SA2_1z_20151019150938100.wav 0
DR2_MDBB0_SI1825.wav DR2_MDBB0_SI1825_1z_20151019150900110.wav 1
DR6_MJDH0_SA2.wav DR6_MJDH0_SA2.wav 0
```

Annexe B : Détails sur les résultats

Résultats individuels des auditeurs

	All	Listener 1	Listener 2	Listener 3	Listener 4	Listener 5	Listener 6	Listener 7	Listener 8	Listener 9	Listener 10	Listener 11	Listener 12
Control pairs fail	81	6	10	15	4	2	0	1	3	17	7	2	14
CMOS	-0,11825	-0,11	-0,13	-0,3	-0,0875	-0,14	-0,088	0,025	-0,225	-0,088	0,0125	-0,25	-0,038
CI at 95%	[-0,15 -0,08]	[-0,3 0,07]	[-0,27 0,023]	[-0,51 -0,086]	[-0,18 0,001]	[-0,26 -0,01]	[-0,15 -0,02]	[-0,025 0,075]	[-0,35 -0,098	[-0,27 0,10]	[-0,17 0,20]	[-0,38 -0,12]	[-0,19 0,11]
T Test		1	0	0	1	0	1	1	0	1	0	0	1	0