Détection automatique des instants de fermeture glottale dans les voix pathologiques

Quentin Robin

Quentin ROBIN
filière SICOM - 2016/2017

Rapport de stage

Détection automatique des instants de fermeture glottale dans les voix pathologiques

Responsable du stage:

Dr. Khalid Daoudi, équipe GeoStat (khalid.daoudi@inria.fr)

Remerciements

Tout d'abord, je tiens à remercier l’INRIA de bordeaux et en particulier l’équipe GeoStat pour m’avoir accueilli durant ces 3 mois de stage.

Je tiens à remercier vivement mon maître de stage, Dr. Khalid Daoudi, chercheur au sein de l'équipe GeoStat, pour son accueil, le temps passé ensemble et le partage de son expertise au quotidien. Grâce aussi à sa confiance, j'ai pu avancer dans ma mission. Il fut d'une aide précieuse dans les moments les plus délicats.

Faire mon stage de deuxième année dans votre centre de recherche a été un plaisir, j'ai pu apprendre beaucoup grâce à vous, et j'ai surtout été conforté dans mon projet professionnel, ce qui est un progrès pour mon cursus d’ingénieur.

Glossaire

GCI: Glottal closure : instant (Instant de fermeture glottal) instant ou les cordes vocales entrent en contact

EGG: Electroglottographie :dispositif permettant d'obtenir une information concernant la position des cordes vocales, il mesure l'impédance au niveau du larynx

Liste des figures

Introduction

Ce rapport constitue une trace écrite du stage de deuxième année de la formation d'ingénieur en traitement du signal de l'institut national polytechnique de Grenoble école Phelma filière Signal Image Communication Multimédia. Le stage se déroule à l'institut National de Recherche en informatique et en automatique (INRIA) de Bordeaux au sein de l'équipe de recherche Geostat. Cette équipe travaille sur la Géométrie et les statistiques dans les données d'acquisition.

Le traitement de la parole est une thématique importante des sciences de l'ingénieur alliant traitement du signal et connaissances médicales. La parole peut être vue comme un vecteur d'information. Cette information est d'une part, celle utilisée par les individus pour communiquer entre eux, c'est le rôle des mots et du langage, et d'autre part, vue comme une information sur l'individu lui même, par exemple si c'est un homme ou une femme, et bien d'autres informations.

Le travail produit pendant le stage est une petite partie d'une large étude sur les voix parkinsoniennes ayant pour objectif d'utiliser la parole pour aider au diagnostic différentiel de la maladie. La maladie de parkinson introduit de nombreux trouble moteur notamment sur la parole. Le stage porte sur l'étude de la vibration des cordes vocales, notamment les instants où elles entrent en contact, appelé Glottal Closure Instant (GCI). L'activité des cordes vocales peut être observée grâce à des électrodes placées sur le cou ou extraite depuis le signal de parole.

Plusieurs méthodes permettent d'identifier ces instants directement depuis le signal de parole. Par exemple la méthode SEDREAMS (1) développée par Thomas Drugman en 2012. Une partie du stage est consacrée à tester ces méthodes sur des bases de données de voix non-pathologiques, puis pathologiques.
Pour évaluer les performances de chacune des méthodes, les GCIs extraits sont comparés avec des GCIs de référence considérés comme vérité terrain. Ainsi, la précision et le taux de bonne et de mauvaise détection peuvent être estimés pour chaque méthode.

Il est indispensable d'avoir une vérité terrain, c'est à dire, la vraie position des instants de fermeture glottale. La vérité terrain provient généralement de l'electroglottographie.

Le texte de ce document est découpé en trois sections :
La première présentera un état de l'art des méthodes existantes et actuelles, c'est à dire, des méthodes publiées dans la littérature depuis moins de 10 ans.
Les deux dernières sections présenteront une nouvelle méthode d'extraction de la vérité terrain développée dans le cadre de ce stage, ainsi que l'évaluation des performances de ces méthodes.

1 Etat de l’art

1.1 Sons voisés non-voisés

Dans un signal de parole, on distingue 2 types de sons: les sons voisés et non-voisés. Les sons voisés correspondent à la prononciation des voyelles (par exemple, “a”), la production de ce type de sons s’accompagne d’une vibration des cordes vocales. Les sons non-voisés, généralement les consonnes, sont produits uniquement par la bouche, ils n’engagent donc pas de mouvement des cordes vocales. Ce rapport traite des instants de fermeture glottale. Ces instants ont donc un sens uniquement pour des sons voisés.
Les sons voisés ont la propriété d’être quasi périodiques. Une période sur le signal de parole correspond à un cycle glottal complet, c’est à dire, à une ouverture et fermeture de la glotte.

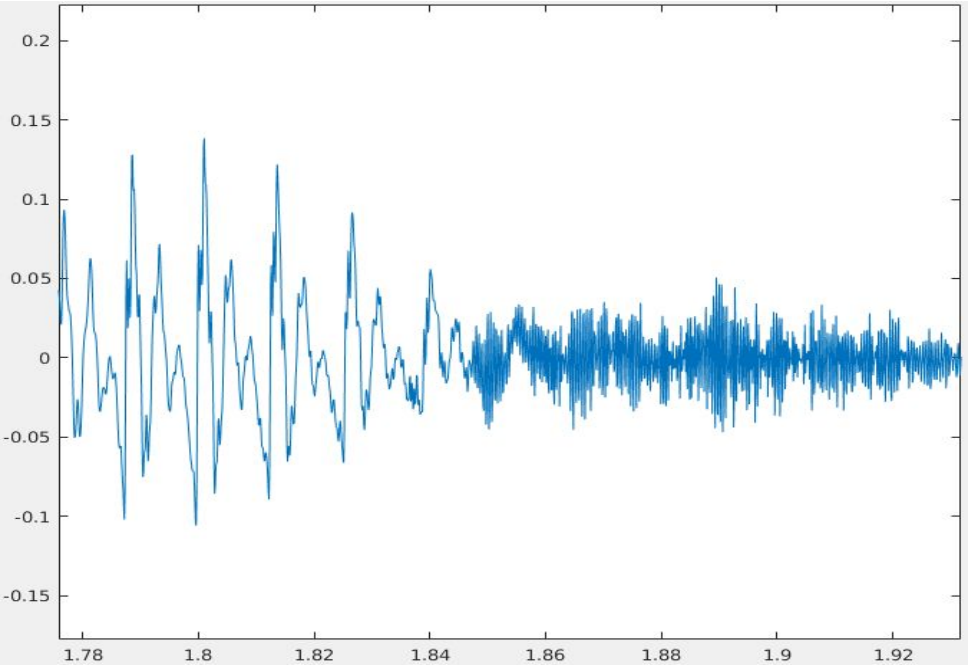


Figure 1 :Signal de parole pour un enregistrement de la base KED

Sur la figure ci-dessus voici le signal d’un son voisé sur la première partie de la courbe, suivi d’un son non-voisé sur la seconde partie de la courbe.

1.2 Instant de fermeture glottale

Voici représenté ci-dessous un cycle glottal complet. Chaque cycle est caractérisé par 2 phases: une ouverture de la glotte (figure 2) et une fermeture (figure 8). L’instant de

fermeture glottale (GCI) est l'instant ou les plis vocaux entrent en contact.

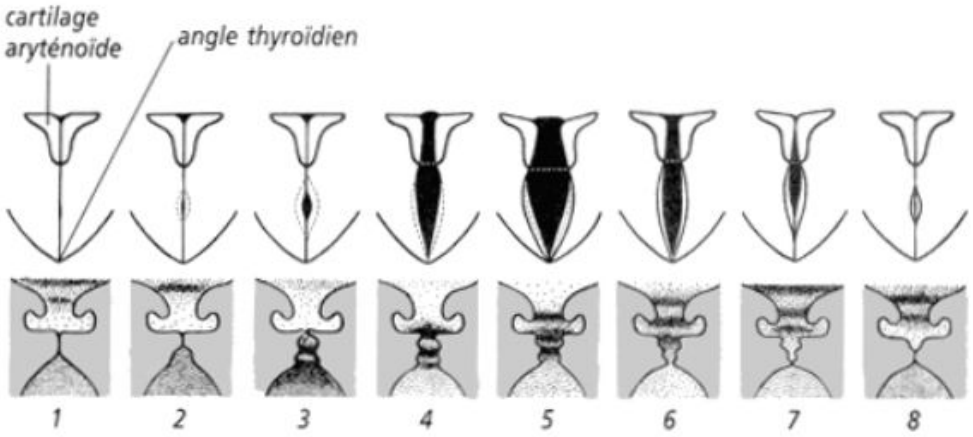


Figure 1.13 – Schéma d'un cycle complet de vibration des plis vocaux. Les figures du haut représentent une vue supérieure de la glotte et les figures du bas une vue sagittale de la glotte (Ormezzano 2000 :103) (d'après Hirano 1981).

Figure 2 : Schéma glottal fermeture / ouverture
Extrait depuis https://tel.archives-ouvertes.fr/tel-00817694/document

Un cycle glottal correspond à une période du signal de parole.
Les instants de fermeture glottal sont utilisés dans plusieurs applications du traitement de la parole: analyse pitch-synchrone, modification de la voix, synthèse, réverbération...etc

1.3 GCIs depuis le signal de parole

Les instants de fermeture glottale peuvent être trouvés depuis le signal de parole. Plusieurs méthodes ont été développées pour réaliser cette tâche.

1.3.1 SEDREAMS

La méthode SEDREAMS développée par Thomas Drugman est une méthode de détection performante et récente. Cette méthode est basée sur l'étude de la prédiction linéaire résiduelle. Voici ci-dessus des courbes et un schéma illustrant cette méthode.

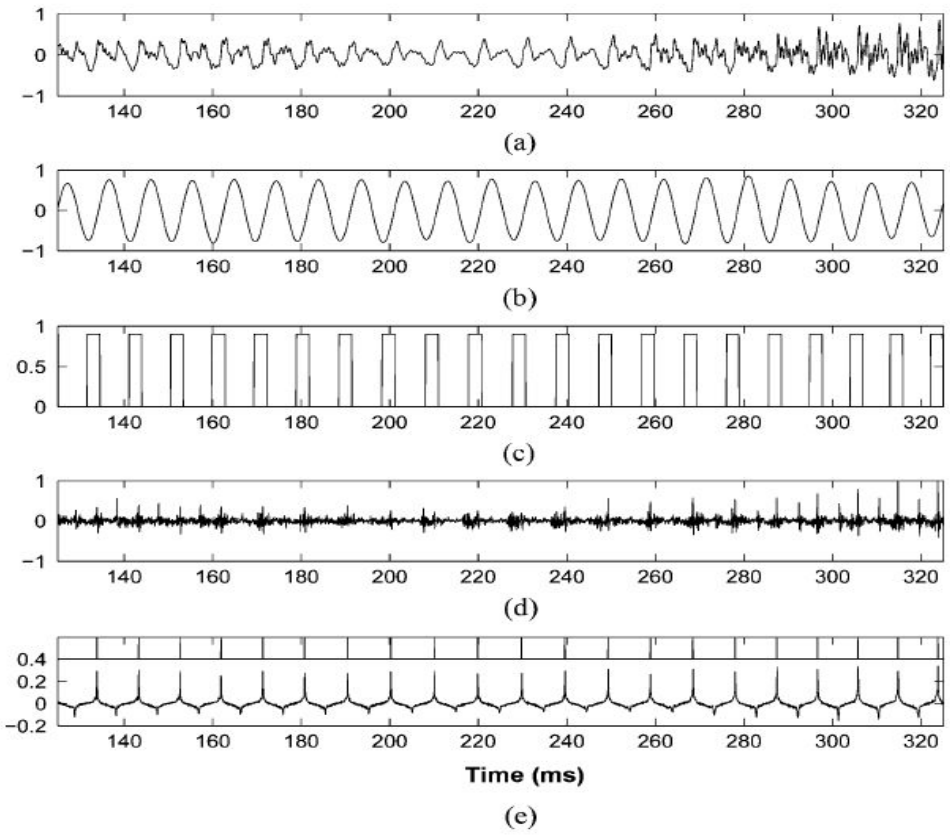
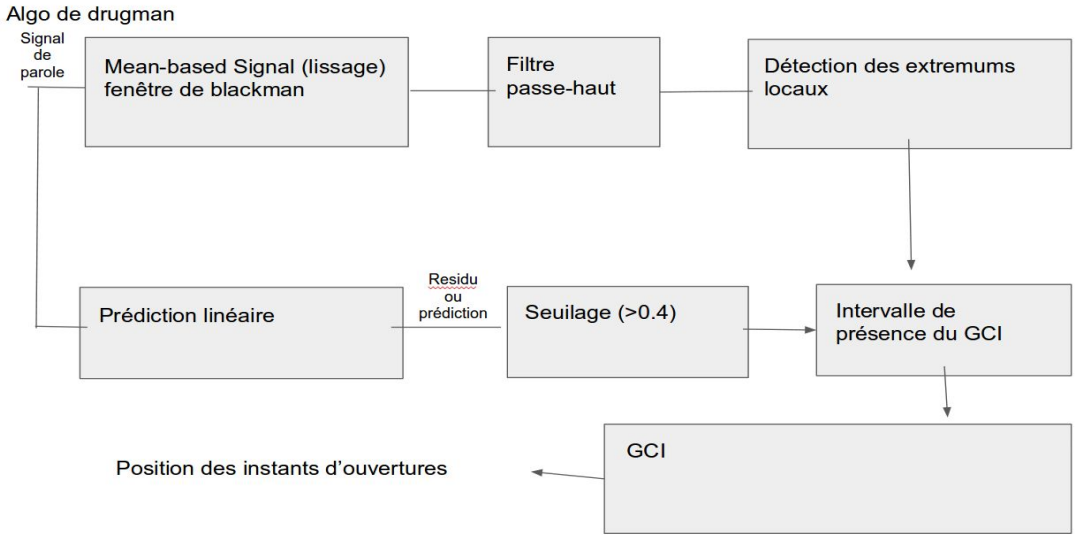


Fig. 3. Illustration of GCI detection using the SEDREAMS algorithm on a segment of voiced speech. (a) The speech signal. (b) The mean-based signal. (c) Intervals of presence derived from the mean-based signal. (d) The LP residual signal. (e) The synchronized dEGG with the GCI positions located by the SEDREAMS algorithm.

Depuis publication de Drugman



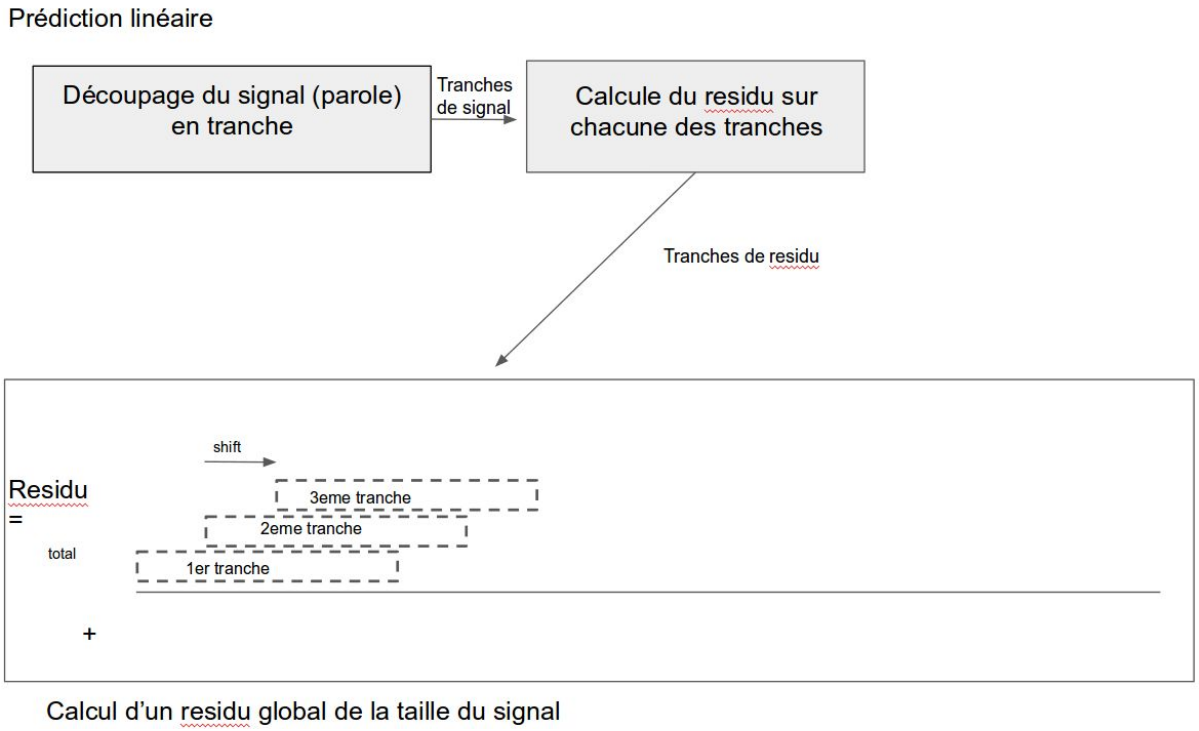


Figure 3 : Schéma méthode SEDREAMS

Les instants de fermeture glottal sont localisés sur les extrema du signal résiduel.

1.3.2 YAGA

YAGA (9) est une autre méthode d'extraction des GCIs depuis le signal parole. Cette méthode est développé par Mark R. P. Thomas, Jon Gudnason et Patrick A. Naylor

cf. https://spiral.imperial.ac.uk/bitstream/10044/1/15494/2/IEEE%20Transactions%20on%20Audio%20Speech%20and%20Language%20Processing_20_1_2012.pdf

1.4 Vérité terrain EGG

Pour pouvoir valider ces méthodes d'extraction des GCIs depuis le signal de parole, il faut avoir une vérité terrain. La vérité terrain c'est la véritable position des GCIs. Les cordes vocales étant situées à l'intérieur du corps humain, elles sont difficilement observables. La vérité terrain est donc délicate à être obtenue. La solution la plus couramment employée consiste à mesurer l'impédance sur le cou des patients. Il s'agit d'electroglottographie.

Sur chacune des bases d'enregistrement étudiées, simultanément à l'enregistrement du signal de parole, un signal en provenance d'électrodes situées sur le cou est enregistré. Ce signal appelé EGG (ElectroGlottoGraphe) nous renseigne sur les mouvements des cordes vocales notamment des phases de fermeture et d'ouverture de la glotte. Le dispositif est appelé un electroglottographe. Il mesure l'impédance électrique à haute fréquence entre 2 électrodes, les mouvements glottaux sont caractérisés par des variations d'impédance.

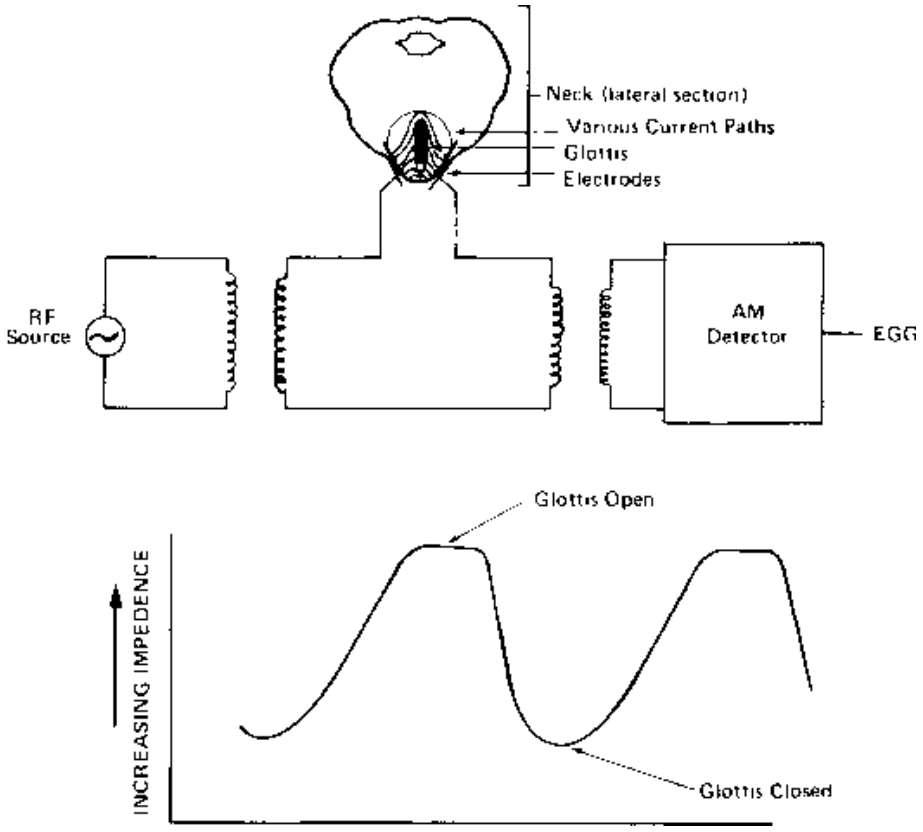


Figure 4 : Schéma electroglottographie

Extrait depuis http://www2.ims.uni-stuttgart.de/EGG/frmst2.htm (The principle of the EGG device (from Childers & Krishnamurthy, 1985:133))

Le signal EGG contient l'information sur les instants de fermeture glottal. La plupart des travaux sur les instants de fermeture glottale le signal EGG est utilisé pour établir la vérité terrain. Il existe plusieurs méthodes pour extraire automatiquement les GCIs depuis le signal EGG.

1.4.1 dEGG

Dans les publications de Vahid Khanagha et Thomas Drugman ainsi que dans la plupart des travaux concernant les GCIs, la vérité terrain est extraite depuis la dérivé du signal EGG, c'est une méthode appelée dEGG.

Voici sur la figure ci-dessous un exemple de signal EGG en bleu et dEGG en rouge. Durant la phase de fermeture glottale l'impédance mesurée par électroglottographie est forte et sur les phases d'ouverture glottale l'impédance est faible. Ce qui nous intéresse est de déterminer la frontière entre une période d'ouverture et de fermeture de la glotte, c'est à dire l'instant de fermeture glottale. Sur les signaux, on observe une variation forte de l'impédance au moment de la fermeture glottale. Les méthodes de détection basées sur la dérivée du signal EGG définissent l'instant de fermeture glottale comme étant l'instant qui maximise la dérivée.

Mathématiquement:

$$t_{GCI} = Max\left(\frac{dEGG}{dt}\right) \quad t \in T0 \text{ avec } T0 \text{ période du cycle glottal}$$

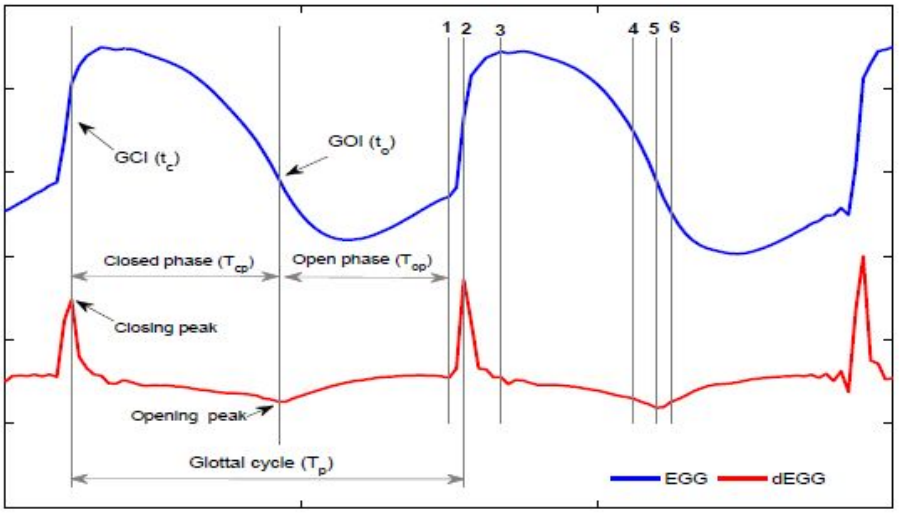


Fig. 1. The EGG and derivative of the EGG (dEGG). (1)-(3) closing phase; (3)-(4) closed phase; (4)-(6) opening phase; (6)-(1): open phase [22]-[27].

Figure 5 : Courbe signal EGG et dEGG

Issue de Effective Glottal Instant Detection and Electroglottographic Parameter Extraction for Automated Voice Pathology Assessment Pranav S. Deshpande and M. Sabarimalai Manikandan, Member, IEEE

Ces instants sont détectés automatiquement par un seuillage et une détection de pic.

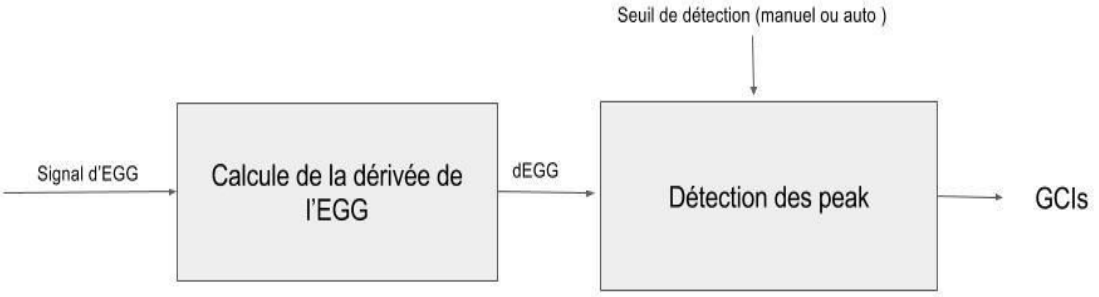


Figure 6 : Schéma résumant la méthode dEGG

Sur les exemples ci-dessus les pics sont dirigé vers le haut. Ce n'est pas le cas pour tous les enregistrements EGG. Les signaux EGG de la base CMU_Artic ne respecte pas cette convention. Les pics sont dirigé vers le bas, c'est l'opposé du signal EGG qui est enregistré.

1.4.2 SIGMA

SIGMA (8) est une autre méthode de détection des GCIs sur l'EGG, le principe de la méthode n'est pas détaillé ici. Deux exécutions de la méthode SIGMA ne donneront pas exactement les mêmes GCIs. Cependant toutes les réalisations sont relativement proches les unes des autres.

2 Une nouvelle méthode d'extraction de la vérité terrain EGG/dEGG

Dans la plupart des publications de recherche publiées à ce jour la vérité terrain est extraite depuis dEGG. Cette vérité terrain n'est pas mise en doute dans les publications.

La méthode dEGG semble relativement bien fonctionner pour les bases CMU_Artic (KED, JMK, BDL et SLT).
Jusqu'à présent aucune publication n'exploite les enregistrements de l'université de Saarland en Allemagne. Cette base est particulièrement intéressante car elle contient des enregistrements de voix pathologique. Un paragraphe détaillant les bases se trouve en dernière partie de ce rapport.
Voici trois problématiques rencontrées avec la base allemande concernant l'extraction de la vérité terrain par la méthode dEGG:

- Un seuil commun à tous les enregistrements semble impossible à trouver ou alors inefficace
- Les signaux sont trop bruités pour estimer de la dérivée du signal
- Certains instants d'enregistrement sont corrompus

Ces problématiques peuvent être étendues aux autres bases.

L'idée est d'écrire une méthode d'extraction de la vérité terrain efficace sur cette base puis de vérifier la robustesse de cette méthode en la testant sur les autres bases. Voici les détails de cette méthode étape par étape.

2.1 Pré-traitement pour supprimer les outliers

Pour quelques signaux issus de la base allemande certains instants de l'enregistrement sont aberrants. C'est ce qu'on remarque par exemple sur le signal 1101-a_lhl-egg.wav. Voici l'enregistrement de EGG de ce signal. Il s'agit de l'enregistrement d'une femme de 18 ans n'ayant aucun trouble de la parole.

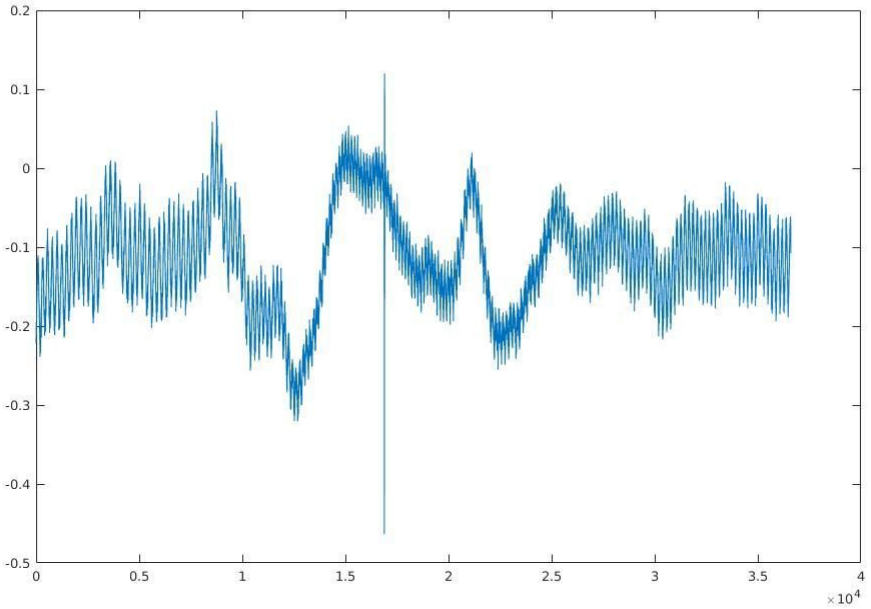


Figure 7: Signal d'EGG 1101-a_lhl-egg.wav

Entre l'échantillon 15k et 20k, on observe un pic aberrant. Pour ne pas compromettre le calcul de la dérivée ou d'autres analyses sur ce signal, la zone aberrante doit être retirée. Pour traiter le problème sans dénaturer le signal (suppression de la tranche, ajout de zéros etc...), la fonction de matlab hampel est utilisée. Cette fonction se charge de détecter la zone aberrante en recherchant les valeurs du signal très supérieures à l'écart type, puis filtre les zones problématiques du signal.

Voici le signal précédent zoomé sur la tranche problématique, en bleu le signal d'origine et en orange le signal traité par la fonction hampel de matlab.

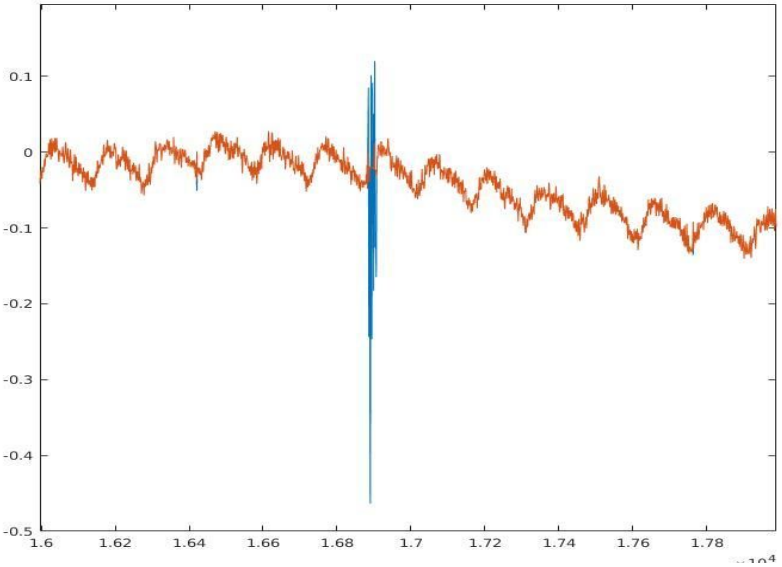


Figure 8 : Signal d'EGG 1101-a_lhl-egg.wav avec et sans suppressions de l'outlier

Les outliers sont donc correctement retirés du signal. La base allemande contient peu d'outlier, moins de 1% des enregistrement sont concernés.

2.2 Filtrage du signal EGG pour réduire le bruit

Voici ci-dessous l'enregistrement d'un EGG partie haute (843-i_h-egg.wav) et la dérivée de l'EGG partie basse de la figure.

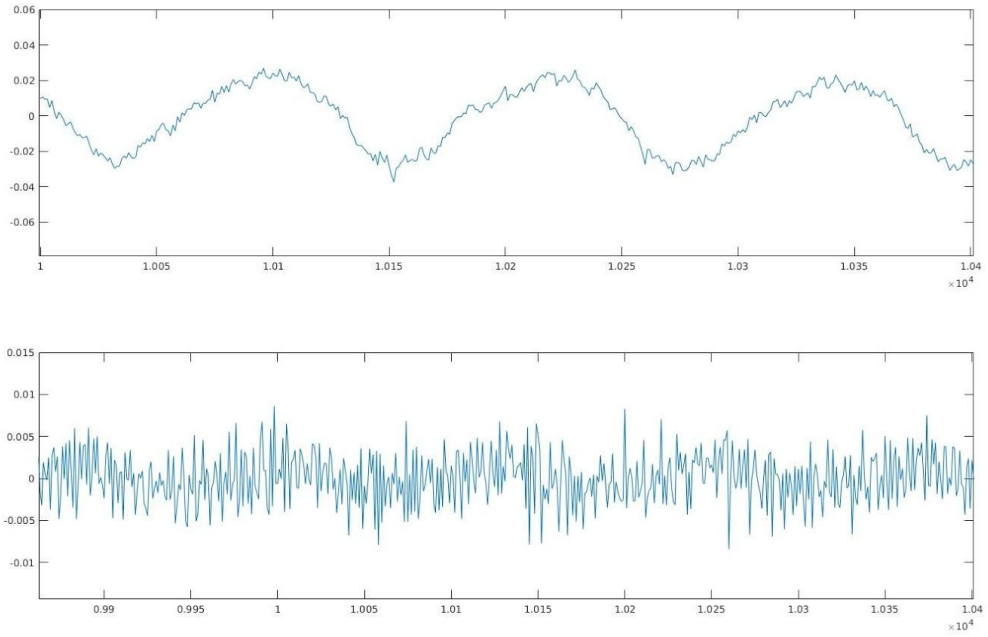


Figure 9 : Signal d'EGG 843-i_h-egg.wav en haut et dérivée du signal en bas

Les signaux EGG issus de la base allemande sont plus bruités que ceux de la base KED. La base allemande est échantillonnée à 50 kHz contre 16 kHz pour la majorité des bases d'enregistrement. Avant l'échantillonnage, les signaux sont filtrés pour éviter les effets de repliement spectral. Ce filtrage permet aussi de limiter la bande du bruit et donc aussi sa puissance. Ici, la fréquence d'échantillonnage est élevée (50 kHz), c'est pour cette raison qu'on observe plus de bruit sur le signal.

Pour la base allemande on a donc une bande de shannon jusqu'à 25kHz. Or l'information recherchée se situe entre 0 et 8 kHz. La bande de fréquence [8 kHz - 25 kHz] ajoute du bruit au signal sans apporter d'information à priori. De plus ce bruit sur le signal d'EGG compromet le calcul de la dérivée (cf. bas de la figure ci-dessus).

$$EGG_{bruité} = EGG_{non-bruité} + bruit$$
$$dEGG_{bruité} = d(EGG_{non-bruité} + bruit) = dEGG_{non-bruité} + d(bruit)$$

Même si le signal est prépondérant sur le bruit, la dérivée du bruit risque d'être prépondérante sur la dérivée du signal.

Le signal va être filtré par un filtre passe bande, dont la bande passante est comprise entre 30 Hz et 8 kHz. Le filtre aura pour effet de limiter le bruit et de couper la composante continue.

Voici le résultat du filtrage:

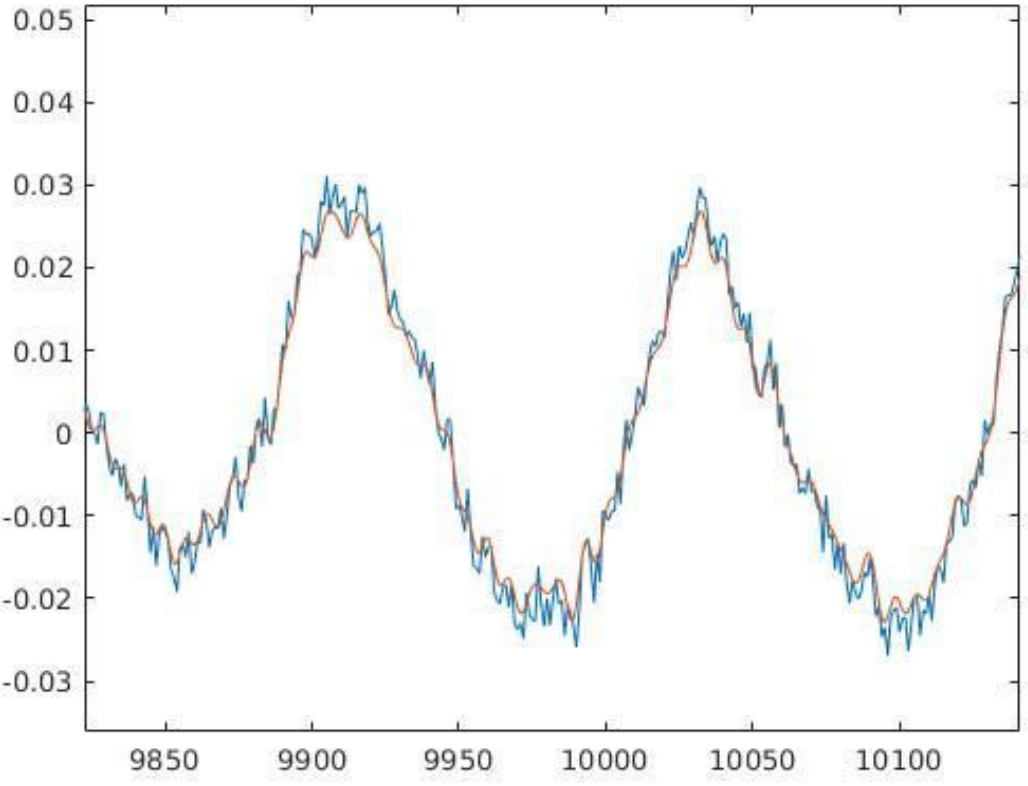


Figure 10 : Signal 843-i_h-egg.wav en bleue et signal filtré en orange

Filtre RIF d'ordre 48 de bande passante [30 Hz - 8k Hz] non causale. RIF pour ne pas avoir d'erreur sur la phase et non causale pour ne pas avoir de retard (de plus le traitement n'est pas fait en temps réel).

Toujours pour 843-i_h-egg.wav voici la dérivée du signal EGG filtré en orange et EGG non-filtré en bleu.

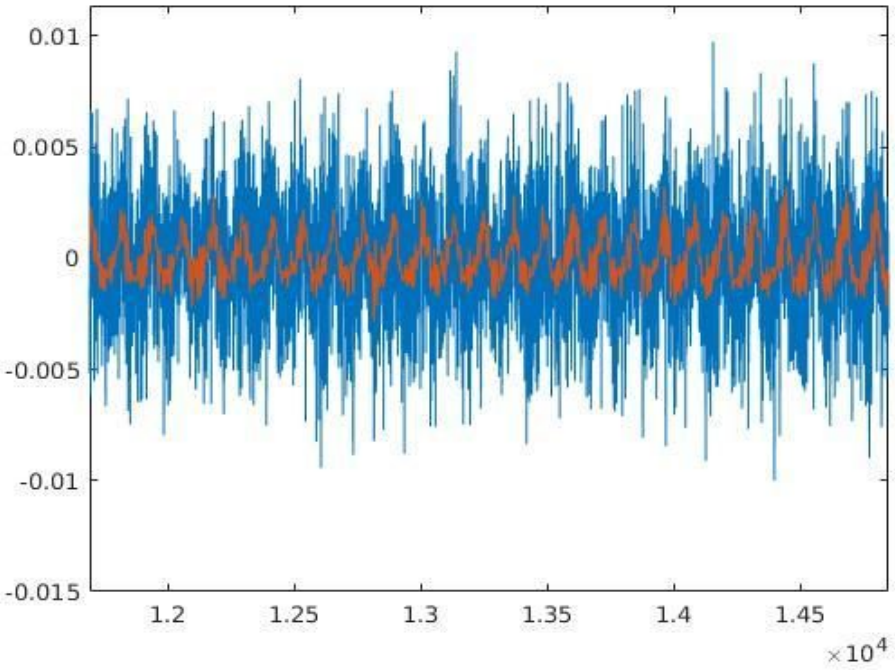


Figure 11 : Dérivée du signal 843-i_h-egg.wav en bleue et dérivée sur signal filtré en orange

Le filtrage permet donc une meilleur estimation de la dérivée ce qui est nécessaire pour extraire les GCIs.

2.3 Découpage du signal en tranche et détection tranche par tranche

Les signaux EGG étant quasi périodiques, la fréquence fondamentale est recherchée comme la fréquence pour laquelle la densité spectrale de puissance est maximale.
La longueur d'une tranche est fixée à 10*Tfondamentale et le shift à 5*Tfondamentale (ces paramètres sont choisis sans justification scientifique mais seulement pour donner de bons résultats).

2.4 Seuil global et seuil local

Le seuil local est calculé comme 28% de la valeur maximale du signal sur la fenêtre.
Le seuil global est calculé comme 16% de la valeur maximale sur l'ensemble du signal.

Le seuil utilisé est le plus grand seuil entre le global et le local.

Imaginons une tranche de signal ne contenant aucun GCIs, sons non-voisés par exemple. Dans cette situation, le seuil local sera très faible car la valeur max sur cette tranche ne correspondra pas à un GCI, si on utilise le seuil local, on aurait une détection de nombreux pic qui ne sont pas des GCIs. Le seuil global empêche ces fausses détections en définissant une amplitude minimale que les pics doivent atteindre pour être détectés comme GCIs.

2.5 Détection des pics

La fonction Matlab findpeaks se charge de la détection.

```
findpeaks(dEGG(fenetre),'MinPeakHeight',dEGGth,'MinPeakDistance',dist_min);
```

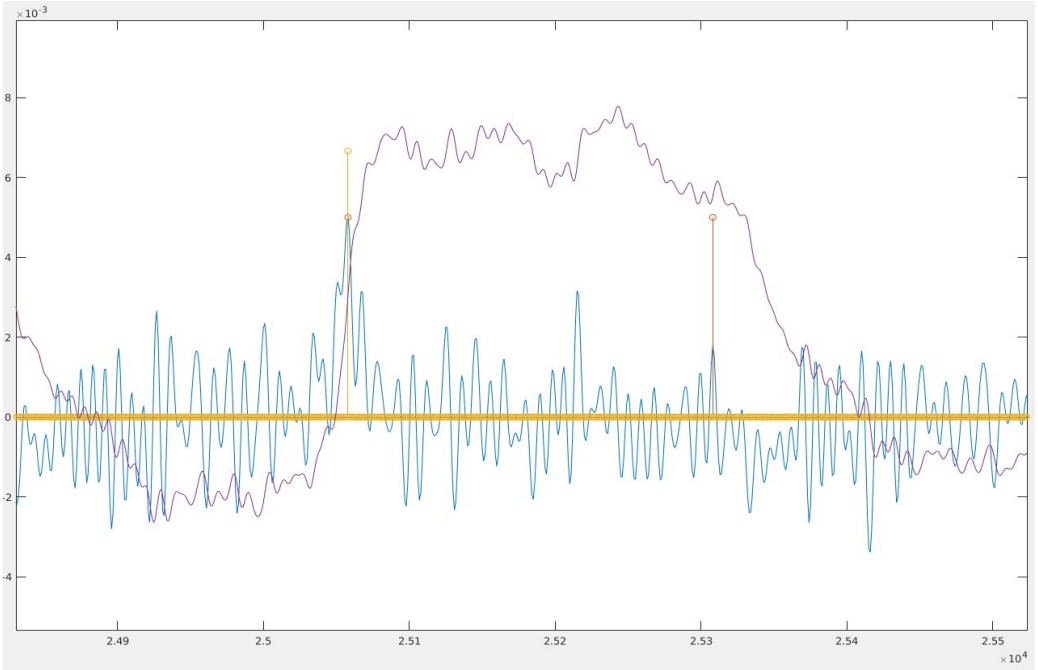
Les arguments de la fonction sont: le seuil (dEGGth) et la durée minimale entre 2 pics. Ce dernier paramètre permet de rechercher seulement un GCI par période.

2.6 Suppression des GCIs trop proches

Le signal EGG est découpé en tranche et les tranches se superposent. Sur les zones de superpositions on risque de détecter plusieurs fois le même GCIs à des instants légèrements différents. Les GCIs distants de moins de T0/3 sont supprimés.

Cette méthode est basée sur la dérivée du signal EGG, c'est une amélioration de la méthode dEGG dans le sens ou elle ne nécessite pas de trouver un seuil manuellement mais elle se charge de trouver localement un seuil adapté au signal.
Cependant malgré tous les efforts fais pour trouver la vérité terrain, le signal EGG 29-a_n-egg.wav reste problématique. Cet enregistrement EGG est un très bon exemple d'enregistrement pour lequel la vérité terrain est difficile à trouver. La suite du rapport a pour objectif d'améliorer la méthode d'extraction de la vérité terrain pour qu'elle fonctionne sur cet enregistrement.

Voici un morceau du signal



29-a_n-egg.wav

En Violet EGG filtré

En bleu dEGG

pic orange GCIs détecté par la méthode

Figure 12 : Courbe avec localisation de GCI trop proche

Le signal est très bruité, l'estimation de la dérivé est donc mauvaise. Sur cette partie du signal on observe un GCI correct et un autre faux.

On pourrait penser que les pics les plus grands correspondent au GCIs mais ce n'est pas le cas. Sur la courbe ci-dessus le pic le plus fort correspond à une mauvaise détection.

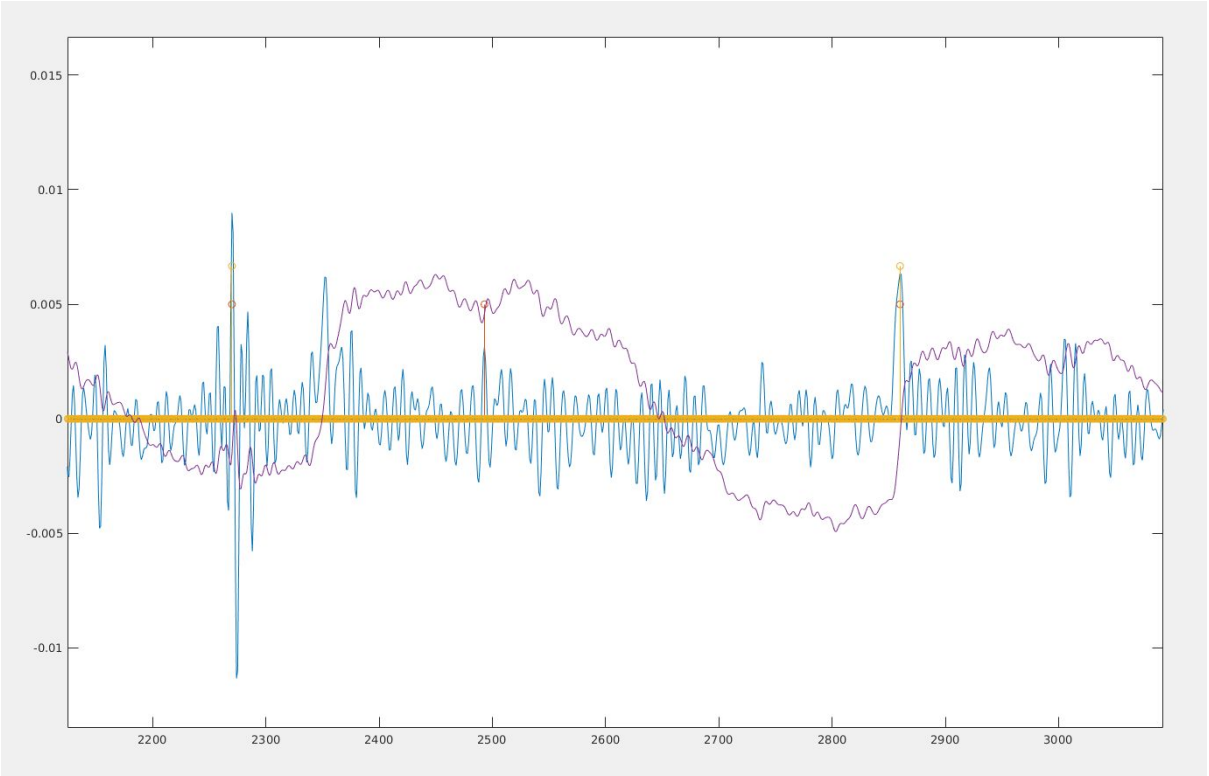


Figure 13 : Exemple de mauvaise detection d'un GCI

Les GCIs semblent être situés à proximité d'un passage par zéro du signal EGG.

2.7 Zeros crossing sur le signal EGG

L'idée est de repérer les positions approximatives des GCIs avec un zeros crossing positif sur le signal EGG dans un premier temps puis dans un second temps positionner précisément le GCIs à partir du signal dEGG.

2.8 Résumé méthode EGG/dEGG

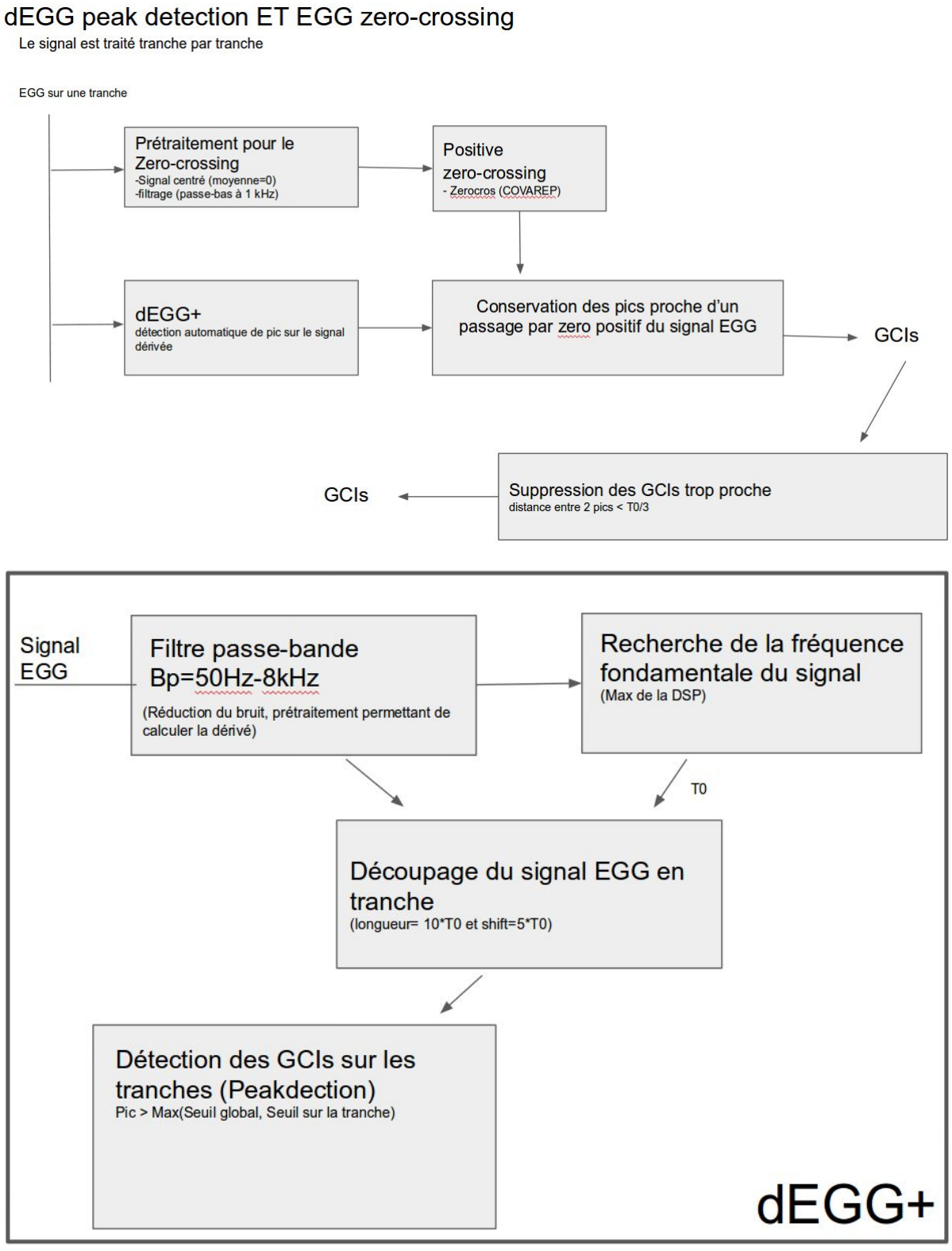
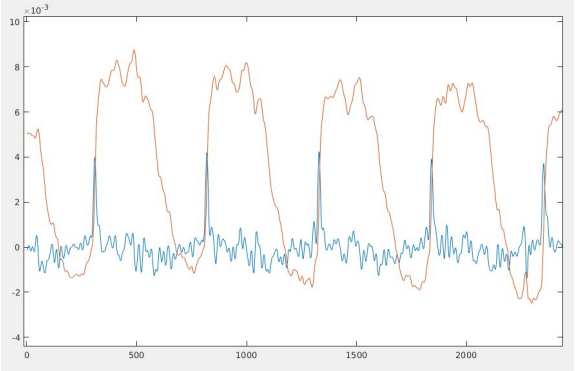


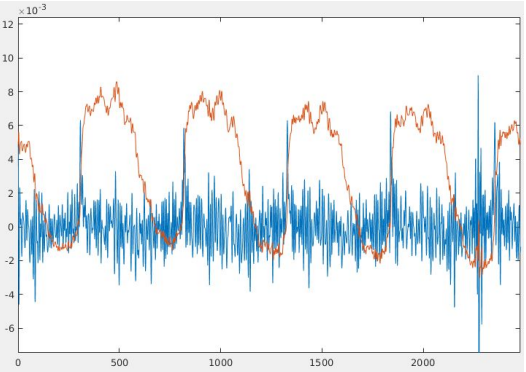
Figure 14 : Résumé méthode EGG /dEGG

2.9 Amélioration possible: Filtrage supplémentaire

Dans l'idée de développer une méthode plus robuste au bruit. Il est envisageable de modifier le filtre du bloc dEGG+ pour filtrer à 1kHz



Filtrage à 1 kHz



Filtrage à 8 kHz

L'inconvénient est une perte de précision.

3 Test des méthodes sur base d'enregistrement

Dans cette partie, les méthodes d'extraction des GCIs depuis le signal de parole sont testées. Les méthodes SEDREAMS et YAGA sont testées. Les GCIs extrait de ces méthodes sont comparé avec les 3 vérités terrain étudié dans ce rapport, SIGMA, dEGG, et EGG/dEGG.

La méthodologie de test employée ici permet d'évaluer 5 paramètres statistiques: Hit rate, FA, MR, IDA, A25. Un test nécessite une base (ensemble d'enregistrement de parole et egg), une méthode d'extraction de la vérité terrain et une méthode de détection sur le signal de parole.

3.1 Bases

Les tests sont réalisés sur 5 bases distinctes, les 4 premières bases sont issues de CMU_Artic: KED, BDL, JMK, SLT. Sur ces bases, les signaux sont échantillonnés à 32 kHz.
http://festvox.org/cmu_arctic/cmu_arctic_report.pdf

La 5eme base provient de l'université allemande de Saarland: http://stimmdb.coli.uni-saarland.de/index.php4#target
Sur cette base, les signaux sont échantillonnés à 50 kHz.
Pour chaque speaker la base contient un enregistrement de la phrase *'Guten Morgen, wie geht es Ihnen?' (Good morning, how are you?)* ainsi que les voyelles "i" "a" "u" prononcées à une hauteur de voix normale, grave et aiguë.

La base permet de sélectionner des enregistrements selon l'âge, le genre et diverses pathologies.
Pour des raisons de simplicité, seuls les enregistrements de “a” neutre ont été retenus pour les tests. Ces enregistrements sont voisés sur toute la durée de l’enregistrement et sont quasi-périodiques. La base contient des enregistrements de patient “healthy” et “Pathological”. Voici ci-dessous la segmentation choisie pour les tests, le nombre d'enregistrement dans chaque partition est mentionné entre parenthèses.

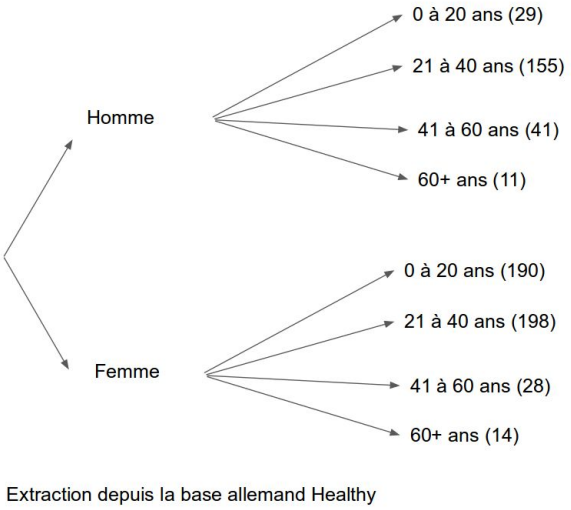


Figure 15 : Partitionnement de la base allemande pour des voix non-pathologiques

Les GCIs extraits à partir des méthodes de localisation sur le signal EGG vont être comparés avec les GCIs localisés sur le signal de parole par la méthode SEDREAMS. L’objectif est de comparer plusieurs méthodes d’extraction de la vérité terrain pour les GCIs. De plus, les performances de la méthode SEDREAMS avec une vérité terrain extraite par la méthode dEGG sur les bases KED, BDL, JMK, SLT est disponible dans la publication de Thomas Drugman. Les statistiques seront donc comparées dans le but de valider la méthodologie employée ici.

3.2 Indicateur statistique

Généralement, les méthodes de location des GCIs sont évaluées par 5 indicateurs statistiques: HR, MR, FAR, IDA, A25. Ces indicateurs sont ceux utilisés par Drugman, nous utiliserons les mêmes ici.

Hit Rate (HR) : hitcnt/Nref

Représente le pourcentage de cycle pour lequel un GCI est détecté sur le cycle

Nref = “Nombre de cycles comptés à partir de l’EGG (vérité terrain)”

hitcnt = “Nombre réussite de localisation”

Miss Rate (MR) : misscnt/Nref

Représente le pourcentage de cycles pour lequel aucun GCI n’est détecté sur le cycle

misscnt = “Nombre de détections manquantes”

False Alarm Rate (FAR): FAcnt/Nref

Représente le pourcentage de cycles pour lequel plusieurs GCIs sont détectés sur le cycle

FAcnt = “Nombre de fausses alarmes”

Ndet = “Nombre de cycles issu de l’algo de detection”

Identification Accuracy (IDA) : Ecart-type de l’erreur

A25ms

Représente le pourcentage de GCIs se situant à une distance inférieure à 0.25ms du GCI vrai

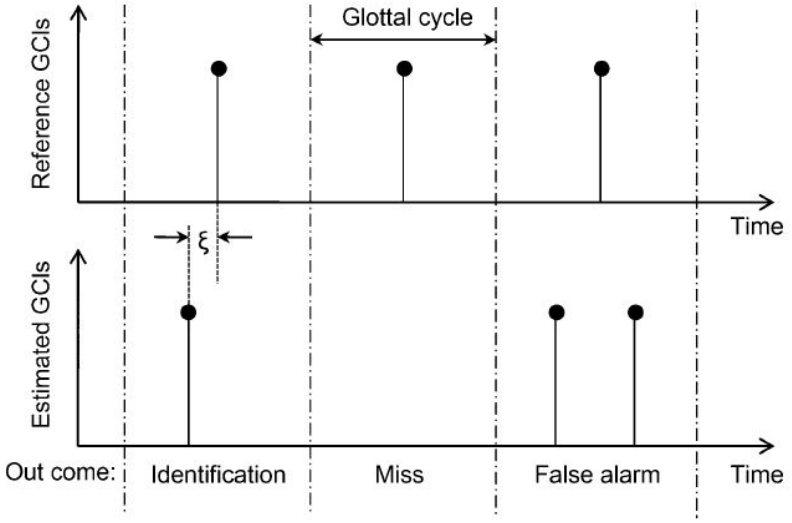


Fig. 6. Characterization of GCI estimates showing three glottal cycles with examples of each possible outcome from GCI estimation [14]. Identification accuracy is characterized by ξ.

Depuis la publication de Drugman

Figure 16 : Indicateur statisque

3.3 Boucle de test

Une boucle parcourt l’ensemble des enregistrements de la base. La statistique est faite pour chaque enregistrement indépendant. C’est la fonction matlab statGCIgeek fournie par Vahid Khanagha qui calcule la statistique. La statistique sur la base entière est calculée comme la moyenne des indicateurs statistiques de chaque enregistrement.

3.4 Shift automatique

Bien que les signaux de parole et d’EGG soient enregistrés simultanément il ne sont pas parfaitement synchronisés. Cette absence de synchronisation peut provenir de plusieurs facteurs, par exemple des retards provoqués par le matériel électronique filtre, conversion analogique numérique ou encore le temps de propagation de la parole dans l’air etc.. Il est

nécessaire de corriger ce décalage entre les voies le plus justement possible. On remarque généralement un décalage d'environ 0.5ms entre l'EGG et la parole. Voici une méthodologie pour estimer ce décalage. La figure ci-dessous montre l'histogramme de l'erreur de localisation des GCIs:

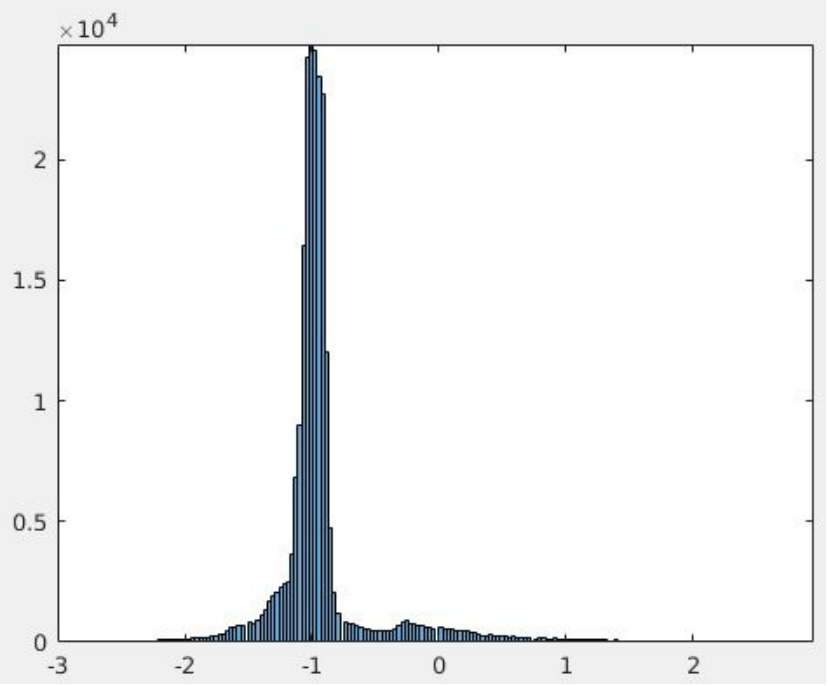


Figure 17 : Histogramme de l'erreur base BDL dEGG SERREAMS en ms

La moyenne de cette erreur est non-nulle. Observons les histogrammes publiés par Drugman. Dans chaque cas la moyenne de l'erreur semble être nulle. Nous supposerons que cette moyenne est uniquement due au décalage entre le signal de parole et l'EGG.

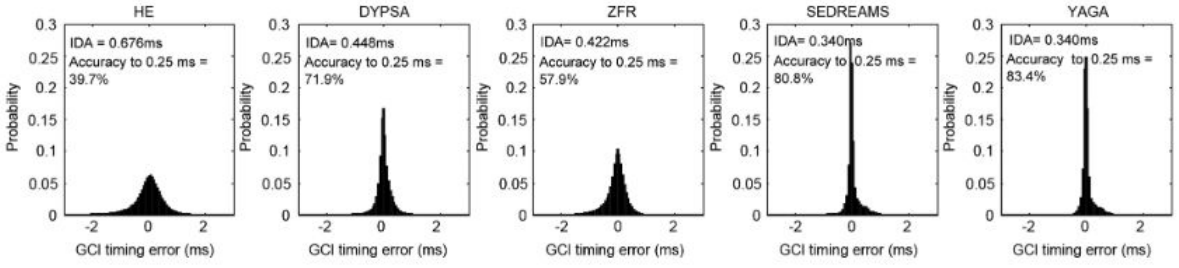


Fig. 9. Histograms of the GCI timing error averaged over all databases for the five compared techniques

Figure 17 : Histogramme de l'erreur issue de la publication de Thomas Drugman (1)

Le décalage entre l'EGG et la parole est estimé automatiquement comme étant la moyenne de l'erreur. L'objectif est de trouver un décalage permettant d'obtenir une erreur moyenne nulle. On pourrait imaginer d'autres méthodes pour trouver automatiquement ce delay shift comme par exemple prendre le maximum de l'histogramme. Ce retard est important à corriger pour obtenir une bonne statistique. La paramètre le plus impacté par ce retard est l'accuracy à 0.25 ms.

3.5 Try/Catch

Chaque base d'enregistrement est constituée de plusieurs centaines d'enregistrements. Il arrive dans de rares cas que l'algorithme de test échoue sur un enregistrement précis. Plusieurs causes d'échec sont possibles: échec d'extraction de la vérité terrain, échec d'extraction des GCIs depuis le signal de parole etc…
Une structure de type Try/Catch est utilisée pour gérer les erreurs dans la boucle de test. Lorsqu'une erreur apparaît, l'enregistrement en cours est ignoré, l'algorithme de test passe à l'enregistrement suivant.
Seuls les enregistrements ne provoquant pas d'erreur sont pris en compte dans la statistique. Le nombre d'enregistrements ignorés est compté et le pourcentage d'enregistrements ignorés par rapport au nombre d'enregistrements total est calculé.

3.6 Post-Traitement: GCI UV

Voici sur ce graphique le signal EGG en orange et en bleu les GCIs localisés par la méthode SEDREAMS.

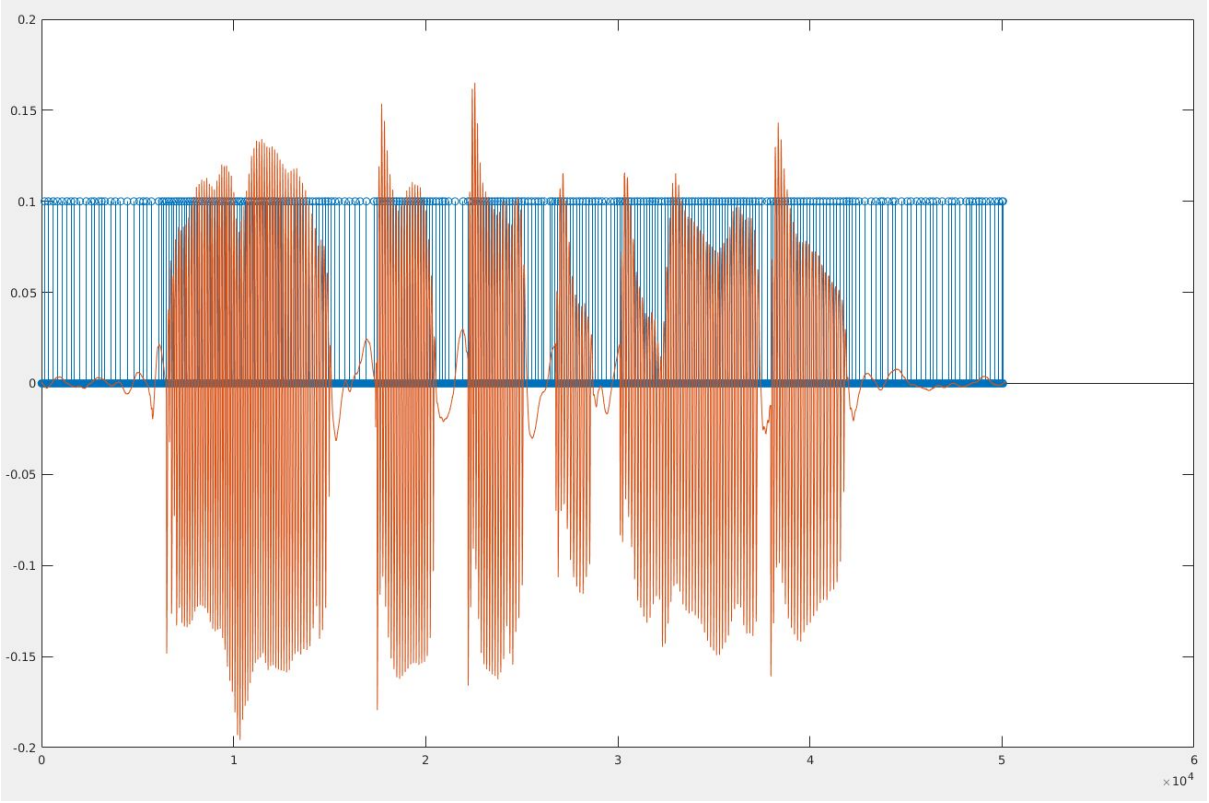


Figure 18 : EGG en orange et en bleu les GCIs localisés par la méthode SEDREAM

On remarque des GCIs localisés sur des parties non-voisées du signal. Comme les sons non-voisés n'engagent pas de mouvement glottal, ces GCIs sont des fausses détections. La méthode SEDREAMS ne restreint pas la recherche des GCIs sur les segments voisés du signal.
Nous avons testé d'ajouter un post-traitement à la méthode SEDREAMS pour supprimer les GCIs appartenant à des segments non-voisés.

La fonction Pitch SRH, écrite par Thomas Drugman, donne la décision voisée non-voisée pour des tranches du signal. Voici sur ce schéma une méthode inspirée de **COVAREP_feature_extraction.m** du répertoire COVAREP pour obtenir la décision voisée non-voisée à chaque instant.

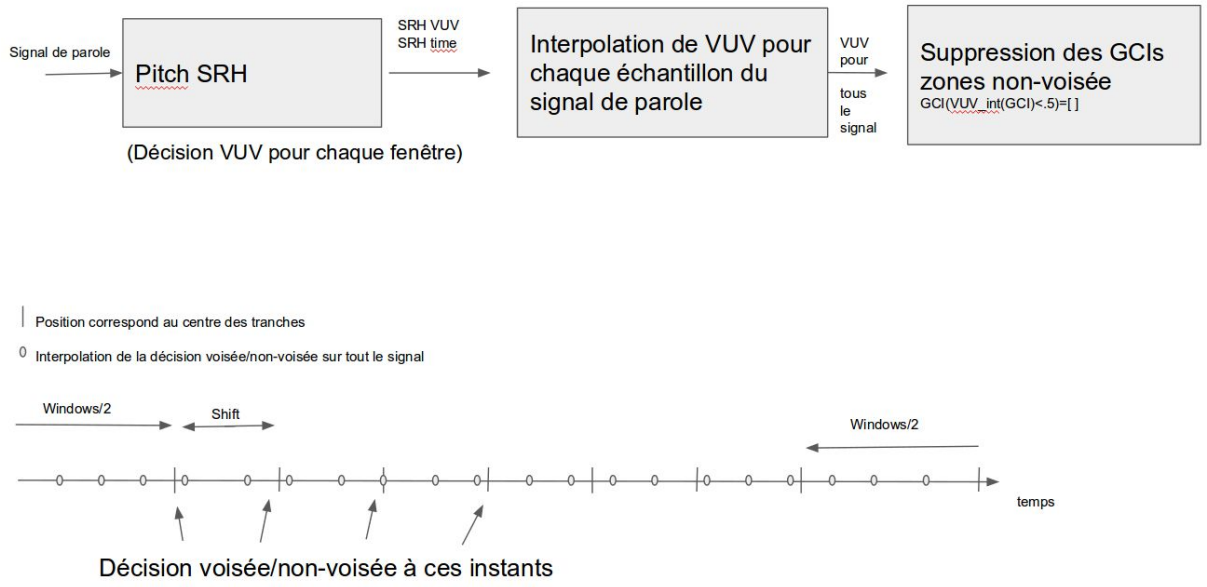


Figure 19 : Traitement des GCIs sur les segment non-voisées

La décision pour chaque instant est obtenue par interpolation.

La méthodologie de statistique analyse seulement les cycles glottaux identifiés dans l'EGG, donc uniquement les parties voisées du signal. Ce post-traitement n'améliorera pas notre statistique. Il faut cependant retenir qu'en absence de ce post-traitment de nombreux GCIs sont positionnés sur les parties non-voisées du signal.

3.7 Résultats pour les bases CMU_Artic

3.7.1 Validation par rapport à la publication de Thomas Drugman

Voici un tableau comparatif des résultats trouvés dans la publication de 2012 publiée par Thomas Drugman (en noir) et des résultats trouvés dans le cadre de ce stage (en bleu).

GLOTAL	HR (en %)	MR (en %)	FAR (en %)	IDA (en ms)	A25 ms (en %)	Taux d'échec (en %	Shift en nb d'ech.
BASE KED							
dEGG (Drugman	98.65	0.67	0.68	0.33	94.65		
dEGG 0.033	98.75	0.09	1.14	0.33	93.10	0.22	-13
BASE BDL							
dEGG (Drugman	98.08	0.77	1.15	0.31	89.35		
dEGG 0.033	97.45	0.66	1.88	0.34	85.23	0	-16
BASE JMK							
dEGG (Drugman	99.29	0.25	0.46	0.42	80.78		
dEGG 0.033	97.97	0.58	1,44	0.47	77.40	0.18	-15
BASE SLT							
dEGG (Drugman	99.15	0.12	0.73	0.30	81.35		
dEGG 0.009	98.44	0.04	1.50	0.28	77.07	0	-16

D'après les résultats statistiques de la publication: Detection of Glottal Closure Instants From Speech
Signals: A Quantitative Review - Thomas Drugman ***en noir***
Figure 20 : Tableau comparatif des résultats

Sur les bases KED, BDL et JMK le seuil pour la méthode dEGG est fixée à 0.033. Pour la base SLT le seuil est fixé à 0.009. La correction de synchronisation apportée entre le signal de parole et EGG est mentionnée dans la dernière colonne du tableau.
Dans la publication de Thomas Drugman, le seuil pour la méthode dEGG et le delay shift ne sont pas communiqués. Nous choisissons ces paramètres pour obtenir les résultats les plus proches possible de ceux de la publication.
Les résultats trouvés ici sont finalement relativement proches par rapport à ceux de la publication ce qui valide le choix des paramètre et notre méthodologie de test.
L'accuracy à 0.25 ms est légèrement plus mauvaise que dans la publication, le plus mauvais cas est pour la base SLT.

3.7.2 Comparatif des méthodes d'extraction de la vérité terrain sur les bases CMU_Artic

Pour chacune des méthodes d'extraction de la vérité terrain: SIGMA, dEGG, EGG/dEGG, les méthodes d'extraction des GCIs depuis le signal de parole: SEDREAMS et YAGA sont comparés. A chaque fois, les GCIs extraits du signal EGG sont comparés avec ceux extraits du signal de parole. De cette façon, les méthodes d'extraction de la vérité terrain sont comparées entre elles. L'idéal serait d'avoir à disposition une vérité exacte sur la position des GCIs pour pouvoir comparer rigoureusement les GCIs. Malheureusement une telle chose n'existe pas. Il est donc impossible de comparer les méthodes d'extraction de la vérité terrain rigoureusement. Cette méthodologie apporte néanmoins un bon aperçu des performances de chacune des méthodes

cf Annexe page 1

Le delay shift est calculé sur chaque bases (KED, BDL, JMK, SLT) pour que la moyenne de l'erreur soit nulle avec les méthodes SEDREAMS et dEGG.

Ce tableau montre bien que le choix de la vérité terrain influence les résultats obtenus pour les algorithmes de détection des GCIs dans le signal de parole. Il y a donc plusieurs moyens d'obtenir une vérité terrain, chacune de ces vérités terrain est différente.

3.8 Résultats sur la base saarland

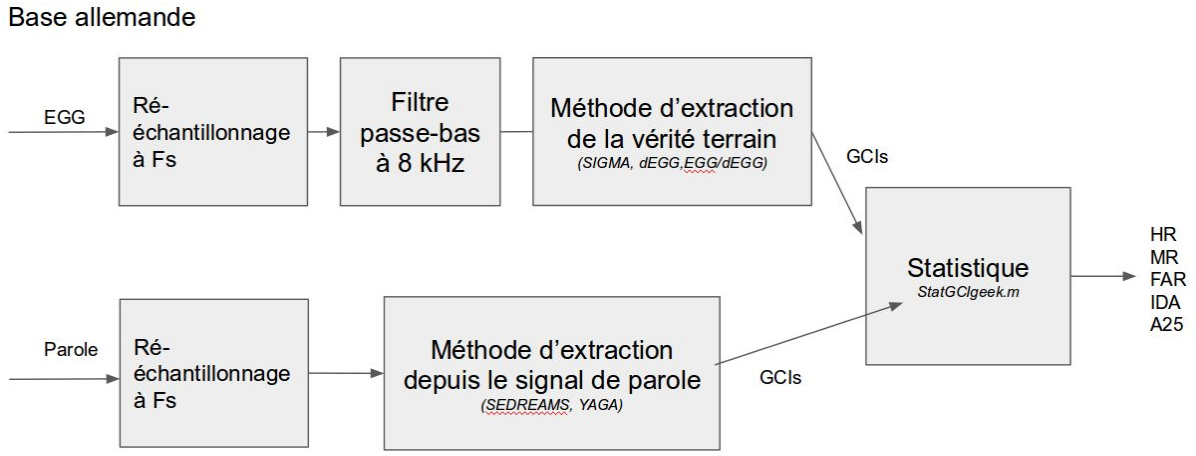


Figure 21 : Chaîne de traitement des signaux pour la base allemande

Le delay shift est calculé une seule fois pour chaque segment de la base. Il est calculé pour annuler la moyenne de l'erreur entre les méthodes SEDREAMS et dEGG. Ce delay shift est conservé pour tous les tests. La conservation du delay shift sur un même segment de la base durant tous les tests, permet de comparer les résultats obtenus par les différentes méthodes. Les résultats de tous les tests se trouve dans les annexes.

3.8.1 Influence de la fréquence d'échantillonnage

Les résultats sont meilleurs lorsque les signaux sont ré-échantillonnés à 16 kHz (voir annexe). Une grosse différence au niveau de l'accuracy à 0.25 ms est observée. La fréquence d'échantillonnage plus faible semble être une amélioration pour la méthode SEDREAMS. L'EGG étant filtré par un filtre passe-bas à 8 kHz, la fréquence d'échantillonnage influence uniquement la méthode d'extraction des GCIs sur le signal de parole.

3.8.2 Taux d'échec

Le méthode d'extraction de la vérité terrain qui engendre le plus fort taux d'échec est la méthode dEGG. Souvent ce taux atteint ⅓ d'enregistrements ignorés.
Le taux d'échec augmente avec la fréquence d'échantillonnage. A une fréquence d'échantillonnage forte, 50 kHz, le taux d'échec est important, lorsqu'on sous-échantillonne les signaux aux fréquences 32 kHz et 16 kHz, ce taux d'échec diminue.

3.9 Diplophonie

Voici une partie dédiée aux voix pathologiques.L'étude se limitera à la diplophonie.
Définition de la diplophonie :"Il s'agit d'un sujet qui émet lorsqu'il tente de parler ou chanter, un son comportant simultanément deux fréquences fondamentales, chacune engendrant ses propres harmoniques" http://dictionnaire.education/fr/diplophonie. La base allemande Saarland contient des enregistrements de 5 patients souffrant de diplophonie, les enregistrements du 5eme patient sont inexploitables. Seuls les 4 premiers patients seront étudiés.

Voici le signal de parole en haut et EGG en bas d'un patient souffrant de ce trouble de la parole.

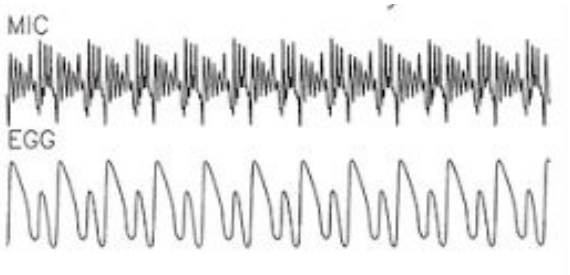


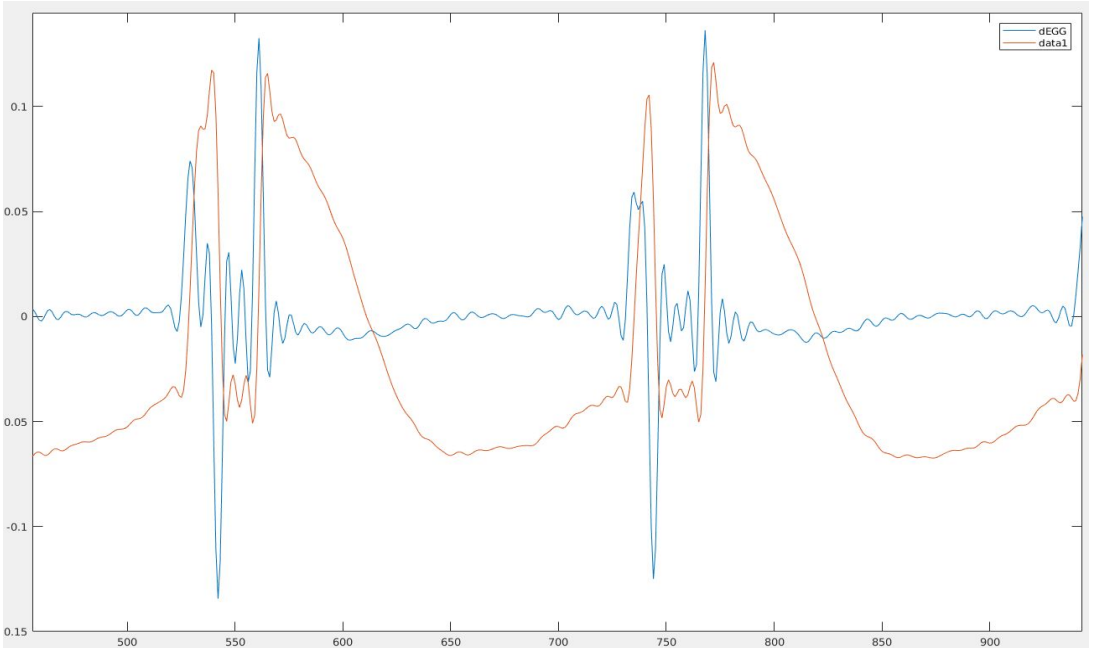
Figure 22 : Signal parole et EGG caractéristique de la diplophonie
D'après Diagnosis and Treatment of Voice Disorders - John S. Rubin,Robert T. Sataloff,Gwen S. Korovin

Une période du signal de parole est constituée de deux phases d'ouverture et de fermeture glottal. Cette allure de EGG est une caractéristique de la pathologie.
Le but de cette partie, est d'évaluer la capacité ou l'incapacité des méthodes étudiées dans les parties précédentes à détecter les GCIs pour ce type de pathologie. Les méthodes testées seront, celles de détection depuis le signal de parole, SEDREAMS et YAGA, MSM (10), et celles d'extraction de la vérité terrain, SIGMA, dEGG et EGG/dEGG.

Le maillon faible de toutes les méthodes d'extraction de la vérité terrain est l'automatisation de la tâche. De toute évidence, il est beaucoup facile et exact de détecter les GCIs manuellement depuis l'EGG. Dans certaines publications, la vérité terrain est extraite manuellement: "GCIs and GOIs were hand-labele" d'aprés publication sur SIGMA (http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4912310)
La méthodologie peut être par exemple de repérer manuellement les pics sur le signal dEGG.

Dans cette partie, la vérité terrain est marquée manuellement. Tout d'abord, la vérité terrain est établie par la méthode SIGMA. Cette vérité terrain est ensuite corrigé manuellement à

partir de l'observation des signaux EGG et dEGG. Les pics positifs du signal dEGG correspondent à des GCIs.
Dans le cadre de cette pathologie, parfois le signal dEGG ne permet pas de positionner certain GCIs. Voici ci-dessous un exemple problématique:



Signal EGG en orange et dEGG en bleu

Figure 23 : Signal EGG et dEGG Diplophonie

Dans ce cas, le GCI est positionné approximativement entre la phase d'ouverture et de fermeture glottale. Sur cette figure, il y a deux instants de fermeture.

Voici ci-dessous, un tableau récapitulatif des performances de chaque méthode pour la diplophonie. Les tests sont réalisés pour la prononciation d'un "a" neutre et d'une phrase.

Diplophonie			seuil dEGG 0.053		Resample à 16 kHz
			Filtrage EGG à 8 kHz		delay shift -12 (VT et parole
a-neutre	HR (en %)	MR (en %)	FAR (en %)	IDA (en ms)	A25 ms (en %)
Methode d'extraction de la vérité terrain					
dEGG	85.52	14.13	0.33	0.26	79.52
SIGMA	82.12	17.72	0.14	0.17	79.94
EGG/dEGG	88.15	11.84	0	0.25	81.28
ZcEGG	97.32	1.80	0.86	0.38	48.06
Methode d'extraction depuis le signal de parole					
SEDREAMS	99.30	0.28	0.40	0.27	78.08
YAGA	88.06	7.63	4.30	1.25	2.35
MSM	86.60	13.32	0.06	0.23	78.71
Phrase	HR (en %)	MR (en %)	FAR (en %)	IDA (en ms)	A25 ms (en %)
Methode d'extraction de la vérité terrain					
dEGG	79.47	20.30	0.22	0.11	77.63
SIGMA	93.03	2.46	4.49	0.11	90.89
EGG/dEGG	84.40	15.59	0	0.12	82.47
ZcEGG	89.45	8.33	2.20	0.48	55.38
Methode d'extraction depuis le signal de parole					
SEDREAMS	98.15	1.22	0.61	0.44	74.03
YAGA	93.61	2.12	4.26	1.21	5.84
MSM	85.55	11.98	2.46	0.73	66.95

Figure 24 : Performance de chaque méthodes pour la diplophonie

Concernant les méthodes d'extraction de la vérité terrain, les méthodes étudiées dans les parties précédentes sont testées ainsi qu'une nouvelle méthode appelé ZcEGG, écrite spécialement pour extraire la vérité terrain de patients souffrant de diplophonie.
Les méthodes dEGG, SIGMA et EGG/dEGG rate environ 10% des GCI. Parmi les trois méthodes c'est la méthode EGG/dEGG qui affiche les meilleurs résultats.

ZcEGG

La méthode ZcEGG est une méthode d'extraction de la vérité terrain basé sur les passages par zero du signal EGG.

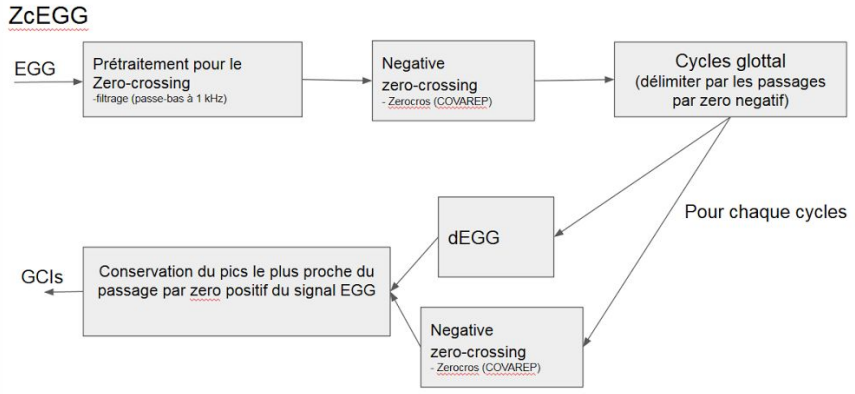


Figure 25 : Schéma ZcEGG

Cette méthode positionne les cycles glottals entre les passages par zero négatif du signal EGG et recherche un GCI sur chaque cycle glottal. Pour chaque cycle, dEGG propose des GCIs candidats, le candidat le plus proche du passage par zero positif est conservé.

Cette méthode est une amélioration pour le HR, MR et FAR. Les cycles glottals sont donc mieux identifiés avec les méthodes dEGG, SIGMA et EGG/dEGG. Cependant, la précision est mauvaise. La moitié des GCIs se situe en dehors de la fenêtre -0.25 ms 0.25ms.

Concernant les méthodes d'extraction depuis le signal de parole, c'est la méthode SEDREAMS qui affiche les meilleurs résultats.

Conclusion

L'objectif initial du stage était de travailler sur les algorithmes de détection des GCIs sur le signal parole, les tester sur des voix pathologiques et tenter d'améliorer ces algorithmes en fonction de résultats obtenus.
Dès les premiers tests, les problématiques autour de la vérité terrain sont apparues. La vérité terrain est indispensable pour pouvoir étudier les algorithmes de détection des GCIs depuis le signal de parole.
La meilleure vérité terrain est celle issue d'un marquage manuel des GCIs sur le signal EGG. Cependant, cela demande un travail long et fastidieux, il est donc impossible d'obtenir la vérité terrain de cette manière pour un gros volume d'enregistrement. La solution est d'extraire la vérité terrain automatiquement, mais l'automatisation de la tâche conduit à des distorsions qui risquent de rendre la vérité terrain fausse.
Dans le cadre de l'étude des instants de fermeture glottal, la construction de la vérité terrain est un problème aussi important et difficile que la détection de ces instants depuis le signal de parole.
Le cas des voix pathologiques est actuellement très peu traité. Avant de pouvoir tester et adapter les algorithmes de détection des GCIs pour les voix pathologiques, il faut être capable d'obtenir une vérité terrain. Établir une vérité terrain pour des voix pathologiques pourrait demander d'autres méthodologies que celles utilisées pour les voix non-pathologiques.
Par exemple, pour la diplophonie, les méthodes existantes d'extraction de la vérité terrain sont moins justes que les méthodes d'extraction depuis le signal de parole. Ce qui montre que les méthodes actuelles d'extraction de la vérité terrain ne sont pas satisfaisantes.

Résumé:

Le traitement de la parole est une thématique importante des sciences de l'ingénieur alliant traitement du signal et connaissances médicales. La parole est utilisée comme vecteur d'informations pour de nombreuses applications industrielles, comme la reconnaissance de la parole, les interfaces homme machine et bien d'autres applications. L'étude de la parole pourrait permettre le diagnostic différentiel de maladies ayant comme symptôme des troubles de la voix. Certaines d'entre elles sont dues à un dysfonctionnement des cordes vocales.
Des méthodes basées sur la parole ont été développées ces dernières années pour identifier les instants ou les cordes vocales entrent en contact, ces instants sont appelés GCIs (Glottal Closure Instant). La vérité terrain est obtenue par l'intermédiaire de l'electroglottographie (EGG). L'EGG fournit un signal, image du mouvement des cordes vocales. De ce signal, plusieurs méthodes permettent d'extraire automatiquement les GCIs. Ces méthodes donnent des résultats différents et parfois faux pour des voix non-pathologiques, et ne sont à priori pas adaptées aux voix pathologiques. On peut donc difficilement obtenir une vérité terrain fiable.
La vérité terrain est pourtant indispensable pour pouvoir développer et étudier des méthodes de détection des GCIs. Ceci a donc constitué le principal axe de travail durant le stage.

Abstract:

Speech processing is an important thematic of the sciences of the engineer allying signal processing and medical knowledge. Speech is used as an information vector for many industrial applications, such as speech recognition, man-machine interfaces and many other applications. The study of the speech could allow the diagnosis of a disease.
Many diseases has voice disorders as symptoms. Some of them are due to a malfunction of the vocal cords.
Speech-based methods have been developed in recent years to identify instants when vocal cords come into contact, these instants are called GCIs (Glottal Closure Instant). The ground truth is obtained via electroglottography (EGG). The EGG provides an image signal of the movement of the vocal cords. From this signal several methods allow us to extract the GCIs automatically. These methods give different and sometimes false results for non-pathological voices and is probably not suitable for pathological voices. It is therefore difficult to obtain a reliable ground truth.
The ground truth is nevertheless essential to be able to develop and study methods of detection of GCIs. This is exactly the main subject I had to work on during my internship.

Biblio

(1) Detection of Glottal Closure Instants From Speech Signals: A Quantitative Review Thomas Drugman Mark Thomas Jon Gudnason Patrick Naylor Thierry Dutoit 14 November 2011

(2) Effective Glottal Instant Detection and Electroglottographic Parameter Extraction for Automated Voice Pathology Assessment Pranav S. Deshpande and M. Sabarimalai Manikandan 17 January 2017

(3) ELECTROGLOTTOGRAPHIE (E.G.G.)
http://www.lpl-aix.fr/~ghio/pedago-EggFR.htm

(4) II Electroglottography
http://www2.ims.uni-stuttgart.de/EGG/frmst2.htm

(5) COVAREP https://github.com/covarep/covarep

(6) http://festvox.org/cmu_arctic/cmu_arctic_report.pdf CMU_Artic database

(7) https://tel.archives-ouvertes.fr/tel-00817694/document Etudes multiparametriques de la voix et de la parole apres cordectomie laser par voie endoscopique de type II-III Lucille Wallet

(8) THE SIGMA ALGORITHM FOR ESTIMATION OF REFERENCE-QUALITY GLOTTAL CLOSURE INSTANTS FROM ELECTROGLOTTOGRAPH SIGNALS
Mark R. P. Thomas and Patrick A. Naylor
http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7080348

(9) Estimation of Glottal Closing and Opening Instants in Voiced Speech using the YAGA Algorithm Mark R. P. Thomas,Member, IEEE, Jon Gudnason
https://spiral.imperial.ac.uk/bitstream/10044/1/15494/2/IEEE%20Transactions%20on%20Audio%20Speech%20and%20Language%20Processing_20_1_2012.pdf

(10) Detection of Glottal Closure Instants Based on the Microcanonical Multiscale Formalism Vahid Khanagha, Khalid Daoudi, and Hussein M. Yahia

Annexe