Caractérisation automatique du rythme de la parole : application aux cancers des voies aéro-digestives supérieures et à la maladie de Parkinson

Jérôme Farinas – Corine Astésano
Robin Vaysse

Journée scientifique TMBI – jeudi 11 mai 2023 – Auditorium J. Herbrand, IRIT

Problématique

La caractérisation automatique du rythme de la parole pathologique

- On souhaite utiliser / créer des outils totalement automatiques pour caractériser le prosodie des patients
- Analyser comment se comporte le rythme des personnes atteintes de pathologies de la voix
  - Quel impact sur la fluence des patients ?
  - Compensent-ils leurs pertes d'articulation par une prosodie particulière ?

Matérialisation du rythme

Structuration de la parole via :

- Matérialité **phonétique**
  - Fréquence fondamentale (F0)
  - Intensité
  - Durée
- Matérialité **phonologique**
  - Structuration en **unités hiérarchiques**

Structuration prosodique

Niveaux de structuration hiérarchique :

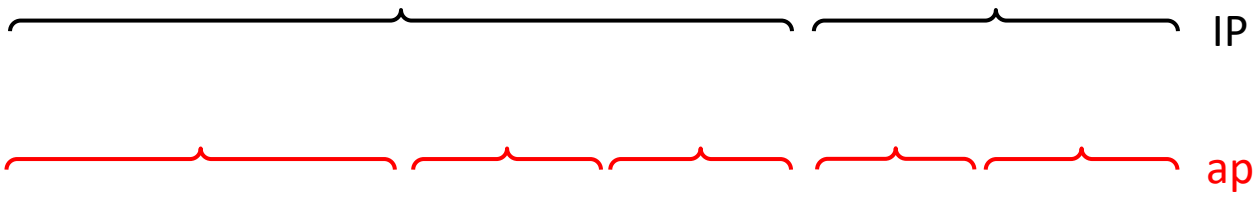
- Le syntagme intonatif (IP, *Intonational phrase*)
-
-

L'ordinateur portable de Gabriel est cassé. Il faudra le changer.

Structuration prosodique

Niveaux de structuration hiérarchique :

- Le syntagme intonatif (IP, *Intonational phrase*)
-
- Le syntagme accentuel (ap, *accentual phrase*) (Jun & Fougeron, 2000)

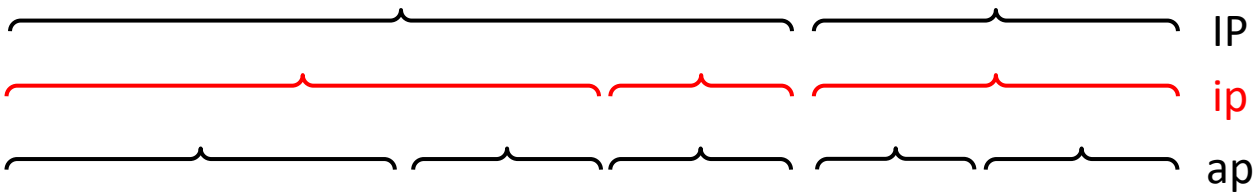


L'ordinateur portable de Gabriel est cassé. Il faudra le changer.

Structuration prosodique

Niveaux de structuration hiérarchique :

- Le syntagme intonatif (IP, *Intonational phrase*)
- Le syntagme intermédiaire (ip, *intermediate phrase*) (Michelas & D'Imperio, 2010)
- Le syntagme accentuel (ap, *accentual phrase*) (Jun & Fougeron, 2000)



L'ordinateur portable de Gabriel est cassé. Il faudra le changer.

Premiers travaux TMBI

- Stage financé par TMBI de **Baptiste Moret** avec Corine Astesano (Octogone-Lordat) et Jérôme Farinas (IRIT) en 2019

- Contributions :
  - Synthèse des modélisations de la prosodie
  - inventaire des outils pour représenter ou modéliser

RAPPORT DE STAGE

Réalisation d'une plateforme logicielle pour l'analyse et la mesure de la dysfluence prosodique en parole pathologique

BAPTISTE MORET, 3ÈME ANNÉE ÉLECTRONIQUE ET TRAITEMENT DU SIGNAL, ENSEEIHT

Octogone Lordat
Université de Toulouse
2-Jean-Jaures
31058 Toulouse, France

Institut de Recherche en Informatique de Toulouse
118 Route de Narbonne
31062 Toulouse, France

Tuteurs de stage :
Mr Jérôme Farinas, IRIT, SAMoVA
Mme Corine Astésano, Octogone Lordat

18 Mars 2019 — 30 Août 2019

Cadre de l'étude

- **Projet ANR RUGBI 2019-2023** (*"Relevant linguistic Units to improve the intelliGiBIlity measurement of speech production disorders")*

  — **Complémenter et expliciter la mesure du déficit d'intelligibilité** de la parole
    - Cancer des Voies Aéro Digestives Supérieures (VADS)
    - Maladie de Parkinson (Park)

  — Recherche **automatique d'unités prosodiques** pertinentes

- **Groupement d'Intérêt Scientifique PAROLOTHEQUE** depuis 2022 **:** accès facilité aux données cliniques et métadonnées (perceptives et automatiques)

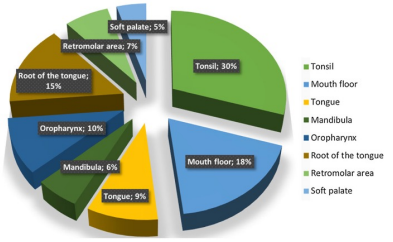
- **Données** :

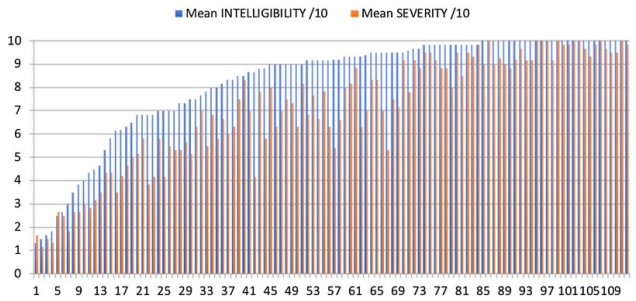
  - **cancer ORL** : 87 patients et 26 témoins

  annotations : sévérité, intelligibilité, prosodie…

  - **Maladie de Parkinson** : 205 patients et 111 témoins

  annotations : sévérité, intelligibilité…





Doctorat Robin Vaysse

Université Fédérale
Toulouse Midi-Pyrénées

THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE
Délivré par l'Université Toulouse 3 - Paul Sabatier

Présentée et soutenue par
Robin VAYSSE

Le 21 mars 2023

Caractérisation automatique du rythme de la parole : application aux cancers des voies aéro-digestives supérieures et à la maladie de Parkinson

Ecole doctorale : **EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse**

Spécialité : **Informatique et Télécommunications**

Unité de recherche :
IRIT : Institut de Recherche en Informatique de Toulouse

Thèse dirigée par
Jérôme FARINAS et Corine ASTESANO

Jury
M. François PELLEGRINO, Rapporteur
Mme Elisabeth DELAIS-ROUSSARIE, Rapporteure
Mme Cécile FOUGERON, Examinatrice
Mme Virginie WOISARD-BASSOLS, Examinatrice
M. Jérôme FARINAS, Directeur de thèse
Mme Corine ASTESANO, Co-directrice de thèse

Contribution 1 : extraction f0 pour voix pathologiques

JASA ARTICLE

Performance analysis of various fundamental frequency estimation algorithms in the context of pathological speech

Robin Vaysse,[1,a),b)] Corine Astésano,[2,c)] and Jérôme Farinas[1]
[1]*IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France*
[2]*Laboratoire de NeuroPsychoLinguistique, Université Toulouse Jean-Jaurès, France*

ABSTRACT:
Reliable fundamental frequency (f_0) extraction algorithms are crucial in many fields of speech research. The current bulk of studies testing the robustness of different algorithms have focused on healthy speech and/or measurements of sustained vowels. Few studies have tested f_0 estimations in the context of pathological speech, and even fewer on continuous speech. The present study evaluated 12 available pitch detection algorithms on a corpus of read speech by 24 speakers (8 healthy speakers, 8 speakers with Parkinson's disease, and 8 with head and neck cancer). Two fusion methods' algorithms have been tested: one based on the median of algorithms and one based on the fusion between the best algorithm for voicing detection and the algorithm that generates the most accurate f_0 estimations on voiced parts. Our results show that time-domain algorithms, like REAPER, are best for voicing detection while deep neural network algorithms, like FCN-f_0, yield better accuracy for the f_0 values on voiced parts. The combination of REAPER and FCN-f_0 yields the best ratio performance/implementation complexity, since it generates less than 4% errors on voicing detection and less than 5% of gross errors in the estimation of the f_0 values for all speaker groups.
 https://doi.org/10.1121/10.0015143
(Received 2 August 2022; revised 29 September 2022; accepted 26 October 2022; published online 29 November 2022)
[Editor: B. Yegnanarayana] Pages: 3091–3101

I. INTRODUCTION

The measurement of the fundamental frequency (f_0) is an essential element of automatic speech processing, particularly in the study of prosody. It is, therefore, crucial to have a good estimate of this parameter. Many algorithms have been developed for estimating the fundamental frequency of healthy speech recorded under good conditions (without noise), which provide very good f_0 approximations (see Sec. II). In the context of pathological speech, the calculation of precise f_0 variations is necessary because most pathologies have an impact on voice quality, more specifically on speakers' inability to maintain a stable fundamental frequency (jitter, shimmer) (Jiménez-Jiménez *et al.*, 1997). In addition, the dynamics of the fundamental frequency in a sentence defines the intonation, which corresponds to the voice "melody." Intonation provides main communicative functions and is a powerful tool for the illocutionary and structural interpretation of the speaker's message (Di Cristo, 2016). Yet, some pathologies can lead to a poor control of intonation that can induce confusion as to the type of sentence the speaker is trying to produce (Le Dorze *et al.*, 1994), which affects both his/her intelligibility and comprehensibility. When we want to model intonation or stress patterns from the f_0, these types of errors can lead to distorted interpretations over large time spans.

It is, therefore, crucial to use an f_0 extraction algorithm, which is as accurate as possible and avoids gross estimation errors (such as dividing by two or doubling the real value of the fundamental frequency) or errors in the detection of voiced or unvoiced areas. When working on large corpora of pathological voice recordings, such as Cesari *et al.* (2018), this issue is even more challenging because the amount of data does not allow for precise manual annotations. The objective of the present study is, therefore, to test several different algorithms in the particular context of pathological voice, such as those resulting from head and neck cancers (H&NC) or Parkinson's disease (PD).

Several performance evaluation studies of pitch detection algorithms have been designed on non-pathological speech (de Cheveigné and Kawahara, 2001; Strömbergsson, 2016) and it seems that auto-correlation function (ACF) from Praat (Boersma and Weenink, 2020) and the YIN algorithms (de Cheveigné and Kawahara, 2002) are good methods for typical, healthy voices. Some studies also looked into the evaluation of noisy speech, which best corresponds to real recording conditions (Jouvet and Laprie, 2017; Luengo *et al.*, 2007). These results show that, while all the evaluated algorithms provide comparable results on healthy speech, an increase in background noise results in a loss of algorithm performance, specifically with regard to the detection of voicing. More specifically, the robust algorithm for pitch tracking (RAPT) and the robust epoch pitch estimator (REAPER) algorithms seem to provide good results on

[a)]Also at: Laboratoire de NeuroPsychoLinguistique, Université Toulouse Jean-Jaurès, France
[b)]Electronic mail: robin.vaysse@irit.fr
[c)]Also at: UMR 5267 Praxiling - Université Paul Valéry Montpellier, France

TABLE I. List of algorithms tested in the present study, with a link to the chosen implementation. The last three columns indicate whether the algorithm works on the signal's time or spectral domain, or whether it uses deep learning.

Algorithm	Implementation	Time domain	Spectral	Neural network
ACF (Boersma, 2000)	Praat	X		
AMDF (Ross et al., 1974)	Snack Sound toolkit (Kåre, 2005)	X		
REAPER (Google-Open-Source, 2015)	https://github.com/google/REAPER	X		
RAPT (Talkin and Kleijn, 1995)	Snack Sound toolkit (Kåre, 2005)	X		
Enhanced RAPT (Ghahremani *et al.*, 2014)	Kaldi (Povey *et al.*, 2011)	X		
Yin (de Cheveigne and Kawahara, 2002)	https://github.com/patrice.guyot/Yin	X		
NDF (Kawahara et al., 2005)	STRAIGHT (Kawahara, 2018)	X	X	
YAAPT (Kasi and Zahorian, 2002)	MATLAB implementation (Zahorian and Hu, 2016)	X	X	
SWIPE (Camacho and Harris, 2008)	Speech signal processing toolkit (Tokuda *et al.*, 2017)		X	
PEFAC (Gonzalez and Brookes, 2014)	VOICEBOX (Brookes, 2018)		X	
CREPE (Kim et al., 2018)	https://github.com/marl/crepe			X
FCN-f_0 (Ardaillon and Roebel, 2019)	https://github.com/ardaillon/FCN-f0			X

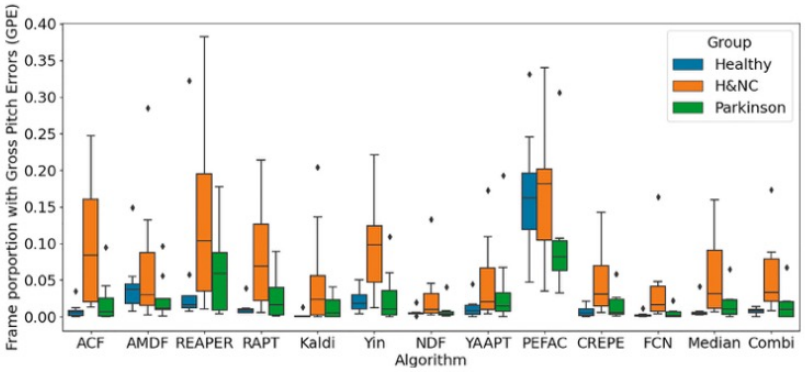
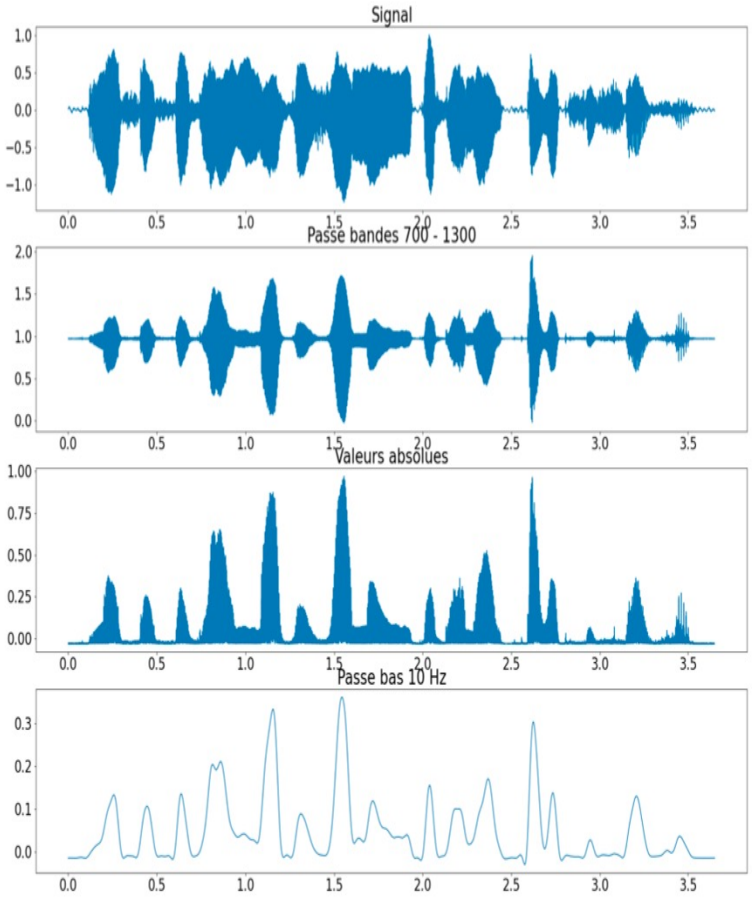


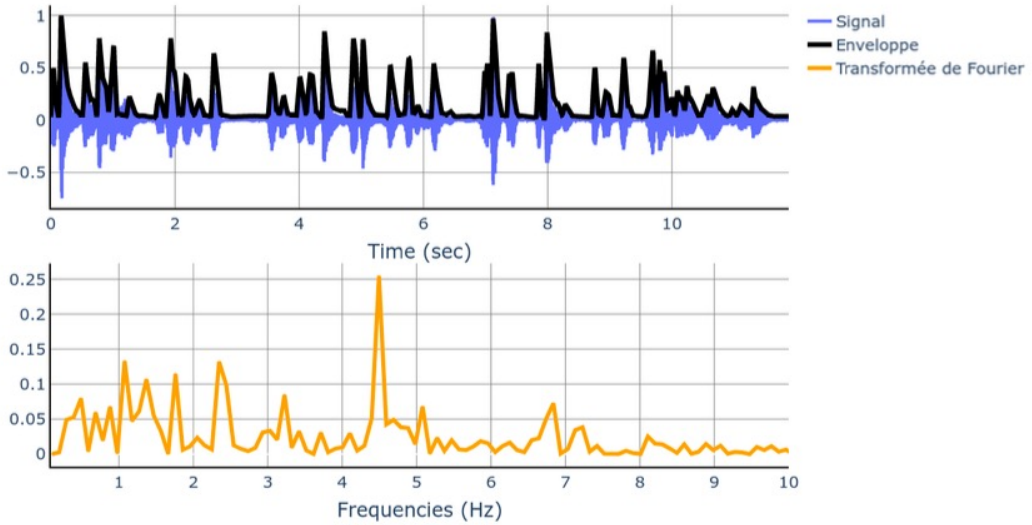
FIG. 5. (Color online) Results on gross pitch errors: Blue boxplots represent GPE for control speakers, orange boxplots are for speakers with H&NC, and green boxplots are for PD patients. Each boxplot represents an error percentage (lower percentages are better).

Algorithme à privilégier pour études cancer :

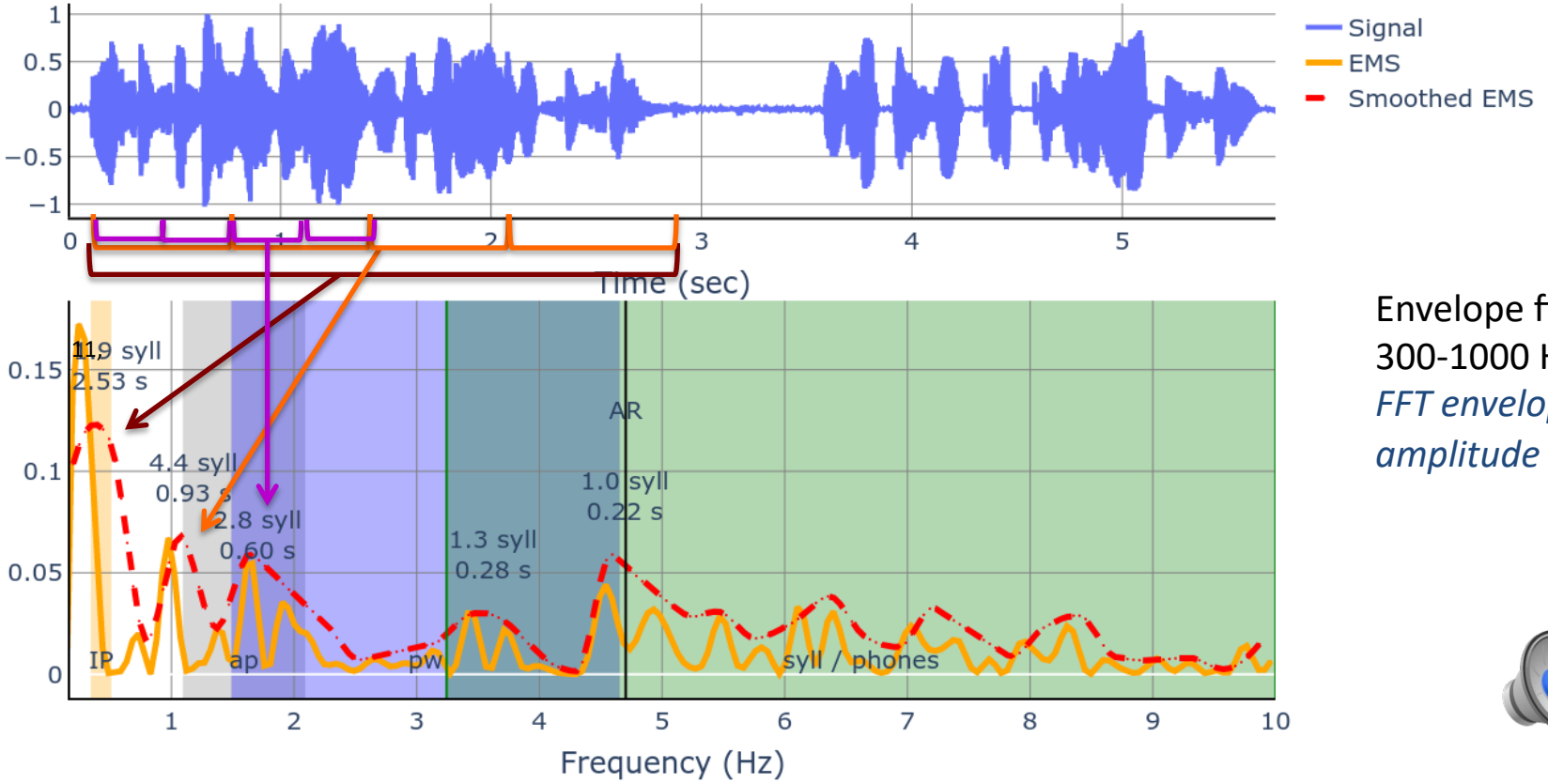
- Détection du voisement : REEPER (time domain)
- Détection des valeurs du f0 : FCN-f0 (neural net.)

Contribution 2 : spectre de modulation d'amplitude





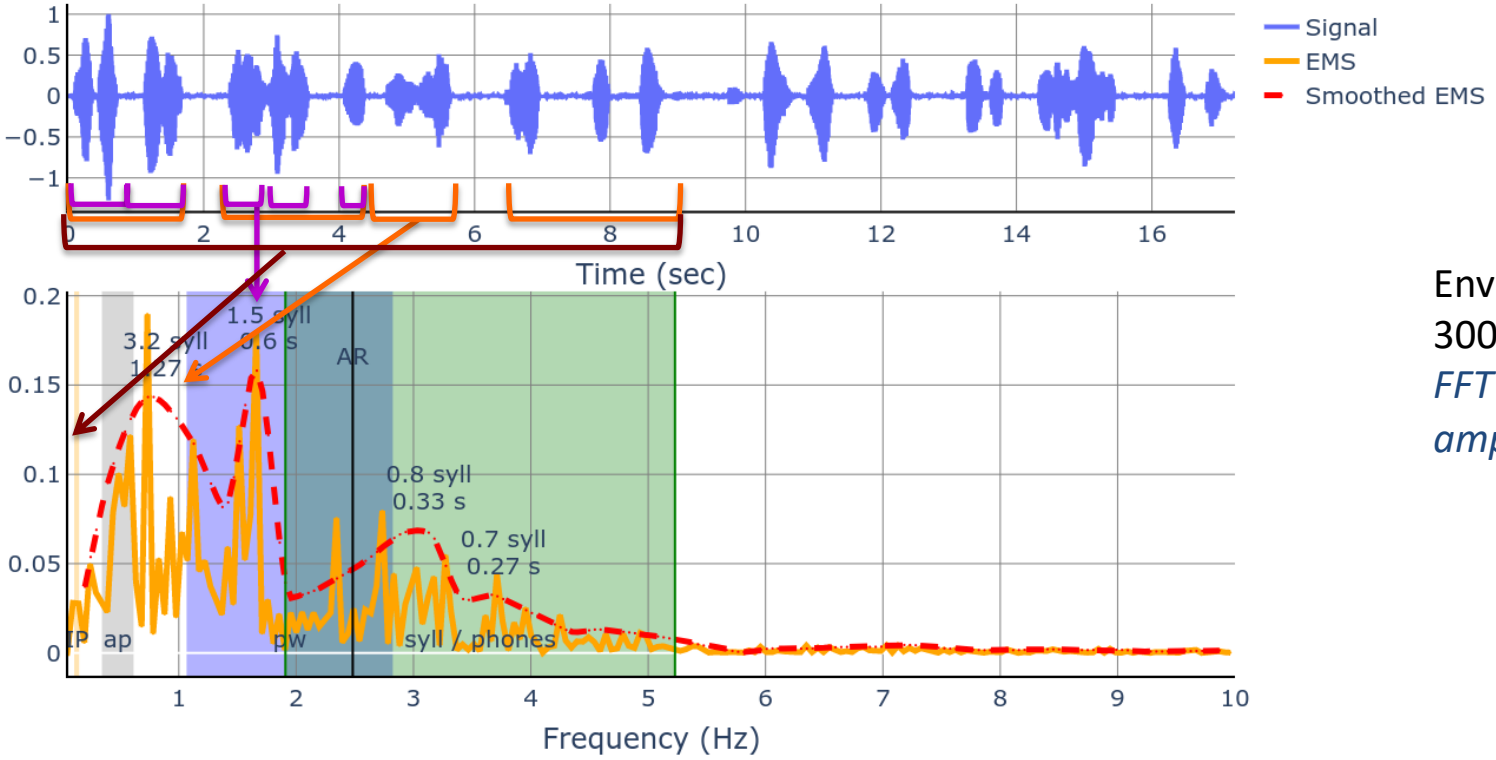
Fast Fourier Transform on Amplitude envelope (locuteur contrôle)



Envelope filtering :
300-1000 Hz
FFT envelope amplitude

‘Normal’ speaking rate (4,3 syll/sec)
PW (1,66 Hz; 600 ms), AP (1,1 Hz; 900 ms) and IP (0,5 Hz; 2,5 s)

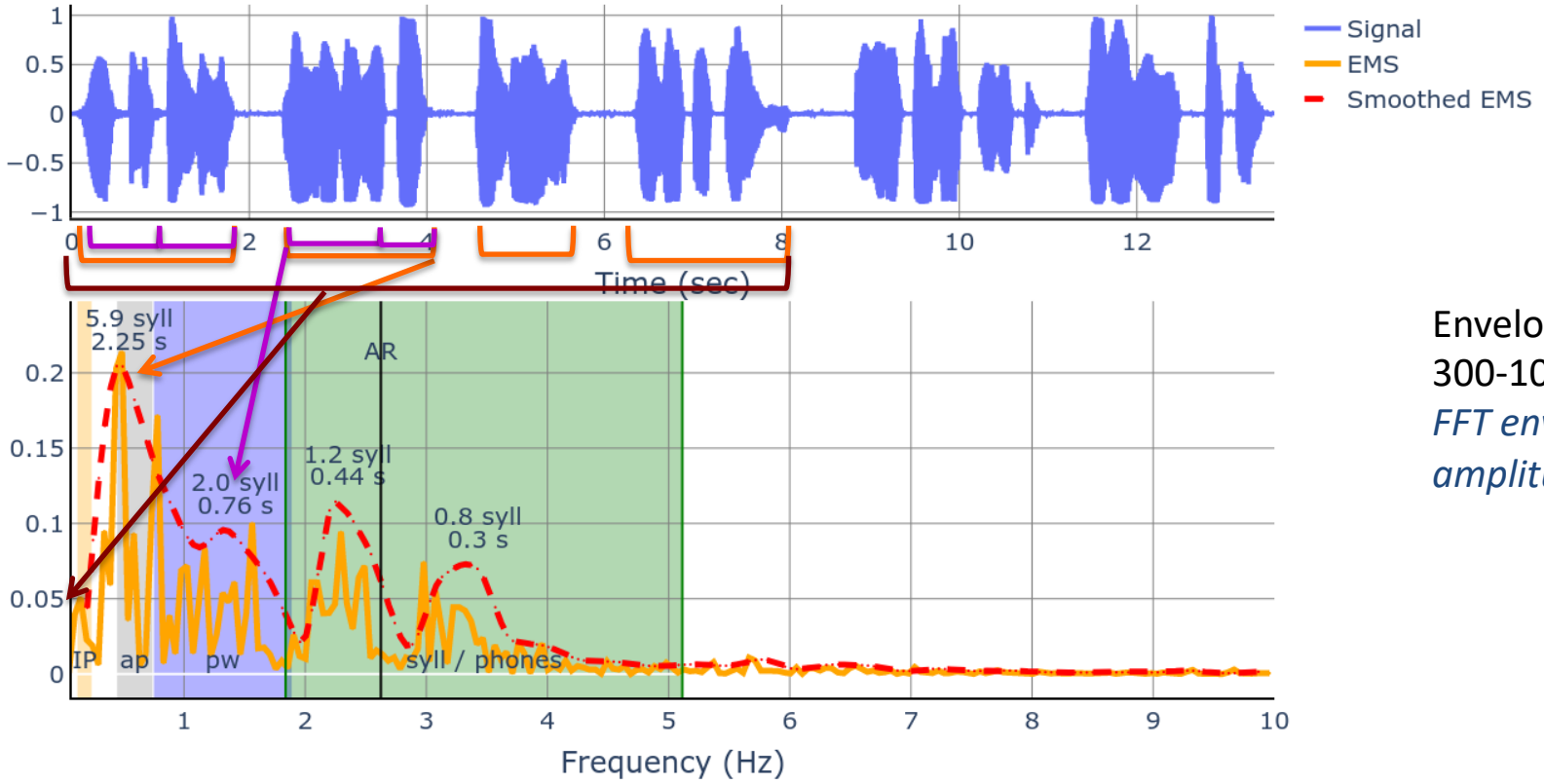
Fast Fourier Transform on Amplitude envelope (patient cancer)



Envelope filtering :
300-1000 Hz
FFT envelope amplitude

Haltingly speaker (word by word; 1,3 syll/sec)
PW (1,66 Hz; 600 ms), AP (0,77 Hz; 1,3 s.) and IP (0,1 Hz; 9 s.)

Fast Fourier Transform on Amplitude envelope (patient cancer, rythme fluide)

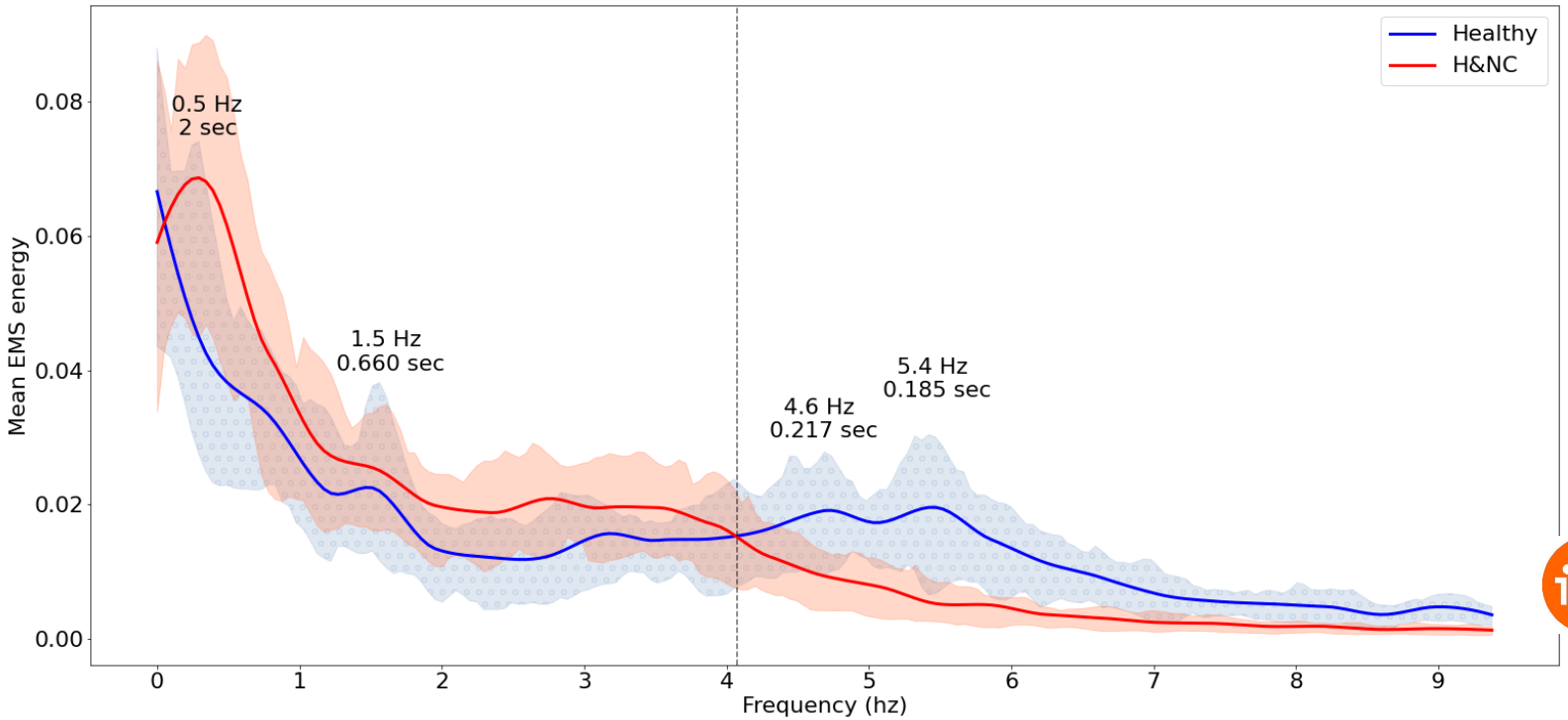


Envelope filtering :
300-1000 Hz
FFT envelope amplitude

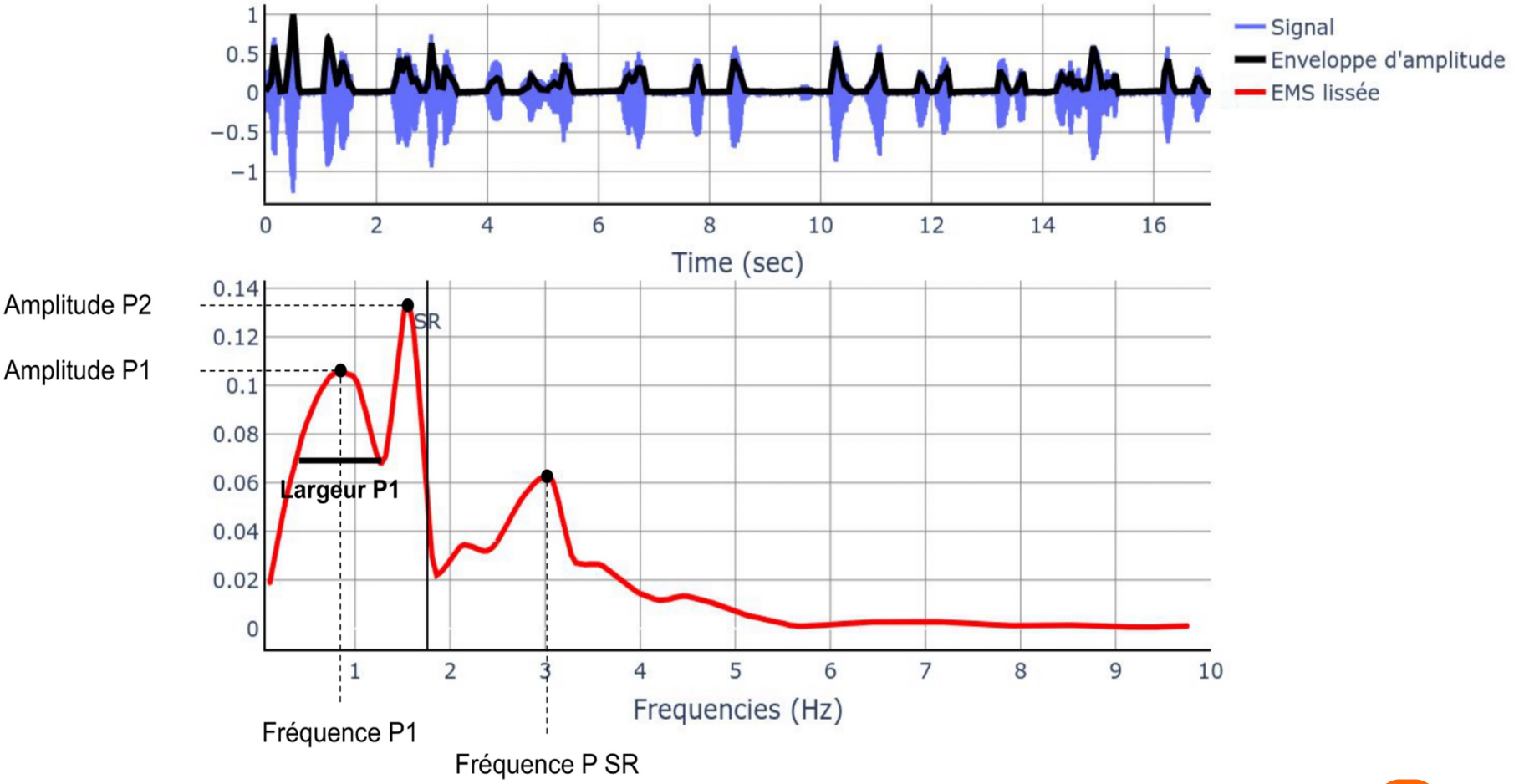
Fluent speaker with slow speaking rate (2 syll/sec)
PW (1,3 Hz; 760 ms), AP (0,44 Hz; 2,3 s.) and IP (0,13 Hz; 8 s.)

Compensation rythmique pour les patients cancer ?

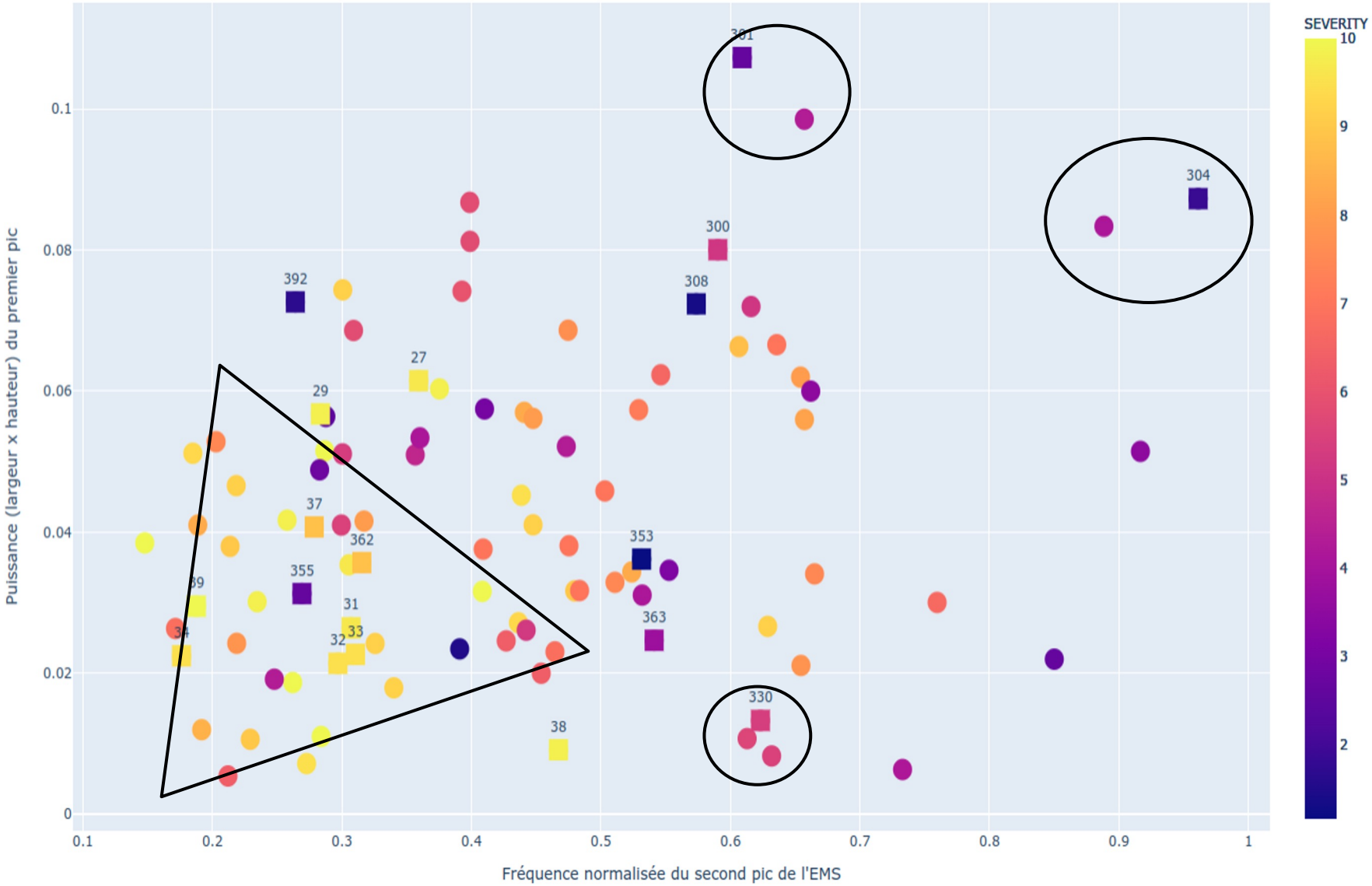
- PW : relativement stable autour de 650 ms
- AP descent en fréquence et remplace les IP “normaux”
- Perte d’information en dessous de la syllable : > 4 Hz = peu d’information sur l’articulation des phonèmes



Contribution 3 : caractérisation rythme



Projection sur tout le corpus



Perspectives

- Etudier l'enchaînement temporel (spectrogramme du rythme)
- Automatiser la detection des structures prosodiques
- Extraire des paramètres stables pour caractériser le rythme
- Etude de l'influence des silences
- Combiner spectre d'amplitude et de l'intonation