ATAXIC SPEECH DISORDERS AND PARKINSON'S DISEASE DIAGNOSTICS VIA STOCHASTIC EMBEDDING OF EMPIRICAL MODE DECOMPOSITION

Marta Campi
Institut Pasteur
marta.campi@pasteur.fr

Gareth W. Peters
University of Santa Barbara
garethpeters@ucsb.edu

Dorota Toczydlowska
University of Technology Sydney
dtoczydlowska@gmail.com

January 7, 2023

ABSTRACT

Medical diagnostic methods that utilise modalities of patient symptoms such as speech are increasingly being used for initial diagnostic purposes and monitoring disease state progression. Speech disorders are particularly prevalent in neurological degenerative diseases such as Parkinson's disease, the focus of the study undertaken in this work. We will demonstrate state-of-the-art statistical time-series methods that combine elements of statistical time series modelling and signal processing with modern machine learning methods based on Gaussian process models to develop methods to accurately detect a core symptom of speech disorder in individuals who have Parkinson's disease. We will show that the proposed methods out-perform standard best practices of speech diagnostics in detecting ataxic speech disorders, and we will focus the study, particularly on a detailed analysis of a well regarded Parkinson's data speech study publicly available making all our results reproducible. The methodology developed is based on a specialised technique not widely adopted in medical statistics that found great success in other domains such as signal processing, seismology, speech analysis and ecology. In this work, we will present this method from a statistical perspective and generalise it to a stochastic model, which will be used to design a test for speech disorders when applied to speech time series signals. As such, this work is making contributions both of a practical and statistical methodological nature.

***K*eywords** Empirical Mode Decomposition, Time-Frequency, Cross-Entropy, Gaussian Processes, Multi-Kernel Learning, Speech Recognition, Parkinson's Disease

1 Introduction

This paper addresses two components of statistical speech analysis for medical diagnostics. The first involves the detection and quantification of what is known as ataxic speech, one of the motor speech symptoms of many neurodegenerative disorders caused by damages to the cerebellar control circuit [1], with applications arising from Parkinson's disease. Secondly, it makes statistical contributions related to the development of non-linear and non-stationary time-series methods based on Empirical Mode Decomposition (EMD), where a stochastic model representation and statistical treatment of EMD are proposed not available just yet. This allows for the formalism of this method within a statistical testing context using Gaussian process models. We will show that the implemented methodology outperforms traditional speech methodologies covering this medical diagnostic task with accuracy scores greater than 80% on the data set collected and provided by King's College given at [2]. The introduction covers the statistical aspects of speech-based diagnostics and the motivation for EMD methods in this context.

The core aim of developing the statistical methods within this work is to utilise the EMD stochastic representations to explore the health diagnostic problem of detecting speech disorders associated with diseases such as Parkinson's. This is particularly topical for diagnostic purposes in settings in which mobility of those potentially afflicted with this debilitating condition are unable to easily commute regularly to hospitals for periodic monitoring of the progression of the disease, either due to illness or due to isolation and lockdown measures resulting from the Covid-19 pandemic

for instance. An early diagnosis through an automatic tool is of primary importance for the therapies to be as effective as possible [3]. Amongst the variety of time deficit impairments which affect multiple human skills and functions [4], significant research has been conducted in speech production (reviews of speech production and analysis are given at [5], [6]), particularly because speech appears to be the biometric feature carrying a great deal of discriminatory power since early symptoms can be detected and, possibly, monitored. Example of such research are [1], [7], [8], [9], [10], [11], [12], [13], [14], [15].

The ultimate ambition is to construct an automatic tool to detect and monitor the disease through a speech signal recorded through a phone or an alternative device (such as a watch or similar). Hence, multiple solutions have been provided recently, relying on machine learning, statistics and deep learning techniques. The ideal tool should address the problem of surveillance the different stages of the disease as well as to understand which factors could affect such a a disease. In speech, many cited methodologies, corresponding to some of the most recent, tend to follow two main approaches: the former refines standard glottal or voice features and then uses a machine learning classifier, while the latter searches for a more complex implementation of the classifier relying on deep learning techniques.

We believe that, when it comes to speech tasks (as highlighted in [16]), two critical aspects should always be considered: the task must be performed by taking gender into account since male and female voices differ a lot in terms of resonant frequencies of the vocal cords and conducting a joint classification would pollute any classifier. Secondly, the employed classifier should be able to guarantee a physical interpretation of the obtained results, i.e. features better performing should reflect the discriminatory power carried by female or male voices. Such a challenge needs the development of ad hoc sophisticated methodologies since an accuracy of at least 80% should be achieved in health diagnostic. This paper aims to fill this gap by designing a stochastic version of the EMD.

1.1 Introduction to Speech Analysis for Medical Diagnostics

It has been known for some time by medical practitioners that numerous neurological and motor-neuron diseases may manifest with symptoms that impact speech through enunciation difficulties, slurring of words, delayed recall leading to unvoiced pauses of unusual duration, stutter or other documented speech disorders such as ataxic speech effects. Many such diseases are degenerative and require continuous monitoring of the patient's status to ensure that treatment regimes adapt to the disease progression for each individual. This can also result in degeneration of speech in such patients, reducing their communication ability over time, thereby making it hard for doctor-patient and caregiver-patient communications.

This paper focuses on the emerging area of medical diagnostic methods that detect symptoms of diseases and disease state progression from speech recordings. The ability to detect speech symptoms can result in earlier diagnostic screening processes, increased care provision when required as a disease progresses, and aid provision in communication. Speech anomalies from such diseases are increasingly being able to be accurately detected in real-sensing environments without the need for specialist recording equipment and sound laboratory environments. This makes the use of such diagnostic modalities offered by speech a new and valuable tool for medical practitioners that can accurately allow insight in standard environments of speech recording, such as a doctor's office, outside in public or via mobile phone. This is made possible using modern time series analysis methods. See examples in Parkinson's disease in [17], Alzheimer's in [18] and a recent tutorial on such applications in [19].

The detection of such speech disorders can be valuable for three reasons: firstly, it can act as one of several prescreening diagnostic tools to prioritise automatically or not further to test for a disorder consistent with a particular speech disorder detected; secondly, it can allow for analysis of automation of one important component of disease progression and monitoring of the ability to communicate effectively for patients with degenerative disorders; and thirdly it can be highly effective methods for remote and telemedicine practices when some diagnostic tests can be achieved through recordings captured from a mobile phone speech record with no special recording environment conditions. The focus of this work will be particularly on particular speech disorders associated with neuro-degenerative disorders such as Parkinson's disease, and the real data case study will be illustrated on analysis of patients at various stages of this disease.

The Parkinson's disease speech diagnostic analysis undertaken in the application framework developed in this paper builds upon the background proposed in [20]. In this work, the authors introduce an alternative method to detect speech abnormalities caused by Cerebellar Ataxia. This corresponds to impaired coordination due to a cerebellum dysfunction, characterised by movement abnormalities such as dysmetria, dysdiadochokinesia, and dyssynergia, amongst others. These abnormalities affect all kinds of movements, including speech, and hence lead to what is termed "ataxic speech". Signs of ataxic speech could be scanning speech ("excess and equal stress"), a reduced speech rate and deviant prosodic (i.e. rhythmical and melodic), modulation of verbal utterances, rhythmical irregularities during (fast) repetitive productions of single or multiple syllables (known as "oral dysdiadochokinesis"), a more significant varia-

tion in pitch and loudness and disturbed articulation of both consonants and vowels with reduced intelligibility (see [21], [22], [23]).

Several medical conditions could generate ataxic speech; in this work, we are interested in speech impairment movements caused by Parkinson's disease (see [24]). Parkinson's disease is a degenerative disorder of the central nervous system resulting from the death of dopamine-containing cells in the substantia nigra, a region of the midbrain. It is the second most common neurodegenerative disorder after Alzheimer's disease [25], [26], [27] and includes both motor (tremor, rigidity, bradykinesia, and impairment of postural reflexes) and non-motor signs (cognitive disorders and sleep and sensory abnormalities). Several studies reported a 70-90% prevalence of speech impairments once the disease makes its appearance (see [24]). Moreover, it might be one of the earliest Parkinson's disease indicators (see [28],) with research showing that 29% of patients consider it one of their greatest obstacles ([29]). Both motor symptoms and speech movements abnormalities worsen with the progression of the disease in a nonlinear fashion ([28], [30]]. At the final stage of the disease, articulation is frequently the most impaired feature (see [24], [31], [32]). Medical treatments or surgical intervention can alleviate the course of the disease; however, there is no definite cure, and, therefore, an early diagnosis is highly critical to lengthen and improve the patient's life ([33], [34]).

1.2 Introduction to Time Series Empirical Mode Decomposition

From a statistical perspective, speech data sets are a collection of complex time series records that can be analysed using advanced time series and speech signal processing methods. Time series decomposition techniques represent a powerful tool for analysing time and frequency domain structure in observed signals such as speech signals. If designed appropriately, the extracted basis components reveal hidden insights into the time-frequency structure of the data generating process more efficiently than the analysis of the original signal per se. The decomposition method of interest in this work is in the class of time-frequency methods ([35], [36]) and is known as the Empirical Mode Decomposition (EMD) ([37]). Two core differences arise when contrasting EMD against more traditional time series decomposition methods. The first is that, unlike most traditional time-frequency decomposition methods, which specify a specific functional form for the basis, such as cosines for Fourier methods or wavelet functions for wavelet methods, the EMD method is not prescriptive of the functional form of the basis used. Instead, it can be considered characteristic since it only specifies the properties that the IMF basis functions must satisfy, not what functional form they should take to achieve these required properties. Secondly, as a result of the flexibility that EMD characterisation offers in allowing for its basis functions to be used, such as splines, the second core difference that results is that the EMD can relax requirements for statistical assumptions such as linearity or stationarity often required for more traditional time series decomposition methods to be applied. Despite these critical practical features of the EMD method, to date, there has been no statistical formalisation of a stochastic representation or stochastic embedding of the empirical algorithmic approach that the EMD method offers, and we address this challenge in this manuscript. This is important to be developed as to undertake statistical analysis tasks such as estimation and inference as well as to incorporate accurately statistical uncertainty quantification in out-of-sample predictions and forecasts, distinct from extrapolation; it is highly beneficial to obtain a stochastic representation of EMD consistent with the path-wise features of the decomposition basis characteristics.

An advantage of the EMD is that the basis functions, known as Intrinsic Mode Functions (IMFs), have characteristic properties that they produce a collection of monocomponent basis functions, see discussion in [38]. Recall, a monocomponent signal is described in the time-frequency (t,f)-domain by one single 'ridge', corresponding to an elongated region of energy concentration. In addition, considering the crest of the 'ridge' as a graph of Instantaneous Frequency (IF) vs time, one requires the IF of a monocomponent signal to be a single-valued function of time. It is precisely this feature that allows one to form the analytic extension of each EMD basis function IMF via a now well-defined Huang-Hilbert transform to explicitly characterise the collection of time-frequency decomposition complement representations obtained as the IFs corresponding to the IMFs. Furthermore, for each IMF, since such a monocomponent signal has an analytic associate of the form as a complex exponential, this can be interpreted as generalising the Fourier representation by allowing the amplitude and phase to become time functions, producing what is known as functional-coefficient basis representations. It is this feature that allows EMD to flexibly adapt to non-stationary and non-linear structures in its time series representation. The EMD method then utilises the fact that a multicomponent signal may be described as the sum of two or more monocomponent signals. The concept of IF has been explored in numerous works, see [39], [40], [41], [35]. A basis decomposition method utilising such characterising features can capture both time and frequency events in a localised fashion, which is extremely useful when there are non-stationarity effects present.

Developing a stochastic representation or embedding along with a family of statistical model representations for the EMD method to complement its algorithmic formulation will be achieved by considering three methodological problem statements addressed in this paper. The first problem is to establish a path-wise statistical model for the IMFs,

satisfying the definitions provided in [37] that will also be consistent with the developed stochastic representation. The second problem statement considers the assumption that the EMD is algorithmically applied to the realisation of a time series signal sampled from an unknown stochastic process. Given the realised time series, the IMFs, per path, are then considered deterministic unknown functions that must be estimated from the samples. Therefore, in PS2, we seek to determine a stochastic version of the IMF decomposition compatible at a population process level with the pathwise representation of the deterministic decomposition being estimated under the solution to PS1.

Given that we will work with spline model representations as the solution to the first problem posed, it becomes natural to consider whether Gaussian Processes (GP) ([42]) stochastic model embeddings will satisfy the solution to PS2, when stochastically embedding the IMFs into a stochastic model representation. In this work's context, a Gaussian process will be considered a continuous-time stochastic process for which all finite-dimensional distributions follow multivariate normal distributions. One may then interpret the GP as a random variable on $L^2([0,1])$ such that the individual sample paths mapping $[0,1] \rightarrow \mathbb{R}$ are considered random functions. In particular, there is a known connection between such functions when they are represented by splines, which under appropriate conditions are known to be suitable sample path realisations for Gaussian processes, see [43]. The challenge will be to ascertain whether this class of GP stochastic models will sufficiently satisfy the requirements imposed on the characteristic properties that such a representation should capture if it is to adequately represent an EMD decomposition as a stochastic representation.

In addition, it is convenient to have a stochastic representation that has practical utility when performing tasks such as estimation, inference and forecasting under the stochastic model representation of EMD decompositions. Otherwise, there is little value practically in developing such a stochastic representation. Fortunately, GPs are a robust inference supervised machine learning technique used in many applications, given that they can be entirely specified by their mean and covariance, or kernel, functions. We will demonstrate that the GP stochastic representation we will develop for EMD basis functions IMFs when aggregated together to represent the original signal, can be considered as a special class of multi-kernel (MKL) GP (see review in [44]) stochastic model representation of the original time series signal. In practice, the EMD is then learning the multi-kernel spectral decomposition in terms of the number of kernel components to consider and their characteristic time-frequency structure for each kernel component. MKL representations can be achieved through multiple strategies developed in the literature. We cite among others hierarchical kernel learning ([45, 46]) or the alternatives proposed by [47], [48], who suggested a multi-kernel frequency approach acting on the power spectral density of the GP kernel function.

The third problem addressed in this work pertains to the suitable selection of the covariance function used to adequately capture the IMFs being stochastically modelled by GPs. Since IMFs will correspond to a collection of non-stationary basis functions, there is a requirement to properly design the family of kernel functions to accurately model the IMF spline representations estimated under the EMD basis extraction procedure, known as sifting. In this regard, in non-trivial applications such as speech analysis focused on in this manuscript, standard parametric kernels such as the Matern kernel and the RBF kernel (see [42]) will not suffice. Instead, we will develop two classes of solutions to this problem that generate two different families of stochastic model GP representations of EMD decompositions. The first is based on a family of data-adaptive kernels known as the Fisher kernel ([49], [50],[51],[52]), which provides a generic mechanism incorporating generative probability models into the development of the covariance operator that will be data-adaptive and act as a flexible time series kernel. The second approach is based on a novel framework to learn optimal partitions of the time-frequency plane that utilises the IFs obtained from the EMD basis IMFs to partition the energy spectrum into localised regions that can then be modelled via localised GPs. One of the challenges with this second approach is how best to learn the time-frequency partition rule. This is solved via a novel application of Cross Entropy optimisation (CEM), which is a stochastic optimisation technique that Rubinstein first presented in 1999 (see [53], [54]). Once the optimal core bandwidths are computed, a new set of frequency band-limited bases we term "band-limited" IMFs (BLIMFs) will be derived. These new set of basis functions are obtained by aggregating the original IMFs sample points according to the location of their IFs within the regions of the computed optimal bandwidths partition. With such a partition model, we can characterise adaptive local bandwidths of the IMFs frequency domain with a kernel function in a Gaussian process setting.

1.3 Contributions, Notation and Structure

The contributions of this work are given as follows:

- A stochastic embedding model representation is developed for the EMD basis decomposition method that is consistent with the characterising properties that the EMD method requires for the IMFs. The focus of this stochastic representation will also be compatible with the setting in which the IMFs are characterised by statistical models comprised of B-spline and P-spline representations, as well as proposing flexible statistical models that readily lend themselves to estimation, inference and statistical forecasting methods for EMD decompositions.

- In order to develop a family of statistical models for the proposed stochastic representation of EMD, a multi-kernel Gaussian Process framework is proposed. The particular features comprise a kernel construction suitable for modelling the non-stationary IMF basis GP spline representations. This uses a time series kernel representation based on a data-adaptive generative model embedding solution constructed via a Fisher Kernel.
- A second, localised stochastic solution, is also developed that defines an optimal set of band-limited basis functions stochastic model representations providing the following advantages: (1) one can focus on modelling specific bandwidths which might be significant for the application of interest; (2) one can formulate a set of basis functions whose marginal distributions are closer to a stationary distribution, compared to the original IMFs. Modelling the covariance function of such basis functions through a certain kernel is less challenging and will provide a more efficient solution for the MKL GP model representation.
- The cross-entropy method is introduced in this context to find an optimal time-frequency partition which is fully data-adaptive.
- A novel solution to speech diagnostics for Parkinson's disease diagnostics and disease progression quantification is developed, which, when compared to state-of-the-art existing methods, is shown to be more sensitive and accurate for both the detection of early onset of Parkinson's disease as well as the quantification of disease progression. The solution is ultimately based on the stochastic EMD representations developed via the MKL GP model representations class.

The following notation will be used throughout: $t_0 < t_1 < \ldots < t_N$ denotes signal observation times; the time series signal is denoted by $s(t) : \mathcal{T} \to \mathbb{R}$ and is observed at $\{s(t_i)\}_{i=1}^N$; the continuous time spline reconstruction of the signal is denoted by $\tilde{s}(t) : \mathcal{T} \to \mathbb{R}$; the L IMF basis function from the EMD method are denoted by $\{\gamma_l(t)\}_{l=1}^L$ such that each satisfies $\gamma_l(t) : \mathcal{T} \to \mathbb{R}$; L generically denoted the total number of IMFs extracted for a given signal; the analytic extension of the l-th IMF will be denoted by $\check{\gamma}_l(t) = \mathcal{H}[\gamma_l(t)]$ where $\mathcal{H}[\cdot]$ denotes the Hilbert transform which produces the analytic signal $z_l(t) = \gamma_l(t) + i\check{\gamma}_l(t)$; $\mathcal{F}[\cdot]$ will denoted the Fourier transform; when extracting IMF basis functions under the EMD method sifting algorithm, we will denote by $\tilde{s}^{U_l}(t)$ the upper envelope used in sifting that is a spline interpolating the maximum of the current best estimate of the l-th IMF and analogously by $\tilde{s}^{B_l}(t)$ the lower envelope of the l-th IMF interpolating the minimum of the current best estimate of the l-th IMF in the iterative IMF extraction algorithm known as sifting; finally, we will denote the collection of frequency band limited IMFs by $\left\{\gamma_m^{(BL)}(t)\right\}_{m=1}^M$ the band-limited IMF construction based on M total specified bandwidths.;

The paper is organised as follows: firstly, a review of the EMD method is shown. We refer to [16] as main reference. Secondly, the EMD stochastic embedding set up is proposed with a set of objectives that must be satisfied. Afterwards, the stochastic embedding is formally developed, with the required notions presented to achieve it. Note that, three different system models will be formulated in this section: one for the stochastic embedding of the original signals and two which are the ones relating to the EMD and proposed in this manuscript. Section 5 presents how to develop a generative embedding kernel based on the Fisher kernel. Furthermore, the formulation of the cross-entropy problem with the derived solution used to formalise an optimal time-frequency partition for the second stochastic embedding is presented. Section 6 introduces the framework of speech based medical diagnostic with a subsection on motivation for Parkinson's speech detection, a subsection standard benchmark model solving this task and the GLRT Test used to test the presence or absence of Parkinson's disease developed in this paper. The last section shows the experiments results and discussion conducted on the speech data for Parkinson's detection.

2 Statistical Model Framework for Empirical Mode Decomposition

This section introduces a formalism required to understand the EMD method and builds upon the work presented in [16]. EMD basis characteristics of IMFs have been defined in [37] through a set of non-constructive properties only and are obtained via a procedure known as sifting, based on a recursive extraction of the signal energy associated with the intrinsic time scales of the original signal. They are therefore ordered according to their number of oscillations or convexity changes, and they furthermore satisfy the property that their sum reproduces the original realised signal path. Hence, the observed time series is reconstructed in principle exactly when the resulting IMFs are estimated or extracted numerically in a manner that perfectly satisfies the characterising properties of the EMD method.

Consider a continuous non-stationary speech signal $s(t)$ observed as a sample recording at times $0 = t_1 < \cdots < t_N = T$. When applying the EMD basis decomposition framework, we first convert the partially observed discrete time signal $s(t)$ into a continuous time analog signal, denote by $\tilde{s}(t)$. To achieve this we use a natural cubic polynomial spline. We will also express the EMD bases $\{\gamma_l(t)\}_{l=1}^L$ as natural cubic splines, derived from representation $\tilde{s}(t)$.

definition 2.1. Given a set of l knots $a = \tau_1 < \tau_2 < \cdots < \tau_l = b$, a function $\tilde{s} : [a, b] \to \mathbb{R}$ is called a cubic polynomial spline if:

- $\tilde{s}(\cdot)$ is a polynomial of degree 3 on each interval (τ_j, τ_{j+1}) $(j = 1, \ldots, l-1)$
- $\tilde{s}(\cdot)$ is twice continuously differentiable

It is then a natural cubic spline when $\tilde{s}''(a) = \tilde{s}''(b) = 0$.

Hence, the speech signal representation $\tilde{s}(t)$ is expressed in the class of truncated power basis, where the knot points are placed at the sampling times ($\tau_i = t_i$)

$$\tilde{s}(t) = a_0 + a_1 t + a_2 t^2 + a_3 \left(t - \tau_1\right)_+^3 + \cdots + a_{3+l-2} \left(t - \tau_{l-1}\right)_+^3 .$$

The coefficients are estimated by standard penalised least squares

$$\sum_{i=1}^{N-1} (s(t_i) - \tilde{s}(t_i))^2 + \lambda \int_{t_i}^{t_{i+1}} \tilde{s}''(t)^2 dt$$

with natural cubic spline constraints $\tilde{s}''(0) = \tilde{s}''(t_N) = 0$ and where $\lambda > 0$ controls smoothness of the representation. In this case, the number of total convexity changes (oscillations) of the analog signal $\tilde{s}(t)$ within the time domain $[0, t_N]$ is denoted by to $L \in \mathbb{N}$. One may now define the EMD decomposition of a speech signal $\tilde{s}(t)$ as follows.

definition 2.2 (Empirical Mode Decomposition)**.** *The Empirical Mode Decomposition of signal $\tilde{s}(t)$ is represented by the finite number of non-stationary basis functions known as Intrinsic Mode Functions (IMFs), denoted by $\{\gamma_l(t)\}$, such that*

$$\tilde{s}(t) = \sum_{l=1}^{L} \gamma_l(t) + r(t) \tag{1}$$

where $r(t)$ represents the final residual (or final tendency) extracted, which has only a single convexity. In general the γ_l basis will have l-convexity changes throughout the domain (t_1, t_N) and each IMF satisfies:

- ***Oscillation*** *The number of extrema and zero-crossing must either equal or differ at most by one;*

$$abs\left(\left|\left\{\frac{d\gamma_l(t)}{dt} = 0 : \quad t \in (t_1, t_N)\right\}\right| - \left|\{\gamma_l(t) = 0 : \quad t \in (t_1, t_N)\}\right|\right) \in \{0, 1\} \tag{2}$$

- ***Local Symmetry*** *The local mean value of the envelope defined by a spline through the local maxima denoted $\tilde{s}^{U_l}(t)$ and the envelope defined by a spline through the local minima denoted by $\tilde{s}^{B_l}(t)$ is equal to zero pointwise i.e.*

$$m_l(t) = \left(\frac{\tilde{s}^{U_l}(t) + \tilde{s}^{B_l}(t)}{2}\right) \mathbb{I}\left(t \in [t_1, t_N]\right) = 0 \tag{3}$$

The minimum requirements of the upper and lower envelopes are:

$$\begin{aligned}
&\tilde{s}^{U_l}(t) = \gamma_l(t), \quad if \quad \frac{d\gamma_l(t)}{dt} = 0 \quad \& \quad \frac{d^2\gamma_l(t)}{dt^2} < 0, \\
&\tilde{s}^{U_l}(t) \geq \gamma_l(t) \quad \forall t \in (t_1, t_N) \\
&\tilde{s}^{B_l}(t) = \gamma_l(t), \quad if \quad \frac{d\gamma_l(t)}{dt} = 0 \quad \& \quad \frac{d^2\gamma_l(t)}{dt^2} > 0, \\
&\tilde{s}^{B_l}(t) \leq \gamma_l(t) \quad \forall t \in (t_1, t_N).
\end{aligned} \tag{4}$$

This definition provides characteristic properties that an IMF basis, $\gamma_l(t)$, under the EMD method should satisfy. Evidently, it is not constructive, i.e. prescriptive of the functional form of the basis. Therefore, in this manuscript, we opt to utilise throughout the same flexible natural cubic spline representation as used to represent the speech signal interpolation $\tilde{s}(t)$ also for the IMFs. Such a B-spline based representation for the realised deterministic basis decomposition that makes up the statistical model for the EMD pathwise representation will be essential to motivate the use of the Gaussian process stochastic model embedding for the stochastic process based representation we develop for the EMD method.

One can note that each IMF carries a unique number of convexity changes that can occur at any time spacings. Typically, the times of convexity change are irregularly spaced and reflect non-stationarity in a local bandwidth of the

frequencies that characterize the signal at that time instant. As a result of this property, one can still order the basis IMF's naturally according to the unique number of total convexity changes they produce in (t_1, t_N).

As outlined in [37], the construction of an IMF basis is directly linked to the concept of local symmetry required to handle non-stationary data. This notion is enclosed by the mean envelope that captures a local time scale, and the definition of a local averaging time scale is hence bypassed. Such a requirement is fundamental to avoid asymmetric waves affecting the concept of instantaneous frequency, formalised below.

2.1 Extraction of EMD Basis Functions Intrinsic Mode Functions (IMFs): The Sifting Procedure

We briefly outline the process applied to extract recursively the IMF basis representations, which is a procedure known as *sifting*, see [55]. To extract the l-th IMF The first step consists of computing extrema of the current signal representation after having removed the previously extracted IMFs by $\tilde{s}_l(t) := \tilde{s}(t) - \sum_{i=1}^{l-1} \gamma_i(t)$, which still admits a spline representation. Using the spline representation of $\tilde{s}_l(t)$ one needs to find the roots of the first derivative $\tilde{s}_l'(t)$ to produce the sequence of time points for successive maxima and minima given by:

$$\left\{t_j^*\right\}_{l=1}^{L} = \left\{ t \in [t_1, t_N] \ : \ a_1 + 2a_2 t + 3 \sum_{i=3}^{3+l-2} a_i \left(t - \tau_1\right)_+^2 = 0 \right\}.$$

Without loss of generality, we assume the maxima occur at odd intervals, i.e. t_{2j+1}^*, and minima occur at even intervals, i.e. t_{2j}^*. The second step of sifting builds an upper ($\tilde{s}^{U_l}(t)$) and lower ($\tilde{s}^{B_l}(t)$) envelope of $\tilde{s}_l(t)$ using two natural cubic splines through the sequence of maxima and the sequence of minima respectively:

$$\tilde{s}^{U_l}(t) = a_0^{U_l} + a_1^{U_l} t + a_2^{U_l} t^2 + \sum_{i=0}^{\lfloor L/2 \rfloor} a_{i+3}^{U_l} \left(t - t_{2i+1}^*\right)_+^3,$$

$$\tilde{s}^{B_l}(t) = a_0^{L_k} + a_1^{B_l} t + a_2^{B_l} t^2 + \sum_{i=0}^{\lfloor L/2 \rfloor} a_{i+3}^{B_l} \left(t - t_{2i}^*\right)_+^3,$$

such that $\tilde{s}^{U_l}(t) \geq \tilde{s}_l(t)\ \forall t$ with $\tilde{s}^{U_l}(t_{2j+1}^*) = \tilde{s}_l(t_{2j+1}^*)$ for all odd t_j^* and strictly greater otherwise; and equivalently $\tilde{s}^{B_l}(t) \leq \tilde{s}_l(t)\ \forall t$ with $\tilde{s}^{B_l}(t_{2j}^*) = \tilde{s}_l(t_{2j}^*)$ for all even t_j^* and strictly less than otherwise. One then utilises these envelopes to construct the mean signal denoted by $m_l(t)$ given in Eq. (3), which will then be used to compensate the current representation of the speech signal by $\tilde{s}_l(t) = \tilde{s}_l(t) - m_l(t)$ at each time point $t \in [t_1, t_N]$. This procedure is then repeated on the compensated signal, where again the current maxima and minima are obtained to produce envelopes which in turn produce a new estimate of the mean $m_l(t)$ which in turn is used in a defluctuation step to compensate the signal $\tilde{s}_l(t)$. This is repeated until the conditions specified in Definition 2.2 for the envelope and mean functions are satisfied, which when achieved produce the current deflucuated version of the signal $\tilde{s}_l(t)$ as the l-th IMF $\gamma_l(t)$. This procedure then repeats again for the $l+1$-th IMF extraction working now on signal $\tilde{s}_{l+1}(t) := \tilde{s}(t) - \sum_{i=1}^{l} \gamma_i(t)$, and the entire sifting process terminates when the $L+1$-st IMF is extracted and it corresponds to the IMF 'tendency' which only has one convexity change in $[t_1, t_N]$ and is often denoted distinctly by $r(t)$, see [16] for an algorithm and further details.

2.2 Obtaining Instantaneous Frequencys (IFs) from IMF Basis Functions

The EMD method extracts a set of basis functions (IMFs), each of which will admit a time-varying frequency structure that can be characterized by their corresponding instantaneous frequeny (IF) signal. The IF of a given IMF basis is extracted in the following stages.

First, one takes the Hilbert Transform of each IMF $\{\gamma_l(t)\}_{l=1}^{L}$, in order to construct a set of analytic extensions $\{\check{\gamma}_l(t)\}_{l=1}^{L}$ via the Hilbert transform as follows:

$$\check{\gamma}_l(t) = \mathcal{H}\left[\gamma_l(t)\right] = \frac{1}{\pi} \lim_{\epsilon \to \infty} \int_{-\epsilon}^{+\epsilon} \frac{\gamma_l(\tau)}{t - \tau} d\tau$$

which then produces the collection of analytic signals $\{z_l(t)\}$ with $z_l(t) = \gamma_l(t) + \check{\gamma}_l(t)$. We observe that when $\gamma_l(t)$ is a proper IMF such that it respects the restrictions defined in (4), its Hilbert transform can be obtained in closed form. The complex analytical signal $z_l(t)$ can be then represented by the polar representation $z_l = a_l(t) e^{\jmath \theta_l(t)}$ with time varying amplitude $a_l(t) = \sqrt{\gamma_l^2(t) + \check{\gamma}_l^2(t)}$ and time varying phase $\theta_l(t) = \arctan \frac{\check{\gamma}_l(t)}{\gamma_l(t)}$.

The instantaneous frequency $\omega_l(t)$ for IMF $\gamma_l(t)$ is then found from the time-varying phased of $z_l(t)$ as the rate of change given by:

$$\omega_l(t) = \frac{1}{2\pi}\frac{d\theta_l(t)}{dt} = \frac{1}{2\pi}\frac{\check{\gamma}_l'(t)\gamma_l(t) - \check{\gamma}_l(t)\gamma_l'(t)}{\gamma_l^2(t) + \check{\gamma}_l^2(t)}.$$

As observed in [37] conditions (4) that characterize the IMF properties are specified to ensure that the instantaneous frequency remains positive and therefore admits a meaningful physical interpretation.

Since, we adopt a statistical model representation for the IMFs based on cubic splines one can utilise this representation of the l-th IMF to obtain the Hilbert transform of the sum of local cubic polynomial transforms, see for details [56]:

$$\check{\gamma}_l(t) = \mathcal{H}\left[\gamma_l(\tau)\right] = \frac{1}{\pi}\sum_{i=1}^{N-1}\check{\gamma}_{l_i}(t) \quad \tau_{i-1} < t \leq \tau_i$$

where $\triangle_i = \tau_i - \tau_{i-1}$ and $\check{\gamma}_{l_i}(t)$ is the Hilbert transform of the i-th polynomial:

$$\begin{aligned}\check{\gamma}_{l_i}(t) = &\left(a_{l_i}t^3 + b_{l_i}t^2 + c_{l_i}t + d_{l_i}\right)\log\left(\frac{t}{t-\triangle_i}\right)\\ &+ a_{l_i}\left(\frac{\triangle_i^2 t}{2} - \triangle_i t^2 - \frac{\triangle_i^3}{3}\right) + b_{l_i}\left(-\triangle_i t - \frac{\triangle_i^2}{2}\right) - c_{l_i}\triangle_i.\end{aligned}$$

Such a representation for the IMF $\gamma_l(t)$ produces a smooth, differentiable, continuous function, it is approximated by the class of polynomial basis in the L^2 space.

3 EMD Stochastic Embedding Set-up

We have shown in Section 2 that working with cubic splines for the representation of the EMD method is advantageous from many perspectives. Firstly it is suitable to represent the interpolated signal $\widetilde{s}(t)$ from the observed time series $\{s(t_i)\}_{i=1}^N$ in an optimal fashion based on minimising mean squared error. Secondly, it allows one to perform the sifting procedure readily when representing the envelope functions and results in a collection of IMF basis functions $\{\gamma_l\}_{l=1}^L$ representations that are also cubic splines. Thirdly, the analytic extension via the Huang Hilbert transform, used to obtain the instantaneous frequency, admits closed form solutions for the representations of the IFs $\{\omega_l\}_{l=1}^L$ which is also characterised readily by cubic splines. Lastly, and most importantly, when considering moving from the path-wise EMD method basis extraction for one of the time series realised trajectories to a stochastic process embedding representation, the representation of IMFs via cubic splines allows one to utilise the established connection between Gaussian processes and B-splines to motivate working with Gaussian process stochastic embeddings.

3.1 EMD Stochastic Embedding Objectives

In developing the stochastic embedding of the EMD, we will distinguish between the deterministic (realised) or empirical EMD decomposition for a given signal trajectory, satisfying at any time $t \in [0,T]$ the property of EMD decomposition

$$s(t) = \sum_{l=1}^{L}\gamma_l(t) + r(t)$$

for IMF $\gamma_l(t)$ satisfying the mathematical characterisation given in Definition 2.2; and the stochastic process embedding of the EMD representation, denoted at any time $t \in [0,T]$, by the random variables (upper case for random variables)

$$S(t) \stackrel{d}{=} \sum_{l=1}^{L}\Gamma_l(t) + R(t)$$

The challenge with developing a stochastic embedding for EMD method is that it will be required to satisfy a few core features:

1. Sample paths of the embedded EMD stochastic process should be able to be consistent with the basis functions for the IMFs obtained from the empirical sample based characteristics that represent the classical EMD method as set-up in Definition 2.2.;

2. Since the EMD method satisfies for each realised sample time-series trajectory $\widetilde{s}(t)$ that

$$\widetilde{s}(t) = \sum_{l=1}^{L} \gamma_l(t) + r(t)$$

then one would naturally require such a property to be inherited at the population stochastic process level such that:

$$\widetilde{S}(t) \stackrel{d}{=} \sum_{l=1}^{L+1} \Gamma_l(t)$$

where we have denoted the stochastic process for $R(t)$ by $\Gamma_L(t)$ to reduce notational burden.

Ideally the representations of processes $\widetilde{S}(t)$ and IMF stochastic processes $\{\Gamma_l(t)\}_{l=1}^{L}$ would satisfy:

(a) Stochastic processes used to model $\widetilde{S}(t)$ and IMF processes $\{\Gamma_l(t)\}_{l=1}^{L+1}$ have known finite dimensional distributions and are from family of known stochastic process models which are easily parameterised and characterised. We will denote this family of models for distributions at time t

(b) Stochastic processes used to model $\widetilde{S}(t)$ and IMF processes $\{\Gamma_l(t)\}_{l=1}^{L+1}$ would also ideally be easily calibrated to realised EMD sample based decompositions via standard estimation methods like maximum likelihood estimation with closed form expressions for the likelihood of the model for the stochastic embedding.

(c) IMF stochastic processes $\{\Gamma_l(t)\}_{l=1}^{L}$ are of the same family of stochastic process model as that which represents the signal stochastic process $\widetilde{S}(t)$. In other words if, for each time t, one has that random variable $\widetilde{S}(t) \sim F \in \mathcal{F}$ is distributed by F in a family of distribution models $\mathcal{F}$ where

$$\widetilde{S}(t) \sim F\left(a; \Psi_{\widetilde{S}}\right) := \int_{-\infty}^{a} \cdots \int_{-\infty}^{a} f_{\Gamma_1,\ldots,\Gamma_{L+1}}\left(\gamma_1, \ldots, \gamma_{L+1}\right) d\gamma_1 \ldots d\gamma_{L+1}$$

with $\Psi_{\widetilde{S}}$ denoting the parameters of the model that indexes the family member from $\mathcal{F}$ and furthermore, where $f_{\Gamma_1,\ldots,\Gamma_{L+1}}$ is the joint distribution of the IMF random variables and tendency at time t, then it also holds that for each $t \in [0, T]$ and $l \in \{1, \ldots, L+1\}$ the distribution of the IMF random variables satisfies that it is also a member of this family of distribution models such that

$$\Gamma_l(t) \sim F\left(s, \Psi_{\Gamma_l}\right) \in \mathcal{F},$$

indexed by parameter vectors Ψ_l.

(d) Another desirable property for the stochastic embedding representation of EMD would be to have the conditional distributions also members of the same family of distributions of $\widetilde{S}(t)$, such that for each $t \in [0, T]$ and any combination of $J \leq L+1$ indexes denoted by subset $\mathcal{K} \subseteq \{1, \ldots, L+1\}$ one has that the random variable

$$\sum_{i \in \mathcal{K}} \Gamma_i(t) | \Gamma_{1,\ldots,L \setminus \mathcal{K}} \sim F\left(s; \Psi_{\mathcal{K}}\right) \sim \mathcal{F}$$

Note: In the case one assumes an independence model approximation for the joint distribution of the IMF random variables and tendency at each time $t \in [0, T]$ such that

$$f_{\Gamma_1,\ldots,\Gamma_L,R}\left(\gamma_1, \ldots, \gamma_L, r\right) = \prod_{l=1}^{L} f_{\Gamma_l}\left(\gamma_l\right) f_R\left(r\right)$$

Then the EMD method decomposition implies that the stochastic representation of the IMFs are closed under convolution. This means that at each time t the random variable for the signal $S(t) \sim F\left(s; \Psi_S\right)$ and the random variables for the IMFs $\Gamma_i \sim F\left(s; \Psi_{\Gamma_i}\right)$ satisfy that

$$F\left(s; \Psi_S\right) = \circledast_{i=1}^{L} F\left(s; \Psi_{\Gamma_i}\right) \circledast F\left(s; \Psi_R\right)$$

such that $F\left(s; \Psi_S\right), F\left(s; \Psi_{\Gamma_1}\right), \ldots, F\left(s; \Psi_{\Gamma_L}\right), F\left(s; \Psi_R\right) \in \mathcal{F}$

4 Developing a Stochastic Embedding of EMD

In this section we develop two approaches for the stochastic embedding of the EMD method which will be consistent with the EMD empirical decomposition whilst also concurrently satisfying the properties set out for such a stochastic representation of EMD given in Section 3.1. To achieve this we will develop two different system models each of which will be based on versions of multi-kernel Gaussian Processes models with specially selected kernel structures. The reference baseline or benchmark model we will compare to these two novel system models for EMD stochastic representation will be a Gaussian process fit directly to the original signal $s(t)$.

Gaussian Processes (GPs) are a highly expressive family of stochastic models widely adopted in machine learning, see [42]. Formally, a Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution, which is entirely described by its mean and kernel covariance function as detailed in Definition 4.1. The positive definite covariance function often referred to as kernel determines the class of functions from which such processes sample paths take support.

definition 4.1 (Gaussian Process (GP))**.** Denote by $f(x) : \mathcal{X} \to \mathbb{R}$ a stochastic process, parametrised with state-space $\{x\} \in \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}^d$. The random function $f(x)$ is a Gaussian Process if all finite dimensional distributions are Gaussian, where for any $n \in \mathbb{N}$, the random vector $(f(x_1), f(x_2), \ldots, f(x_n))$ is jointly normally distributed. We can therefore interpret a GP formally defined by the following class of random functions:

$$f := \{f(\cdot) : \mathcal{X} \to \mathbb{R} : f(\cdot) \sim \mathcal{GP}\left(\mu(\cdot, \boldsymbol{\psi}_f), k(\cdot, \boldsymbol{\theta}_f)\right)\} \tag{5}$$

with $\mu(\cdot, \boldsymbol{\psi}_f) : \mathcal{X} \to \mathbb{R}$, $k(\cdot, \boldsymbol{\theta}_f) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$,

$$\begin{aligned} \mu(\cdot, \boldsymbol{\psi}_f) &= \mathbb{E}\left[f(\cdot)\right] \\ k(\cdot, \boldsymbol{\theta}_f) &= \mathbb{E}\left[\left(f(\cdot) - \mu(\cdot, \boldsymbol{\theta}_\mu)\right)\left(f(\cdot) - \mu(\cdot, \boldsymbol{\theta}_\mu)\right)\right] \end{aligned} \tag{6}$$

The properties of the functions, i.e. smoothness, periodicity, etc., are determined by the sufficient statistic given by the covariance kernel function.

Before introducing these GP models, we will motivate theoretically why the class of GP models is suitable for a stochastic embedding that will be shown to be both meaningful for regularised spline representations of IMFs as well as suitable to satisfy the properties outlined for such a stochastic embedding of EMD discussed in Section 3.1.

4.1 Spline Representations of an IMF and Reproducing Kernel Hilbert Spaces

In order to make explicit the connection between using spline models to represent the path-wise empirical EMD decomposition of $\widetilde{s}(t)$ and the stochastic embedding via a multi-kernel Gaussian process, we will recall briefly known connections between splines and Gaussian Processes (GPs). Splines may be viewed as limits of interpolations related to stationary Gaussian processes. Hence, we will explore further this connection as follows.

Consider seeking to recover the l-th unknown IMF function $\gamma_l(t)$ for $t \in [0, T]$ based on current sifting defluctuation step data $\tilde{s}_l(t) := \tilde{s}(t) - \sum_{i=1}^{l-1} \gamma_i(t)$ at time points $t_1, \ldots, t_N$ denoted as observations here generically by $y_i := \tilde{s}_l\left(t_i\right)$. That is one has data $\{t_i, y_i\} \in \mathcal{T} \times \mathbb{R}$ and we seek the function representation for the l-th IMF $\gamma_l(t) : \mathcal{T} \to \mathbb{R}$ that minimizes the objective given generically in Equation 7, for instance which may be the familiar penalised residual sum-of-squares,

$$Q(\gamma_l) = \sum_{i=1}^{N} L\left(y_i, \gamma_l\left(t_i\right)\right) + \lambda J\left(\gamma_l\right) \tag{7}$$

where L is a loss function, $\lambda \geq 0$ is regularisation strength and J is a functional imposing smoothness on the IMF representation γ_l. One can connect the regularised spline solution to GPs by considering Reproducing Kernel Hilbert Spaces (RKHS) to explore the unifying framework to motive the GP stochastic embedding model, see details in [57] and more recent works in [58, 43, 59].

A Hilbert space $\mathcal{H}$ is an inner-product space which is complete in the metric induced by its norm. For every Hilbert space of functions on a set $\mathcal{T}$, one may define for each $t \in \mathcal{T}$ the evaluation functional $f : t \mapsto f(t)$. If every evaluation functional in the Hilbert space is bounded, then one obtains a Reproducing Kernel Hilbert Space (RKHS). Note L^2 is not an RKHS since the Dirac-delta function is not in L^2. In an RKHS the Riesz representation theorem states that one may find, for each t a representer $k_t \in \mathcal{H}$ such that

$$f(t) = \langle f, k_t \rangle.$$

Then one can define a function known as the kernel $k : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ by $k(s,t) = k_s(t)$. This function will be unique to a given RKHS $\mathcal{H}$ and has the properties of symmetry, nonnegative definiteness and satisfies the reproducing property $\langle k(\cdot,s), k(\cdot,t)\rangle = k(s,t)$.

To understand why the RKHS space and reproducing kernel K are introduced, consider the space of all finite linear combinations of functions $\{k(\cdot,s)|s \in \mathcal{T}\}$ with the inner product given by $\langle k_s, k_t\rangle = k(s,t)$ along with linearity. It is then the case that k is a kernel for this space with the property, according to the Representer Theorem, that solutions to the regularised empirical risk given in Equation 7 take the form

$$f(\cdot) = \sum_{i=1}^{N} \alpha_i k(\cdot, t_i)$$

for $\alpha_i \in \mathbb{R}$ for all $i \in \{1,\ldots,N\}$. The conditions under which such a representer theorem exists are studied in [60].

Given these results one may then link the estimation problem for representing each IMF to the case of polynomial smoothing splines, used to represent the IMF basis functions under the EMD method proposed. To see this consider, without loss of generality $\mathcal{T} = [0,1]$, penalty function $J(\gamma_l) = \int_0^1 \left(\gamma_l^{(m)}(t)\right)^2 dt$ which acts to penalise irregularity and induce smoothness in the spline representation of IMF basis. One can then construct an RKHS whose norm corresponds to this smoothing penalty J. Hence, the kernel needs to be made explicit.

Using Taylor's theorem in one dimension with integral remainder term to express the IMF function γ_l, which is assumed to have at least $m-1$ order absolutely continuous derivative in $[0,1]$ and $\gamma_l^{(m)} \in L^2[0,1]$, then

$$\gamma_l(t) = \sum_{i=1}^{m-1} \frac{t^i}{i!}\gamma_l^{(i)}(0) + \int_0^1 \frac{(t-s)_+^{m-1}}{(m-1)!}\gamma_l^{(m)}(s)ds,$$

where $(\cdot)_+$ is the positive part only and zero otherwise. If functions with this series representation with the first $m-1$ derivatives being 0 at $t=0$ are denoted by $\mathcal{W}_m^0$, then for $\gamma_l \in \mathcal{W}_m^0$ one has

$$\gamma_l(t) = \int_0^1 G_m(t,s)\gamma_l^{(m)}(s)ds$$

where $G_m(t,s) := (t-s)_+^m/(m-1)!$. Now observe that one can obtain an RKHS space from $\mathcal{W}_m^0$ with the inner product

$$\langle f, g\rangle = \int_0^1 f^{(m)}(s)g^{(m)}(s)ds$$

and kernel $k_1(t,s) = \int_0^1 G_m(t,r)G_m(s,r)dr$. Now if one defines the null space of the penalty function as $\mathcal{H}_0 = span\left(\{\varphi_i(t)\}_{i=1}^m\right)$ with $\varphi_i(t) = t^{i-1}/(i-1)!$. Then the kernel for $\mathcal{H}_0$ is $k_0(t,s) = \sum_{i=1}^m \varphi_i(s)\varphi_i(t)$. As shown in [57] the space $\mathcal{W}_m$ of functions with $m-1$ absolutely continuous derivatives and m derivatives can be written as a direct sum $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{W}_m^0$ with kernel $k = k_1 + k_0$. Furthermore, $J(\gamma_l)$ will be the square norm of the projection $P\gamma_l$ of γ_l onto $\mathcal{W}_M^0$ so the PRSS estimation objective in Equation 7 with $J(\gamma_l) = \int_0^1 \left(\gamma_l^{(m)}(s)\right)^2 ds$ becomes

$$Q(\gamma_l) = \sum_{i=1}^{N} L\left(y_i, \gamma_l(t_i)\right) + \lambda\|P\gamma_l\|^2 \tag{8}$$

for $\gamma_l \in \mathcal{H}$. By Representer Theorem, the solution is the generalised form given by

$$\gamma_l^{\lambda}(s) = \sum_{i=1}^{N} \alpha_i k_1(s,t_i) + \sum_{j=1}^{m} \beta_j \varphi_j(s)$$

is comprised of two parts: an unpenalized component of $\mathcal{H}_0$ and a linear combination of the projections onto $\mathcal{W}_m^0$ of the representers of evaluation at the N time points $t_1,\ldots,t_N$. For the squared error loss $L\left(y_i,\gamma_l(t_i)\right) = L\left(y_i - \gamma_l(t_i)\right)^2$ the solution corresponds to the natural polynomial spline, see discussion in [61] and [62].

Hence, we have been able to motivate the spline representation of the IMF as the solution to a generalised estimation problem in an RKHS regularised function space. Now we will endeavour to connection this through the RKHS theory to the Gaussian process embedding.

4.2 Relating Spline Representations of an IMF and a Gaussian Processes Stochastic Embedding

Now we will treat $\Gamma_l(t)$ as a random function modelled by a GP and we will illustrate the mathematical connection between the spline representation on the pathwise EMD method decomposition of an IMF and the stochastic embedding developed in this work via GP models.

For Gaussian process prediction with likelihoods that involve the observed values of the IMF γ_l at N training points, extracted by the EMD method sifting algorithm, the empirical loss $L\left(y_i, \gamma_l\left(t_i\right)\right)$ can be expressed according to the negative log-likelihood. Then the analog of the representer theorem, as detailed in [63] is given as follows.

Since the predictive distribution of $\Gamma_l(t_*)$ at test point t_* given observations $y_1, \ldots, y_N$ is given by

$$p\left(\gamma_l(t_*)|y_1, \ldots, y_N\right) = \int p\left(\gamma_l(t_*)|\gamma_l(t_1), \ldots, \gamma_l(t_N)\right) p\left(\gamma_l(t_1), \ldots, \gamma_l(t_N)|y_1, \ldots, y_N\right) d\gamma_l(t_1) \ldots d\gamma_l(t_N)$$

which in the GP case is expressed in terms of the GP covariance kernel k by

$$\begin{aligned} \mathbb{E}\left[\gamma_l(t_*)|y_1, \ldots, y_N\right] &= \left[k(t_*, t_1), \ldots, k(t_*, t_N)\right]^T K^{-1} \mathbb{E}\left[\gamma_l(t_1), \ldots, \gamma_l(t_N)|y_1, \ldots, y_N\right] \\ &= \sum_{i=1}^{N} \alpha_i k(t_*, t_i) \end{aligned} \tag{9}$$

with $[\alpha_1, \ldots, \alpha_N] = K^{-1}\mathbb{E}\left[\gamma_l(t_1), \ldots, \gamma_l(t_N)|y_1, \ldots, y_N\right]$ where K is the $N \times N$ Kernel matrix (Gram matrix).

One then obtains the regularized solution to Equation 7 from a GP perspective by noting that for the specific choice of loss and penalty given by

$$Q\left(\gamma_l\right) = \frac{1}{\sigma_N^2} \sum_{i=1}^{N} \left(y_i - \gamma_l(t_i)\right)^2 + \frac{1}{2}\|\gamma_l\|_{\mathcal{H}}^2$$

where the loss function is set to the negative log-likelihood in which σ_N^2 is the Gaussian noise model variance. The solution for the estimated IMF using this regularized estimation produces $\widehat{\gamma}_l = \operatorname{argmin}_{\gamma_l} Q\left(\gamma_l\right)$ which if one substitutes $\gamma_l(t) = \sum_{i=1}^{N} \alpha_i k(t_*, t_i)$ and uses the fact of RKHS space $\langle k(\cdot, t_i), k(\cdot, t_j)_{\mathcal{H}} = k(t_i, t_j)$ can be re-expressed by an estimation objective explicitly in terms of the GP model as follows:

$$\begin{aligned} Q(\boldsymbol{\alpha}) &= \frac{1}{2}\boldsymbol{\alpha}^T K \boldsymbol{\alpha} + \frac{1}{2\sigma_N^2}|\boldsymbol{y} - K\boldsymbol{\alpha}|^2 \\ &= \frac{1}{2}\boldsymbol{\alpha}^T \left(K + \frac{1}{2\sigma_N^2} K^2\right) \boldsymbol{\alpha} - \frac{1}{\sigma_N^2} \boldsymbol{y}^T K \boldsymbol{y} + \frac{1}{2\sigma_N^2} \boldsymbol{y}^T \boldsymbol{y}. \end{aligned}$$

Rewriting the objective in this manner expresses it as a parameter optimization problem in terms of coefficient vector $\boldsymbol{\alpha}$, this is the advantage of knowing that a Representer Theorem can be applied. If one then minimizes Q w.r.t. vector of coefficients $\boldsymbol{\alpha}$ one obtains

$$\widehat{\boldsymbol{\alpha}} = \left(K + \sigma_N^2 \mathbb{I}_N\right)^{-1} \boldsymbol{y}$$

which gives the prediction at test point t_*

$$\gamma_l(t_*) = \left[k(t_*, t_1), \ldots, k(t_*, t_N)\right]^T \left(K + \sigma_N^2 \mathbb{I}_N\right)^{-1} \boldsymbol{y}$$

which is exactly the predictive mean given in Equation 9.

Now to explicitly recover the solution to the smooth spline interpolation for the IMF representation obtained via solving Equation 8 using $m = 2$ and the regularised GP solution just presented we can use the result of [64] which shows that in this case if one considers a random function representation of the IMF given by

$$\gamma_l(t) = \sum_{j=0}^{1} \beta_j t^j + f(t)$$

where $\boldsymbol{\beta} \sim N\left(\mathbf{0}, \sigma_\beta^2 \mathbb{I}\right)$ and $f(\cdot)$ a GP with covariance $\sigma_f^2 k_{sp}(t, t')$ given by

$$k_{sp}(t, t') = \int_0^1 (t - s)_+ (t' - s)_+ ds = \frac{|t - t'| \min(t, t')^2}{2} + \frac{\min(t, t')^3}{3}.$$

Then to complete the example of the regularizer in the cubic spline case, we must remove penalties on polynomial terms in the null space by making taking $\sigma_\beta \to \infty$. This produces the final predictive mean solution for the GP representation of the cubic spline characterisation of the IMF given by

$$\overline{\gamma}_l(t_*) = [k(t_*, t_1), \ldots, k(t_*, t_2)]^T K_y^{-1} \left(\boldsymbol{y} - H^T \overline{\boldsymbol{\beta}}\right) + [(1, t_*)]^T \overline{\boldsymbol{\beta}}$$

with Kernel covariance matrix K_y corresponding to elements $\sigma_f^2 k_{sp}(t_i, t_j) + \sigma_N^2 \delta_{ij}$ evaluated at all training points, H the matrix collecting the vector of polynomial basis terms $(1, t)$ at training points and kernel least squares coefficient estimator given by

$$\overline{\boldsymbol{\beta}} = \left(H K_y^- 1 H^T\right)^{-1} H K_y^{-1} \boldsymbol{y}.$$

From this solution, one can see that the resulting solution for the predictive mean function for the GP representation of the IMF for γ_l will have a cubic polynomial form.

4.3 Gaussian Processes Based Stochastic EMD Embeddings

Having established how the GP representations is connected mathematically to the empirical path-wise cubic spline representation for an IMF in the EMD method, we now generalise the stochastic embedding from a single IMF to the entire collection of IMFs under two different system models proposed. Each of these will be designed to satisfy the properties proposed for the stochastic embedding objectives set out in Section 3.1.

To achieve the desired embedding, consider first the stochastic process associated with the observed sampled signal converted from samples $\{s(t_1), \ldots, s(t_N)\}$ to spline $\tilde{s}(t)$ which when considered as the realisation of stochastic process will be denoted by $S(t)$ and $\widetilde{S}(t)$ respectively. The reference model used for comparison to the stochastic EMD models will involve directly modelling the process $\widetilde{S}(t)$ without the EMD method signal decomposition information, via a GP model given in System Model 1 (SM1).

4.3.1 System Model 1 (SM1): Gaussian Process for $\widetilde{S}(t)$

For SM1 there is a choice to calibrate the GP model directly to observations of the process $S(t)$ or to set up the model alternatively as follows, using the values of $\widetilde{s}(t)$ for estimation of the GP model. This second choice will often be both more aligned as a reference model to the EMD method stochastic embedding as well as more robust to noise due to the regularisation that can be adopted when obtaining $\widetilde{s}(t)$. Therefore, under SM1 the GP model for signal $S(t)$ is obtained via

$$S(t) \stackrel{d}{=} \widetilde{S}(t) + \epsilon(t)$$

where we treat $\widetilde{S}(t)$ as a GP

$$\tilde{S}(t) \sim \mathcal{GP}\left(\mu(t; \boldsymbol{\psi}_{\widetilde{S}}); k(t, t'; \boldsymbol{\theta}_{\widetilde{S}})\right), \tag{10}$$

with $\mu(t; \boldsymbol{\psi}_{\widetilde{S}})$ and $k(t, t'; \boldsymbol{\theta}_{\widetilde{S}})$ representing the mean and kernel functions respectively, $\boldsymbol{\psi}_{\widetilde{S}}$ and $\boldsymbol{\theta}_{\widetilde{S}}$ are the sets of hyperparameters of the mean and the kernel respectively. The additive error $\epsilon(t)$ corresponds to a regression error based on using the spline representation $\widetilde{s}(t)$ for the representation and potentially calibration of the SM1.

4.3.2 System Model 2 (SM2): Gaussian Processses for IMFs $\{\Gamma_l(t)\}_{l=1}^L$

When the EMD is applied to signal $\widetilde{s}(t)$ and the set of basis functions are extracted, each IMF $\gamma_l(t)$ will be considered as the realised path of the stochastic process denoted as $\Gamma_l(t)$ and the one for the residual $r(t)$ denoted as $R(t)$. This will produce the following stochastic embedding of the EMD given:

$$\tilde{s}(t) \begin{cases} \nearrow \gamma_1(t) & \Gamma_1(t) \sim \mathcal{GP}(\mu_1(t; \boldsymbol{\theta}_{\mu_1}), k_1(t, t'; \boldsymbol{\theta}_1)) \\ \vdots & \\ \searrow \gamma_L(t) & \Gamma_L(t) \sim \mathcal{GP}(\mu_L(t; \boldsymbol{\theta}_{\mu_L}), k_L(t, t'; \boldsymbol{\theta}_L)) \end{cases}$$

The SM2 representation of the original stochastic process for $S(t)$ is then given by the GP:

$$S(t) \stackrel{d}{=} \widetilde{S}(t) + \epsilon(t)$$

with

$$\widetilde{S}(t) \stackrel{d}{=} \sum_{l=1}^{L} \Gamma_l(t) + R(t)$$

where $\epsilon(t) \sim N(0, \sigma_\epsilon)$ and $\Gamma_l(t)$ represents the GP for IMF l and there are $l = 1, \ldots, L$ of them and $R(t)$ represents the GP on the residual tendency component. This general structure will form the basic structure for the two stochastic embeddings proposed for the EMD method and we will refer to these two models as System Model 2 (SM2) and System Model 3 (SM3). Therefore one can see that the resulting model is still a GP model but differs from the baseline benchmark model in Eq. (10) as follows

$$\tilde{S}(t) \sim \mathcal{GP}\left(\sum_{l=1}^{L} \mu(t; \boldsymbol{\psi}_{\Gamma_l}) + \mu(t; \boldsymbol{\psi}_R); \sum_{l=1}^{L} k(t, t'; \boldsymbol{\theta}_{\Gamma_l}) + k(t, t'; \boldsymbol{\theta}_R) + \sigma_\epsilon \delta_{t,t'}\right) \tag{11}$$

It is apparent that the proposed GP model for the stochastic embedding of the EMD method differs from a direct GP model on the signal as detailed in reference model directly in how the sufficient statics are designed. The key point of the stochastic embedding of the EMD method GP framework is that the kernel of the GP is now comprised of a multi-kernel framework, where each kernel can be specifically calibrated to the extracted EMD's basis functions. Furthermore, it is trivially to verify that this stochastic embedding of the EMD method satisfies the objectives set-out in Section 3.1.

4.4 Treatment of the Residual Tendency Stochastic Embedding

As detailed in Section 3 last component extracted by the EMD corresponds to the residual or tendency component $r(t)$. By definition, this last component has only one convexity within the domain $[0, T]$. Therefore, it is possible, without loss of generality, to partition it in two subregions $[0, s]$ and $[s, T]$ in which monotonicity applies locally in each. Consequently one could then impose the following structure on the GP model for $R(t)$ over each region that enforces a stochastic monotonicity as discussed in [65], producing an isotonic restriction on the Gaussian Process. This is achieved by imposing derivative constraints on the sufficient statistics. Effectively, this utilises the fact that a derivative of a Gaussian process is a Gaussian process ([63]) and therefore a convexity constraint will result in conditions on the mean as outlined below:

$$\mathbb{E}\left[\frac{\partial \mu(t; \boldsymbol{\theta}_R)}{\partial t}\right] = \begin{cases} \frac{\partial \mathbb{E}[R(t)]}{\partial t} > 0, & \forall t \in [0, s] \\ \frac{\partial \mathbb{E}[R(t)]}{\partial t} < 0, & \forall t \in (s, T]. \end{cases}$$

One can then consider to impose these conditions at all out-of-sample points $R(t_*)$ in such a manner that on average one preserves monotonicity. Given the conditional distribution for $R(t_*)|R(t_1), \ldots, R(t_N)$ one imposes the following conditions on the predictive distribution:

$$\begin{aligned} \mathbb{E}\left[\frac{\partial R(t_*)}{\partial t}|R(t_1), \ldots, R(t_N)\right] &= \frac{\partial k(t_*, \boldsymbol{t})}{\partial t_*}\left(K + \sigma_\epsilon^2 \mathbb{I}\right)^{-1}\left[R(t_1), \ldots, R(t_N)\right]^T > 0 \\ Var\left[\frac{\partial R(t_*)}{\partial t}|R(t_1), \ldots, R(t_N)\right] &= \frac{\partial^2 k(t_*, \boldsymbol{t})}{\partial t_* \partial t_*} - \frac{\partial k(t_*, \boldsymbol{t})}{\partial t_*}\left(K + \sigma_\epsilon^2 \mathbb{I}\right)^{-1} \frac{\partial k(\boldsymbol{t}, t_*)}{\partial t_*} > 0 \end{aligned}$$

where $\boldsymbol{t} = [t_1, \ldots, t_N]^T$ and

$$\mathbb{C}\text{ov}\left[\frac{\partial r(t)^{(i)}}{\partial t}, r(t)^{(i)}\right] = \frac{\partial}{\partial t}\mathbb{C}\text{ov}\left[r(t)^{(i)}, r(t)^{(i)}\right], \quad \mathbb{C}\text{ov}\left[\frac{\partial r(t)^{(i)}}{\partial t_i}, \frac{\partial r(t)^{(j)}}{\partial t_j}\right] = \frac{\partial}{\partial t_j}\mathbb{C}\text{ov}\left[r(t)^{(i)}, r(t)^{(j)}\right].$$

There exists a second option for the stochastic embedding of EMD to treat the tendency, which involves rewriting the model in a conditional form as follows:

$$\tilde{S}(t)|r(t) \sim \mathcal{GP}\left(\sum_{l=1}^{L} \mu(t; \boldsymbol{\theta}_{\Gamma_l}) + r(t); \sum_{l=1}^{L} k(t, t'; \boldsymbol{\theta}_{\Gamma_l}) + \sigma_\epsilon \delta_{t,t'}\right).$$

Under this formulation, the monotonicity of the tendency is obtained using the EMD methods pathwise extracted tendency function $r(t)$. This is equivalent to developing an empirical Bayes formulation of the stochastic EMD embedding, see discussion in [66].

4.5 Adaptive Band-Limited IMF Partitions

Consider the extracted instantaneous frequencies (IFs) $\omega_1(t), \omega_2(t), \ldots, \omega_L(t)$ which were constructed from the IMFs $\gamma_1(t), \ldots, \gamma_L(t)$ as described in Section 2.2. The EMD method extracts these functions in decreasing order according to the oscillation index of the IMFs, i.e. $osc[\omega_1(t)] > osc[\omega_2(t)] > \cdots > osc[\omega_L(t)]$, where $osc[\cdot]$ is an operator that counts the number of turning points ie. convexity changes of a signal. Notice, that in non-stationary settings, the number of oscillations will not correspond to particular stationarity in the frequency plane, and in fact the IMFs can have time-varying IFs that move around the frequency plane but remain ordered in general by their oscillation. Therefore, in order to use the EMD extracted IMFs for a stochastic embedding that is aligned with a traditional notion of bandwidth based analysis, we develop the concept of the Band Limited IMFs (BLIMFs). This allows for the development of a stochastic representation of an EMD signal decomposition that is guaranteed to be characteristic of a particular frequency band. This leads to the third system model (SM3) which is formulated based on the idea of aggregating the IMFs samples whose IFs lie within the same frequency band. Such newly formulated Quasi-IMFs are named band-limited IMFs and denoted as BLIMFs and are then modelled according to the same GP. To define the model, one needs first to introduce a partition rule which identifies different local frequency bandwidths.

In order to develop SM3 based on BILMFs we need to first present the formalism of what we refer to as an adaptive partition of the (time,frequency) plane based on the EMD extracted instantaneous frequencies (IFs) $\omega_1(t), \omega_2(t), \ldots, \omega_L(t)$. We will construct a partition based on the observed IF samples, denoted by $\{\mathbf{p}_{l,n}\}_{l=1,n=1}^{L,N}$ where $\mathbf{p}_{l,n} = (t_n, \omega_l(t_n)) \in \Pi := \mathcal{T} \times \mathcal{I}$ with time interval $\mathcal{T} = [t_0, t_N]$ and frequency interval $\mathcal{I} = [\omega_0, \omega_M] = [\min_{n,l} \omega_l(t_n), \max_{n,l} \omega_l(t_n)]$, where Π denotes the partition region. In developing the BLIMFs, a criteria and estimation objective will be established that will allow for the definition of an optimal partition, denoted by Π^*, for the collection of empirical samples $\{\mathbf{p}_{l,n}\}_{l=1,n=1}^{L,N}$. To define Π^* we will segregate Π into an $M \times D$ partition. The partition of M non-overlapping bandwidths, denoted $\{\mathcal{I}_m\}_{m=1}^{M}$, in the frequency domain satisfy

$$\mathcal{I} = \bigcup_{m=1}^{M} \mathcal{I}_m, \text{ s.t. } \bigcap_{m=1}^{M} \mathcal{I}_m = \emptyset \text{ and } |\mathcal{I}| = \sum_{m=1}^{M} |\mathcal{I}_m|.$$

Within each bandwidth $\mathcal{I}_m$ a time domain partition is sought, that can be unique to each bandwidth, corresponding to D total time partitions per bandwidth. This produces a set of time partitions for the m-th bandwidth given by

$$\mathcal{T} = \bigcup_{d=1}^{D} \mathcal{T}_{m,d}, \text{ s.t. } \bigcap_{d=1}^{D} \mathcal{T}_{m,d} = \emptyset \text{ and } |\mathcal{T}| = \sum_{d=1}^{D} |\mathcal{T}_{m,d}|.$$

As noted, it is not necessary that $|\mathcal{T}_{m,d}| = |\mathcal{T}_{m',d}|$ for $m \neq m'$ and $m, m' \in \{1, \ldots, M\}$. From this formulation of time partitioned bandwidths we can arrive at a partition of Π by defining MD rectangles, each denoted by $\Pi_{m,d} = \mathcal{I}_m \times \mathcal{T}_{m,d}$ for $m = 1, \ldots, M$ and $d = 1, \ldots, D$ which are non-overlapping and satisfy

$$\Pi = \bigcup_{m,d}^{D} \Pi_{m,d}, \text{ s.t. } \bigcap_{m,d} \Pi_{m,d} = \emptyset \text{ and } |\Pi| = \sum_{m,d} |\Pi_{m,d}|.$$

See a diagramatic example of such a partition in Figure 1. In this illustration the frequency domain is partitioned into three intervals and the time domain into four intervals.

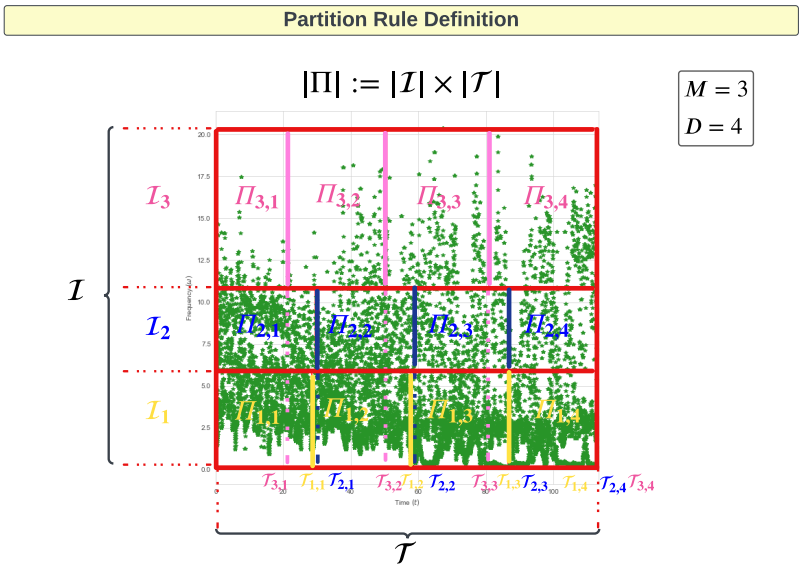


Figure 1: Partition Rule Definition showing how the empirical IFs samples $\{\mathbf{p}_{l,n}\}_{l=1,n=1}^{L,N}$ (colored in green) within region Π are partitioned into 12 time-frequency sub-regions that are defined by running the CEM method deriving Π^*. Note that, for this figure, we used only the first three IMFs, hence the first three IFs. This means that $L = 3$ in the Figure. The three IFs corresponds to the first three IFs of a speech segment used within the application of interest. Therefore, as it will be later in the paper highlighted, we consider speech segments with length $N = 5000$ samples.

4.5.1 System Model 3 (SM3): Gaussian Processses for BLIMFs $\left\{\Gamma_m^{(BL)}(t)\right\}_{m=1}^{M}$

Given a partition Π^* with M bandwidth we can develop the BLIMFs as follows

$$\begin{cases} \gamma_1^{(\mathrm{BL})}(t) = \gamma_1(t)\mathbb{1}_{\{\omega_1(t)\in\bigcup_{d=1}^{D}\Pi_{1,d}^*\}} + \ldots + \gamma_K(t)\mathbb{1}_{\{\omega_L(t)\in\bigcup_{d=1}^{D}\Pi_{1,d}^*\}} \\ \gamma_2^{(\mathrm{BL})}(t) = \gamma_1(t)\mathbb{1}_{\{\omega_1(t)\in\bigcup_{d=1}^{D}\Pi_{2,d}^*\}} + \ldots + \gamma_K(t)\mathbb{1}_{\{\omega_L(t)\in\bigcup_{d=1}^{D}\Pi_{2,d}^*\}} \\ \vdots \\ \gamma_M^{(\mathrm{BL})}(t) = \gamma_1(t)\mathbb{1}_{\{\omega_1(t)\in\bigcup_{d=1}^{D}\Pi_{M,d}^*\}} + \ldots + \gamma_K(t)\mathbb{1}_{\{\omega_L(t)\in\bigcup_{d=1}^{D}\Pi_{M,d}^*\}} \end{cases} \tag{12}$$

these extracted BLIMFs in turn lead to the band-limited stochastic embedding of EMD method that we denoted as System Model 3 (SM3) given as follows

$$\begin{array}{ccccccc} & \gamma_1(t) & \to & \omega_1(t) & & \Gamma_1^{(\mathrm{BL})}(t)|\Pi=\Pi^* & \sim \mathcal{GP}(\mu_1^{\mathrm{BL}}(t;\boldsymbol{\theta}_{\mu_1^{\mathrm{BL}}}), k_1^{\mathrm{BL}}(t,t';\boldsymbol{\theta}_{k_1^{\mathrm{BL}}})) \\ \tilde{s}(t) & \ldots & & \longrightarrow \quad \Pi^* \quad \longrightarrow & \ldots & & \ldots \\ & \gamma_L(t) & \to & \omega_L(t) & & \Gamma_M^{(\mathrm{BL})}(t)|\Pi=\Pi^* & \sim \mathcal{GP}(\mu_M^{\mathrm{BL}}(t;\boldsymbol{\theta}_{\mu_M^{BL}}), k_M^{\mathrm{BL}}(t,t';\boldsymbol{\theta}_{k_M^{\mathrm{BL}}})) \end{array}$$

where $\Gamma_l^{(BL)}(t)$ denote the stochastic GP embedding of the l-th BLIMF. We note that since the BLIMF construction satisfies that

$$\tilde{s}(t) = \sum_{m=1}^{M-1} \gamma_m^{(\mathrm{BL})}(t) = \sum_{i=1}^{L} \gamma_i(t)$$

one can see that there will be no loss of information. However, the advantage will be in bandwidth selectivity as well as producing a frequency band-limited multi-kernel GP formulation where under SM3 one represents the stochastic process $\tilde{S}(t)$ via multi-kernel representation given by

$$\widetilde{S}(t)|\Pi^* \overset{d}{=} \sum_{m=1}^{M} \Gamma_m^{(\mathrm{BL})}(t) \sim \mathcal{GP}\left(\mu_s(t;\boldsymbol{\theta}_{\mu_s}), k_s(t,t';\boldsymbol{\theta}_{k_s})\right),$$

where $\mu_s(t;\boldsymbol{\theta}_{\mu_s}) = \sum_{m=1}^{M} \mu_m^{BL}(t)$ and $k_s(t,t';\boldsymbol{\theta}_{k_s}) = \sum_{m=1}^{M} k_m^{BL}(t,t';\boldsymbol{\theta}_{k_M^{BL}})$.

To demonstrate such a construction, consider the illustration in Figure 2. The left panels show the first three IMFs $\gamma_1(t), \gamma_2(t), \gamma_3(t)$ extracted on a given speech signal. The x-axis represents the time (in seconds). Only three IMFs have been considered in this example since, for speech analysis in general, the first 3 IMFs capture the majority of the frequency content (corresponding to formant frequencies, i.e. the frequencies at which the vocal folds vibrate) required to describe, capture or classify voices in general (see [16]). The right panels present the first three BLIMFs, which are obtained according to the model given in Eq. (12). It is possible to observe how the time sample points have been reassigned within a new basis since its related frequency sample points fell into a different sub-region.

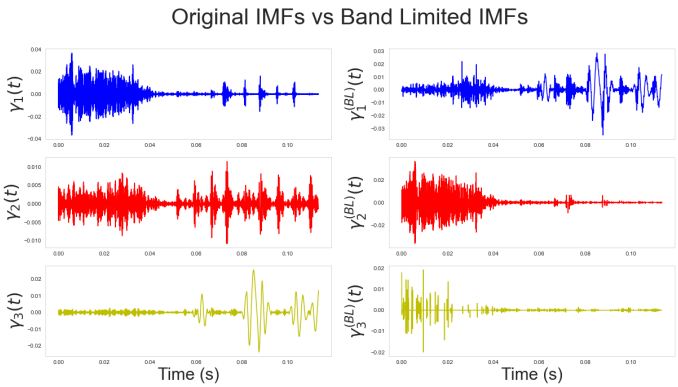


Figure 2: Comparison of the original extracted IMFs (left panels) and the obtained band-limited IMFs. (right panels). The original signal is a segment of the speech signals considered in section 7. The x-axis represents time and is given in seconds. It corresponds to 0.13 seconds, or, 130 milliseconds approximately (given that the speech segments is 5000 samples recorded at 44.kHz). The y-axis shows the amplitudes of the IMFs (left panels) and the band-limited IMFs (right panels).

5 Time Series Covariance Functions for Multi-kernel GP Stochastic EMD Embeddings

In this section we discuss how to develop a generative embedding kernel based on the Fisher kernel first proposed in [49] and [67]. This kernel family has the advantage that it can be developed to produce a time series kernel for a GP that will adapt to the local structure of the observed process being modelled, see discussions on GP time series kernels in [68]. It does this through a generative embedding mechanism that transfers the observed signal into a model space and then develops a subsequent sequence of feature vectors captured by the covariance operator that makes up the kernel. When the feature vectors represent summary statistics of a fitted model over the observed signal, such as the Fisher score, one produces the Fisher kernel embedding. We will use this Fisher kernel structure for SM1, SM2 (per IMF) and SM3 (per BLIMF). We begin this section by presenting the Fisher kernel basic details. We then subsequently discuss how we obtain the partition Π^* for SM3 definition of the optimal BLIMFs.

5.1 Generative Embedding Kernel

The idea of a generative embedding kernel is to map the original time series data into a model derived sequence of feature vectors that form an embedded time series representations. Think of, for instance, a time series of summary statistics. When the summary statistics are based on a model representation, this is known as a generative embedding as the model generates the feature time series upon which the GP kernel is designed from the original input time series data. In [49] a generative embedding approach was developed where the kernel used was termed a Fisher kernel. It was given this name as the final stage of the generative embedding map was determined by the gradient of the log-likelihood of the parameters of an underlying generative model, which subsequently defined a new feature space called the Fisher score space. It describes how that parameter contributes to the process of generating a particular input data. The gradient maintains all the structural assumptions that the model encodes about the generation process.

The Fisher kernel has been successfully employed within speech verification and recognition tasks by [69] and [70]. Its role in this work consists of detecting voice disturbances in displacement, direction, and velocity to differentiate between healthy and ill subjects. The adopted generative models used to produce the Fisher score feature space were intentionally kept simple and utilised basic time series models to represent the generative model embedding selected

to produce the speech signal IMF based feature vectors. The model for the generative embedding of the l-th IMF will be denoted by $g(\gamma_l(t);\boldsymbol{\theta}_k)$ with model parameters $\boldsymbol{\theta}_k$. Such generative models are not designed to be perfect representations of the original time series but rather to capture summary features of the IMF over time that, in turn could produce an adaptive Fisher kernel structure that could adapt locally to a time varying frequency characteristics of each IMF.

One defines the Fisher score at time t, denoted by $U_{\boldsymbol{\theta}_k}(t)$ as follows:

$$U_{\boldsymbol{\theta}_k}(t) = \nabla\boldsymbol{\theta}_k \ln g(\gamma_l(t);\boldsymbol{\theta}_k)$$

where $\nabla\boldsymbol{\theta}_k$ denotes the gradient operator with respect to $\boldsymbol{\theta}_k$ of the time t of the log-likelihood term $\ln g(\gamma_l(t);\boldsymbol{\theta}_k)$. In so doing, one constructs an embedding into a generative model feature space which allows one to subsequently define the Fisher kernel via the inner product in this space:

$$k(t,t') = U_{\boldsymbol{\theta}_k}(t)^\intercal\ \mathcal{I}^{-1}\ U_{\boldsymbol{\theta}_k}(t')$$

where $\mathcal{I}$ is the Fisher Information Matrix $\mathcal{I} := \mathbb{E}\left[U_{\boldsymbol{\theta}_k}(t)\ U_{\boldsymbol{\theta}_k}(t)^\intercal\right]$. Hence, the Fisher score is a feature mapping such that $U_{\boldsymbol{\theta}_k}(t)$ maps $\gamma_l(t)$ into a feature vector that is a point in the gradient space of the manifold $M_{\boldsymbol{\Theta}_k}$, see [49]. The gradient $U_{\boldsymbol{\theta}_k}(t)$ defines the direction δ which maximizes $\ln g(\gamma_l(t);\boldsymbol{\theta}_k)$ while traversing the minimum distance in the manifold given by $D(\boldsymbol{\theta}_k, \boldsymbol{\theta}_k+\delta)$, where $D(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|$. This latter gradient is usually known as natural gradient and is obtained from the ordinary gradient via $\phi_{\boldsymbol{\theta}_k}(t) = \mathcal{I}^{-1}U_{\boldsymbol{\theta}_k}(t)$. Hence, the mapping $\gamma_l(t) \to \phi_{\boldsymbol{\theta}_k}(t)$ is called the natural mapping and the natural kernel associated to it corresponds to the inner product between these feature vectors relative to the local Riemannian metric. Note that the information matrix is asymptotically immaterial and so often one works with the simplified kernel given by setting $\mathcal{I} = \mathbb{I}$.

5.2 Adaptive Gaussian Kernel Design through Optimal Time-Frequency EMD Partitions

In SM3, where the BLIMFs are used to define the inputs to the GP models, one has a choice to either select the desirable time-frequency partitions Π^* based on apriori information about the signal spectrum or frequency bands of interest over time. Alternatively, in many settings, such apriori beliefs about the partition may not be available and one instead seeks an optimal partition Π^* according to a desirable data-driven criterion. This section develops a solution to the optimal data-driven partition rule for SM3.

Many possible objectives could be considered. The one considered in this work is to determine the optimal partition for a given number of bandwidths that achieves empirical coverage of the sample IFs per time-frequency slot with most uniform coverage over Π. Such a partition is based on a discretised representation of the time-frequency plane that uses the IFs samples so that these can be allocated to frequency bandwidths whose distribution is as close as possible to uniform such that each band selected will have equivalent total spectral energy contributions from each BLIMF. This problem corresponds to a combinatorial search which becomes highly computational when it comes to standard optimisation techniques like simulated annealing, tabu search, MCMC algorithms. In this section an effective solution is proposed using the cross-entropy method (CEM) of [71] which has been shown to be highly effective in solving hard COPs.

A core component of CEM is that it exploits an Importance Sampling (IS) framework to approximate the optimal solution. In the main literature of CEM minimising the Kullback–Leibler (KL) divergence, the distributions are commonly referred to as the target (true) distribution treated as an ideal model for the data (in this case, a uniform distribution) and an empirical distribution (an approximation of the true distribution), in this case, based on the empirical distribution of the sample IFs obtained from a given partition rule. An overview of the process of constructing IMFs followed by IFs then an optimal partition rule Π^* via CEM followed by construction of the subsequent BLIMFs given the partition rule is provided in Figure 3.

5.2.1 Formulation of the Time-Frequency Partition Optimisation Problem

This subsection formalises the optimisation problem that estimates the optimal partition Π^*. A given partition of Π according to M frequency bands is structured according to an increasing sequence of parameters $\omega_1,\ldots,\omega_{M-1}$, defining frequency bandwidth subintervals of $\mathcal{I}$. In addition, for each bandwith there are D time partitions determined, for the m-th bandwidth, by an increasing sequences of parameters $s_{m,1},\ldots,s_{m,D-1}$, which defines the subintervals of $\mathcal{T}$. Hence, we denote the set of parameters to be estimated to determine the partition by vector:

$$\boldsymbol{\psi} = [\omega_1,\ldots,\omega_{M-1}, s_{1,1},\ldots,s_{1,D-1},\ldots,s_{m,1},\ldots,s_{m,D-1},\ldots,s_{M,1},\ldots,s_{M,D-1}] \in \Psi. \tag{13}$$

We will next introduce the CEM importance sampling structure. Consider $\mathcal{X} = \{(m,d)\}_{m=1,d=1}^{M,D}$, the set of DM tuples and a random variable $X : \mathcal{X} \to \mathbb{R}$ with a target uniform density $\pi(x)$ given on support $\mathcal{X}$ by:

$$\text{Target: } \pi(x) = \prod_{m,d} \pi_{m,d}^{\mathbf{1}_{\{x=(m,d)\}}} \ \text{ for } \ \pi_{m,d} = \mathbb{P}\big(X=(m,d)\big) = \frac{|\Pi_{m,d}|}{|\Pi|}.$$

such that the probability of drawing tuple (m,d) is proportional to the area of rectangle $\Pi_{m,d}$ versus Π. Given a current estimate of the partition Π^* one can also construct the empirical distribution from N time samples of the L set of IFs denoted by $\hat{\pi}(x)$ such that

$$\text{Empirical}: \ \hat{\pi}(x) = \prod_{m,d} \hat{\pi}_{m,d}^{\mathbf{1}_{\{x=(m,d)\}}} \ \text{ for } \ \hat{\pi}_{m,d} = \hat{\mathbb{P}}\big(X=(m,d)\big) = \frac{|\mathcal{P}_{m,d}|}{LN},$$

where $\mathcal{P}_{m,d} = \Big\{\omega_l(t_n) \in \Pi^*_{m,d} : l \in \{1,\ldots,L\}, n \in \{1,\ldots,N\}\Big\}$. Therefore, the probability of drawing tuple (m,d) reflects the proportion of the number of points $p_{l,n} = (t_n, \omega(t_n))$ that lay within the rectangle $\Pi^*_{m,d} \subset \Pi^*$ to the overall sample size. Furthermore, the distribution $\widehat{\pi}(x)$ is clearly then a function of the parameter vector Ψ, which has parameters that satisfy the conditions for each bandwith:

$$\Psi = \begin{cases} \omega_1,\ldots,\omega_{M-1} \in (\omega_0,\omega_M) \text{ such that } \omega_0 < \omega_1 < \ldots < \omega_{M-1} < \omega_M, \\ s_{1,1},\ldots,s_{1,N_1-1} \in (t_0,t_N) \text{ such that } t_0 < s_{1,1} < \ldots < s_{1,D-1} < t_N, \\ \vdots \\ s_{m,1},\ldots,s_{m,N_m-1} \in (t_0,t_N) \text{ such that } t_0 < s_{m,1} < \ldots < s_{m,D-1} < t_N, \\ \vdots \\ s_{M,1},\ldots,s_{M,N_M-1} \in (t_0,t_N) \text{ such that } t_0 < s_{M,1} < \ldots < s_{M,D-1} < t_N. \end{cases}$$

and characterise the partition Π^*. From these definitions, it is clear that under these definitions one has that $\pi_{m,d}$ and $\hat{\pi}_{m,d}$ are valid probabilities and satisfy

$$\sum_{m,d} \pi_{m,d} = 1 \text{ and } \sum_{m,d} \hat{\pi}_{m,d} = 1.$$

The optimization objective can then be formed under the CEM which in this problem formulation involves selecting the support of X in such a way that the Kullback-Leibler divergence,

$$KL(\hat{\pi},\pi) = \int_{x\in\mathcal{X}} \pi(x) \log\left(\frac{\pi(x)}{\hat{\pi}(x)}\right) dx,$$

measuring the similarity between the two proposed distributions target and empirical partitioned density, is minimised based on determining an optimal choice of the parameters that define the partition $\psi^\star$, given as follows:

$$\psi^* = \underset{\psi\in\Psi}{\operatorname{argmin}}\, KL(\hat{\pi},\pi;\psi) = \underset{\psi\in\Psi}{\operatorname{argmax}} -KL(\hat{\pi},\pi;\psi) \tag{14}$$

Since this is a discrete problem, this objective can be simplified as follows:

$$\begin{aligned} KL(\hat{\pi},\pi;\psi) &= \sum_{m=1}^{M}\sum_{d=1}^{d} \pi\left(x=(m,d)\right) \log\left(\frac{\pi\left(x=(m,d)\right)}{\hat{\pi}\left(x=(m,d)\right)}\right) \\ &= \log LN - \log|\Pi| + \frac{1}{|\Pi|}\sum_{m=1}^{M}\sum_{d=1}^{d}\Big\{|\Pi_{m,d}|\Big(\log|\Pi_{m,d}| - \log|\mathcal{P}_{m,d}|\Big)\Big\}. \end{aligned} \tag{15}$$

The derivation of this is provided in Appendix A.

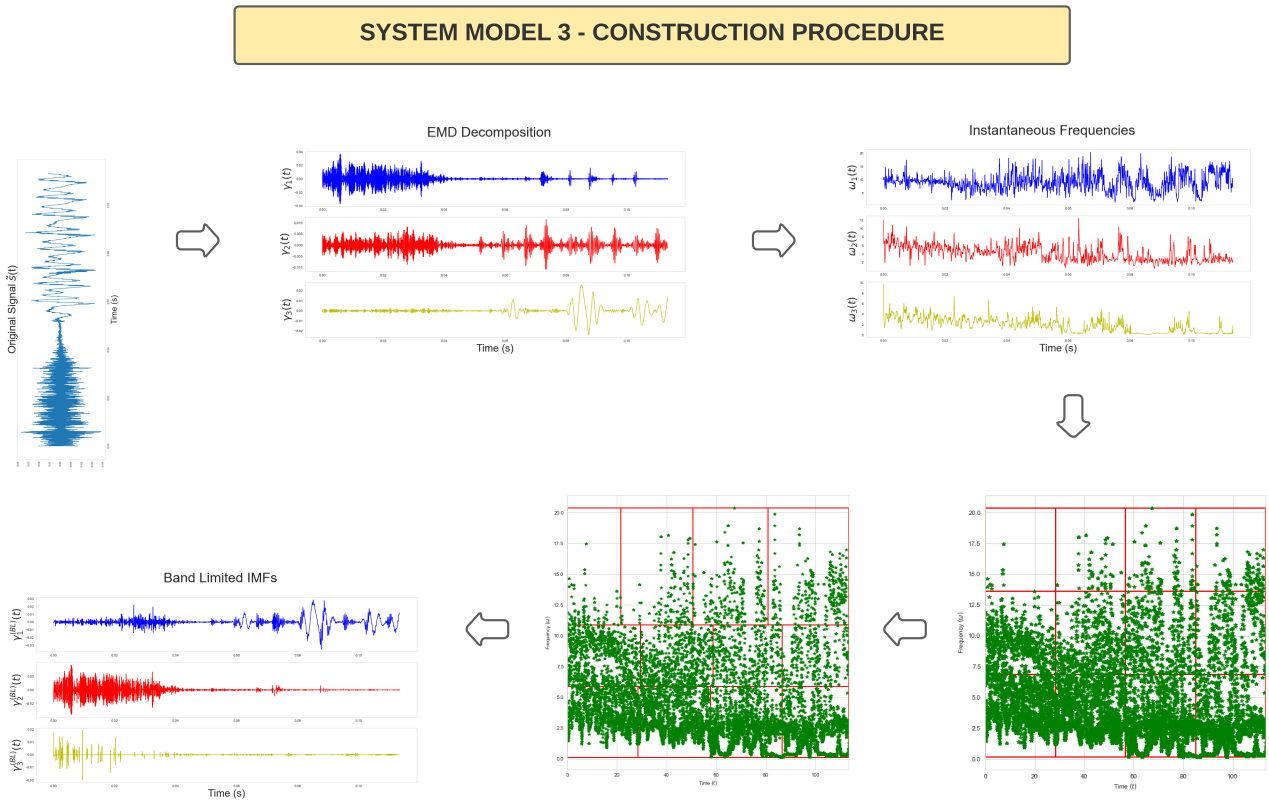


Figure 3: Figure presenting the steps required for the implementation of System Model 3. The first plot represents the original interpolated signal $\tilde{s}(t)$. This is a segment of speech signal used within the experiments section and corresponds to 0.13 seconds of speech. The x-axis corresponds to time (measures in seconds) and the y-axis to the amplitude. In the following plots, equivalent settings for the axes apply. Afterwards, the EMD is applied and the first three IMFs $\gamma_1(t)$, $\gamma_2(t)$, $\gamma_3(t)$ are plotted. The related IFs $\omega_1(t)$, $\omega_2(t)$, $\omega_3(t)$ are extracted and plotted. After, the empirical sample points of the IFs are passed to the CEM method. The fourth step of this procedure is the initial partition Π^0 used to initialise the cross-entropy algorithm, while the fifth step represents the CEM estimated optimal partition Π^*. Lastly, the reconstructed BLIMFs are provided.

5.2.2 Kernel Density Estimator Smoothing of Kullback-Leibler Divergence in Optimal Partitioning Problem

For a given current estimate of the partition Π^*, it can arise for a given emprical sample of the IFs that certain sub-rectangles $\Pi^*_{m,d}$ might not contain any of the sample points $\mathbf{p}_{l,n} = (t_n, \omega_l(t_n)) \in \Pi$. As a result, the corresponding set $\mathcal{P}_{m,d}$ will be empty, i.e. $\mathcal{P}_{m,d} = \emptyset$. Consequently, the probabilities $\hat{\pi}_{m,d}(x) = \frac{|\mathcal{P}_{m,d}|}{LN}$ equal zero and their logarithms used to calculate $KL(\hat{\pi}, \pi; \psi)$ in Eq. (15) tend to infinity. To avoid these numerical difficulties one can approximate $\hat{\pi}_{m,d}(x)$ by a kernel density estimator $\hat{\pi}^e_{m,d}(x; k, h)$ parametrised by kernel $k : \Pi \times \Pi \to \mathbb{R}$ and bandwidth $h > 0$ such that

$$\hat{\pi}^e_{m,d}(x; k, h) = \int_{\Pi_{m,d}} \hat{\pi}(\mathbf{p}; k; h) d\mathbf{p} = \int_{\omega_{m-1}}^{\omega_m} \int_{s_{m,d-1}}^{s_{m,d}} \hat{\pi}(\mathbf{p}; k; h) d\mathbf{p},$$

where $\hat{\pi}(\mathbf{p}; k; h) : \Pi \to [0, 1]$ is a kernel density estimator of points $\mathbf{p} = (t, \omega(t)) \in \Pi$ specified on a sample set $\mathbf{p}_{n,l}$

$$\hat{\pi}(\mathbf{p}; k; h) = \frac{1}{Nh} \prod_{n=1}^{N} \prod_{l=1}^{L} k\left(\frac{\mathbf{p} - \mathbf{p}_{n,l}}{h}\right) \quad \text{such that} \quad \int_{\Pi} \hat{\pi}(\mathbf{p}; k, h) d\mathbf{p} = 1.$$

By using the above, the objective function of the partitioning problem in (15) is reformulated to be the Kullback-Leibler divergence between $\pi(x)$ and

$$\hat{\pi}^e(x;k,h) = \prod_{m,d} \left(\hat{\pi}^e_{m,d}(x;k,h)\right)^{\mathbf{1}_{\{x=(m,d)\}}}, \tag{16}$$

given by

$$\begin{aligned} KL(\psi) := KL(\hat{\pi}^e, \pi; \psi) &= \int_{x\in\mathcal{X}} \pi(x) \log\left(\frac{\pi(x)}{\hat{\pi}^e\left(x;k,h\right)}\right) dx \\ &= \sum_{m=1}^{M}\sum_{d=1}^{d} \pi\left(x=(m,d)\right) \log\left(\frac{\pi\left(x=(m,d)\right)}{\hat{\pi}^e(x=(m,d);k,h)}\right) \\ &= -\log|\Pi| - \log C \\ &+ \frac{1}{|\Pi|}\sum_{m=1}^{M}\sum_{d=1}^{d} |\Pi_{m,d}|\Big(\log|\Pi_{m,d}| - \log\frac{\hat{\pi}^e(x=(m,d);k,h)}{C}\Big) \end{aligned} \tag{17}$$

with $C > 0$ and set to a very small number, ie $C = 10^{-100}$. The derivation of the above is provided in Appendix B.

5.3 Stochastic Optimisation of Optimal Time-Frequency Partition via Cross Entropy

Given the formulated objective function for the partition problem defined in (14) one can now define the CEM approach to stochastic optimisation used to solve for the optimal partition given the IFs. Recall, such an objective utilises the $KL(\cdot)$ divergence as a similarity measure between two distributions, empirical and target. This must be optimised with respect to the vector of parameters ψ. The CEM process to undertake this stochastic optimisation is developed by considering the level sets of the objective function $\{\psi : KL(\psi) \geq \zeta\}$ for $\zeta \in \mathbb{R}$, such that at the point that $\zeta = \widehat{KL} = \text{argmax}_{\psi\in\Psi} KL(\psi)$, we have $\{\psi : KL(\psi) \geq \zeta\} = \{\psi^\star\}$. We can formulate the importance sampling solution to achieving this outcome through a sequence of K intermediate solutions each based on a progressively less relaxed level set constraint i.e. $\zeta_1 < \zeta_2 < \cdots < \zeta_K$ where $\zeta_K \approx \text{argmax}_{\psi\in\Psi} KL(\psi)$ and at each iteration one updates the importance distribution to increase the chance of sampling solutions that are feasible according to the current level set constraint. Next we define the IS formulation of the CEM stochastic optimisation solution. This will involve defining an IS sampling distribution for the parameters $\boldsymbol{\psi}$ as given in Eq. (13) that make up the specification of the current estimate of the optimal partition Π^*. In order to achieve this we consider a family of probability measure $\{\mathbb{P}_{\varphi'} : \varphi' \in \Phi\}$ with support Ψ that admits a density $\{f_\varphi : \varphi \in \Phi\}$ also parametrised by $\varphi \in \Phi$. Let $\mathbb{E}_\varphi$ denote the expectation taken with respect to $\mathbb{P}_\varphi$. Let us fix φ and ζ and define a rare event probability problem:

$$\mathbb{P}_\varphi\left[KL(\psi) \geq \zeta\right] = \mathbb{E}_\varphi\left[\mathbb{I}_{\{KL(\psi)\leq\zeta\}}\right] = \int_\Psi \mathbb{I}_{\{KL(\psi)\leq\zeta\}} f_\varphi(\psi) d\psi$$

Instead of approximating this probability naively by sampling from f_φ, the importance sampling method is used. Let $g_{\varphi'}$ denote the importance sampler with $\varphi' \in \Phi$. Importance sampling approximates the rare event probability by

$$\begin{aligned} \mathbb{P}_\varphi\left[KL(\psi) \geq \zeta\right] &= \int_\Psi \mathbb{I}_{\{KL(\psi)\leq\zeta\}} f_\varphi(\psi) d\psi = \int_\Psi \mathbb{I}_{\{KL(\psi)\leq\zeta\}} \frac{f_\varphi(\psi)}{g_{\varphi'}(\psi)} g_{\varphi'}(\psi) d\psi \\ &= \mathbb{E}_{\varphi'}\left[\mathbb{I}_{\{KL(\psi)\leq\zeta\}} \frac{f_\varphi(\psi)}{g_{\varphi'}(\psi)}\right] \approx \frac{1}{S}\sum_{i=1}^{S}\left\{\mathbb{I}_{\{KL(\psi^i)\leq\zeta\}} \frac{f_\varphi(\psi^i)}{g_{\varphi'}(\psi^i)}\right\} \end{aligned}$$

where vectors ψ^i for $i = 1, \ldots, S$ are iid samples generated from IS density $g_{\varphi'}(\psi)$. The optimal importance sampler densities ($g_{\varphi'}$) parameters φ' are then obtained progressively in the CEM iterations for a given level set ζ by:

$$\begin{aligned} \varphi^\star &= \underset{\varphi'\in\Phi}{\text{argmax}} \int_\Psi \mathbb{I}_{\{KL(\psi)\leq\zeta\}} f_\varphi(\psi) \log\frac{f_\varphi(\psi)}{g_{\varphi'}(\psi)} d\psi \\ &\approx \underset{\varphi'\in\Phi}{\text{argmax}} \frac{1}{S}\sum_{i=1}^{S} \mathbb{I}_{\{KL(\psi^i)\leq\zeta\}} \log g_{\varphi'}(\psi^i) \end{aligned} \tag{18}$$

where vectors ψ^i for $i = 1, \ldots, S$ are iid samples generated from $f_{\varphi'(\psi)}$. Notice that the last line of 18 corresponds to the maximum likelihood estimation (MLE) of φ' when the samples are $\{\psi^i : KL(\psi^i) \geq \zeta\}$. The CEM starts from an initial sampling distribution $g_{\varphi_0^\star}$ and iteratively updates the threshold $\hat{\zeta}$ and the sampling distribution $g_{\varphi'}$. For a detailed introduction to cross-entropy, the reader should refer to [54].

5.4 Design of the Cross Entropy Importance Sampling Distribution

In this manuscript the optimisation problem is over a discrete support and so we have utilised a Multinomial distribution for the importance sampling distribution. In order to specify this distribution, consider a discretisation of the intervals $\mathcal{I}$ and $\mathcal{T}$. The importance sampling distribution must reflect the distribution of discrete random variables that partition the rectangle Π. Consider regular dense grids of $\mathcal{I}$ and $\mathcal{T}$ constructed by:

1. Partition of $\mathcal{I}$ into small N_ω intervals of size $\Delta_\omega = \frac{\omega_M - \omega_0}{N_\omega}$, and we define $\mathcal{I}_{n_\omega}^{grid} = \omega_0 + [n_\omega - 1, n_\omega]\Delta_\omega$ for $n_\omega = 1, \ldots, N_\omega$, therefore $|\mathcal{I}_a^{grid}| = \Delta_\omega$;
2. We partition $\mathcal{T}$ into small N_τ intervals of size $\Delta_\tau = \frac{t_N - t_0}{N_\tau}$, and we define $\mathcal{T}_{n_\tau}^{grid} = \omega_0 + [n_\tau - 1, n_\tau]\Delta_\tau$ for $n_\tau = 1, \ldots, N_\tau$, therefore, $|\mathcal{T}_\tau^{grid}| = \Delta_\tau$.

Now define the probabilistic model to partition $\mathcal{I}$ into M subintervals, $\mathcal{I}_m$ for $m = 1, \ldots, M$ according to an (M)-dimensional multinomial random vector $\mathbf{X}$ with entries X_m on the support of $\{0, \ldots, N_\omega\}$ which indicate how many subsequent grids $\mathcal{I}_{n_\omega}^{grid}$ are connected to construct partitions $\mathcal{I}_m$ and corresponding break points $\omega_{m-1}, \omega_m \in \mathcal{I}$. Therefore, the multinomial random vector $\mathbf{X}$ models the number of grid points out of N_ω that belong to each of M intervals with probabilities of being in an interval being $0 \leq p_1, \ldots, p_M \leq 1$ for $\sum_{m=1}^M p_m = 1$. The distribution function of $\mathbf{X}$ is formulated as

$$\pi(\mathbf{x}; \mathbf{p}) = \pi(x_1, \ldots, x_M; p_1, \ldots, p_M) = \frac{N_\omega!}{\prod_{m=1}^M x_m!} \prod_{m=1}^M p_m^{x_m}.$$

for $\mathbf{p} = [p_1, \ldots, p_M]$. Recall that $\sum_{m=1}^M X_m = N_\omega$ since $\mathbf{X}$ divides N_ω points into M subsets. For instance, for realisations of $X_1,\ X_2$ such that $x_1 = 2$ and $x_2 = 5$, the partitions $\mathcal{I}_1 = [\omega_0, \omega_1]$ and $\mathcal{I}_2 = [\omega_1, \omega_2]$ are given by

$$\omega_1 = \omega_0 + \Delta_\omega x_1 \text{ and } \omega_1 = \omega_1 + \Delta_\omega x_2 = \omega_0 + \Delta_\omega(x_1 + x_2)$$

This example gives an intuition for the general rule

$$\omega_m = \omega_0 + \Delta_\omega \sum_{m'=1}^{m} x_{m'} \text{ for } m = 1, \ldots M - 1.$$

and defines the approach to sample $W_1, \ldots, W_{M-1}$ via change of variables such that $W_m = \omega_0 + \Delta_\omega \sum_{m'=1}^m X_{m'}$ for $m = 1, \ldots M - 1$. The realisation of $W_1, \ldots, W_{M-1}$, denoted by $\omega_1, \ldots, \omega_{M-1}$, represent the break points defining partitions $\mathcal{I}_1, \ldots, \mathcal{I}_M$. Also, we recall that ω_0 and $W_M = \omega_M$ are fixed.

We model M independent not identical partitions of the time-domain interval $\mathcal{T}$ into D subintervals by following the same steps. We define M independent multinomial random variables that are D-dimensional, each, denoted by $\mathbf{X}'_m$ for $m = 1, \ldots, M$, which entries $X'_{m,d}$ on the support of $\{0, \ldots, N_\tau\}$, for $d = 1, \ldots, D$, specify how many subsequent grids $\mathcal{T}_{n_\tau}^{grid}$ are connected to construct partitions $\mathcal{T}_{m,d}$ of $\mathcal{T}$ and determine break points $s_{m,d-1}, s_{m,d} \in \mathcal{T}$. We denote their distributions by $\pi(\mathbf{x}'_m; \mathbf{p}'_m)$ for $\mathbf{p}'_m = [p'_{m,1}, \ldots, p'_{m,D}]$ such that $\sum_{d=1}^D p'_{m,d} = 1$. For every $m = 1, \ldots, M$ this construction satisfies $\sum_{d=1}^D X'_{m,d} = N_\tau$ and

$$s_{m,d} = t_0 + \Delta_\tau \sum_{d'=1}^{d} x'_{m,d'} \text{ for } d = 1, \ldots, D-1,\ m = 1, \ldots M.$$

where $x'_{m,d}$ is a realisation of $X'_{m,d}$. Therefore, the random variables $S_{m,1}, \ldots, S_{m,D-1}$ for $m = 1, \ldots, M$ are defined via change of variables such that $S_{m,d} = t_0 + \Delta_\tau \sum_{d'=1}^d X'_{m,d'}$ for $d = 1, \ldots D-1$ with realisations $s_{m,1}, \ldots, s_{m,D-1}$ representing the break points of the partitions $\mathcal{T}_{m,1}, \ldots, \mathcal{T}_{m,D}$. Again, we recall that t_0 and $S_{m,D} = t_N$ are fixed for every $m = 1, \ldots, M$.

We can now connect this formulation back to the IS framework in the previous section as follows. Given this model, the joint distribution of $\Psi = [W_1, \ldots, W_{M-1}, S_{1,1}, \ldots, S_{M,D-1}]$ can we written as

$$g(\psi; \varphi) = C\ \pi(\mathbf{x}_m; \mathbf{p}) \prod_{m=1}^{M} \pi(\mathbf{x}'_m; \mathbf{p}'_m).$$

Using this IS distribution we can now rewrite the IS parameter estimation rule under CEM framework, according to Eq. (18) as follows, using

$$\log g(\psi; \varphi) = \log C + \log(N_\omega!) + \sum_{m=1}^{M} \{\log(x_m!) + x_m \log(p_m)\}$$

$$+ M\log(N_\omega!) + \sum_{m=1}^{M}\sum_{d=1}^{D}\left\{\log(x'_{m,d}!) + x'_{m,d}\log(p'_{m,d})\right\}.$$

to obtain the estimation equation for the IS parameters with constraint imposed on $\mathbf{P} = [\mathbf{p}, \mathbf{p}'_1, \ldots, \mathbf{p}'_M] \in [0,1]$ under a Lagrangian constrained parameter estimation given as follows:

$$\begin{aligned}\Lambda(\mathbf{P}, \boldsymbol{\lambda}) = &\sum_{s=1}^{S}\left\{\mathbf{1}_{\left\{KL(\hat{\pi},\pi;\psi^{(s)})\leq\gamma\right\}}\left(\log C + \log(N_\omega!) + \sum_{m=1}^{M}\left\{\log(x_m^{(s)}!) + x_m^{(s)}\log(p_m)\right\}\right.\right.\\ &\left.\left. + M\log(N_\omega!) + \sum_{m=1}^{M}\sum_{d=1}^{D}\left\{\log(x_{m,d}^{'(s)}!) + x_{m,d}^{'(s)}\log(p'_{m,d})\right\}.\right)\right\}\\ &+ \lambda\left(1 - \sum_{m=1}^{M} p_m\right) + \sum_{m=1}^{M}\lambda_m\left(1 - \sum_{d=1}^{D} p'_{m,d}\right).\end{aligned}$$

where $\mathbf{P}$ represents the IS distribution parameters to be estimated and vector $\boldsymbol{\lambda} \in \mathbb{R}^{M+1}$ are the Lagrangian multipliers. If one then seeks the First Order Conditions for this Lagrangian, one obtains the system of equations that admit a feasible solution as follows:

$$\begin{cases} \frac{\partial\Lambda(\mathbf{P},\lambda)}{\partial p_1} = \sum_{s=1}^{S}\left\{\mathbf{1}_{\left\{KL(\hat{\pi},\pi;\psi^{(s)})\leq\gamma\right\}}\frac{x_1^{(s)}}{p_1}\right\} - \lambda = 0 \\ \vdots \\ \frac{\partial\Lambda(\mathbf{P},\lambda)}{\partial p_M} = \sum_{s=1}^{S}\left\{\mathbf{1}_{\left\{KL(\hat{\pi},\pi;\psi^{(s)})\leq\gamma\right\}}\frac{x_M^{(s)}}{p_M}\right\} - \lambda = 0 \\ 1 - \sum_{m=1}^{M} p_m = 0 \end{cases}$$

$$\Rightarrow \begin{cases} p_1^* = \frac{1}{\lambda}\sum_{s=1}^{S}\left\{\mathbf{1}_{\left\{KL(\hat{\pi},\pi;\psi^{(s)})\leq\gamma\right\}}x_1^{(s)}\right\} \\ \vdots \\ p_M^* = \frac{1}{\lambda}\sum_{s=1}^{S}\left\{\mathbf{1}_{\left\{KL(\hat{\pi},\pi;\psi^{(s)})\leq\gamma\right\}}x_M^{(s)}\right\} \\ \sum_{m=1}^{M} p_m = 1. \end{cases}$$

These solutions to the IS distribution parameter estimates can be further simplified by noting that since $\sum_{m=1}^{M} p_m = 1$ and $\sum_{m=1}^{M} x_m^{(s)} = N_\omega$ one can obtain:

$$\frac{1}{\lambda}\sum_{s=1}^{S}\left\{\mathbf{1}_{\left\{KL(\hat{\pi},\pi;\psi^{(s)})\leq\gamma\right\}}\sum_{m=1}^{M} x_m^{(s)}\right\} = 1 \Rightarrow \lambda = N_\omega\sum_{s=1}^{S}\mathbf{1}_{\left\{KL(\hat{\pi},\pi;\psi^{(s)})\leq\gamma\right\}}$$

and finally

$$\hat{p}_m = \frac{\sum_{s=1}^{S}\left\{\mathbf{1}_{\left\{KL(\hat{\pi},\pi;\psi^{(s)})\leq\gamma\right\}}\frac{x_m^{(s)}}{N_\omega}\right\}}{\sum_{s=1}^{S}\mathbf{1}_{\left\{KL(\hat{\pi},\pi;\psi^{(s)})\leq\gamma\right\}}} \tag{19}$$

Following the same steps, we have that

$$\hat{p}'_{m,d} = \frac{\sum_{s=1}^{S}\left\{\mathbf{1}_{\left\{KL(\hat{\pi},\pi;\psi^{(s)})\leq\gamma\right\}}\frac{x_{m,d}^{'(s)}}{N_\tau}\right\}}{\sum_{s=1}^{S}\mathbf{1}_{\left\{KL(\hat{\pi},\pi;\psi^{(s)})\leq\gamma\right\}}} \tag{20}$$

Note that the support of the random variables introduced in this subsection includes zero, and this may lead to the situation that some partitions are of zero length. If that happens, the breakpoints $\omega_1, \ldots, \omega_M$ and $s_{1,1}, \ldots, s_{M,D-1}$ are not admissible as they may not form increasing sequence. Consequently, they do not belong to the feasible set Ψ. To address this difficulty, we may consider two procedures

1. sample directly from the conditional distribution

$$X_1, \ldots, X_M | X_1 \neq 0, \ldots, X_M \neq 0$$
$$X'_{1,1}, \ldots, X'_{M,D} | X'_{1,1} \neq 0, \ldots, X'_{M,D} \neq 0.$$

2. sampling from the Multinomial distribution and and force non zero realisation by removing any realisations that contain 0 entry to meet the conditions of the feasible set.

An algorithm for the CEM method based on this IS distribution construction is provided in the Appendix C.

6 Application: Speech Based Medical Diagnostics

In this section, we introduce how we will adopt the aforementioned Stochastic Embedding of EMD method into a medical signal processing application based on diagnostics of Parkinson's Disease using patient speech signals. An emerging area of speech analysis is developing in the domain of medical diagnostics; see examples in Parkinson's disease in [17], Alzheimer's in [18] and a recent tutorial on such applications in [19]. It has been known for some time by medical practitioners that many symptoms of numerous neurological and motor-neuron diseases may manifest with impacts on speech through enunciation, slurring, delayed recall leading to unvoiced pauses of unusual duration, stutter or other various effects. Many such diseases are degenerative and require continuous monitoring of the patient's status to ensure that treatment regimes adapt to the disease progression for each individual.

6.1 Motivation for Speech Based Parkinson Disease Diagnostics

The Parkinson's disease (Parkinson's disease) speech diagnostic analysis undertaken in the application framework developed in this paper builds upon the background proposed in [20]. In this work, the authors introduce an alternative method to detect speech abnormalities caused by Cerebellar Ataxia. This corresponds to impaired coordination due to a dysfunction of the cerebellum, characterised by movement abnormalities such as dysmetria, dysdiadochokinesia, and dyssynergia, amongst others. These abnormalities affect all kinds of movements, including speech, and hence lead to what is termed "ataxic speech". Signs of ataxic speech could be scanning speech ("excess and equal stress"), a reduced speech rate and deviant prosodic (i.e. rhythmical and melodic), modulation of verbal utterances, rhythmical irregularities during (fast) repetitive productions of single or multiple syllables (known as "oral dysdiadochokinesis"), a more significant variation in pitch and loudness and disturbed articulation of both consonants and vowels with reduced intelligibility (see [21], [22], [23]).

Several medical conditions could generate ataxic speech; in this work, we are interested in speech impairment movements caused by Parkinson's disease (see [24]). Parkinson's disease is a degenerative disorder of the central nervous system resulting from the death of dopamine-containing cells in the substantia nigra, a region of the midbrain. It is the second most common neurodegenerative disorder after Alzheimer's disease [25], [26], [27] and includes both motor (tremor, rigidity, bradykinesia, and impairment of postural reflexes) and non-motor signs (cognitive disorders and sleep ande sensory abnormalities). Several studies reported a 70-90% prevalence of speech impairments once the disease makes its appearance (see [24]). Moreover, it might be one of the earliest Parkinson's disease indicators (see [28],) with research showing that 29% of patients consider it one of their greatest obstacles ([29]). Both motor symptoms and speech movements abnormalities worsen with the progression of the disease in a nonlinear fashion ([28], [30]]. At the final stage of the disease, articulation is frequently the most impaired feature (see [24], [31], [32]). Medical treatments or surgical intervention can alleviate the course of the disease; however, there is no definite cure, and, therefore, an early diagnosis is highly critical to lengthen and improve the patient's life ([33], [34]).

6.2 Comparative Benchmark Models for Parkinson Disease Speech Analysis

Among the various empirical tests considered for PD dysfunctions evaluation, there are also speech and voice tests, where an expert is subjectively assessing the patient's ability to perform a range of tasks with a perceptual judgement relying on standardised clinical scales. The standard metric specifically designed to follow PD progression, introduced in 1987, is called the "Unified Parkinson's Disease Rating Scale" (UPDRS) and corresponds to a questionnaire which combines several sections to produce a comprehensive and flexible tool to monitor the course of Parkinson's and the degree of disability, see [72] and [73]. A UPDRS assessment produces an integer number providing information about the stage of symptoms. Speech has two explicit labels in this questionnaire, namely UPDRS II-5 and UPDRS III-18, ranging between 0-4, with 0 representing the less severe stage given as "Normal speech" and 4 being the most severe stage given as "Unintelligible most of the time".

One challenge with such a survey based diagnosis is that even for expert specialist doctors, there is difficulty to find a consistent reference baseline that is standardised across the profession. This leads to a desire to develop a standardised objective uniform tool assessing PD ataxic speech such that it would ideally identify acoustic disturbances in displacement, direction and rate (or velocity), see discussion in [20]. As highlighted in [27], the final goal of such a tool would be to detect the presence of the disease and, afterwards, to surveil it to track a patient's advancement through the different stages of the PD disease. In [34], a noninvasive telemonitoring solution is constructed by exploiting

linear and nonlinear signal processing algorithms to extract useful clinically features. The authors propose a mapping between dysphonia measures and stages of UPRDS. In these works and, in general, different tasks have been used to evaluate PD speech progressions: voice sustained phonation, rapid syllable repetition, variable reading of short sentences, longer passages and freely spoken spontaneous speech. Moreover, multiple speech features referring to different voice characteristics (as acoustic, prosodic, and glottal features) have been considered. The reader might refer to [26], [27] and [34] as main references for further description of both tasks and features.

In speech classification tasks, be it speaker identification, disease detection or other use cases, numerous studies have shown that the majority of the discriminatory power in detection of speech variations arises from a type of individuals "vocal signature" or "vocal figure print" known as the individual speech formants structure. Speech formants are a concentration of speech acoustic energy, usually occurring at approximately each 1,000Hz frequency band, directly related to the oscillatory modes of resonance of an individual vocal tract structure [74]. Several alternatives can be employed to extract aspects of formant feature information, often based on basis decomposition techniques [75, 76, 77] aiming to separate the signal into components whose frequency spectra could be preferably dominated by a single non-overlapping formant frequency. A widely used technique is to adopt warped filter basis extraction methods applied to windowed raw speech signal segments. A popular choice in practice is the Mel Frequency Cepstral Coefficients (MFCCs), see [78]. The MFCCs capture magnitude-based cepstral information, measuring in practice the short-term power spectrum of a speech signal based on a linear cosine transform of a log power spectrum through a nonlinear mel scale of frequency [16]. These features have successfully used in health diagnostic for ataxic speech ([26], [27], [34]). We are interested in the background proposed in [20].

The main contribution of [20] is to consider phase-based cepstral features combined with the magnitude cepstrum as a human signature to detect speech abnormalities of ataxic speech. While the magnitude cepstrum has been widely used in the analysis of ataxic speech (see [79], [80]), the phase cepstrum has often been discarded for two main reasons: the difficulty in phase wrapping and the conventional view of the human auditory system as "phase deaf". This perspective has recently changed, with several studies testifying that the change of sound phase has an instead significant impact on auditory perception ([81], [82], [83]). Specifically, [20] made use of the modified group delay function (MGD) ([84]) to derive phase-based cepstral coefficients (MGDCCs) and combines them with magnitude cepstrum based features, i.e. the MFCCs ([85], [86]). A Random Forest and an SVM framework are used to assess the discrimination power of these features in detecting ataxic speech.

The work in this paper will extend and significantly enhance the features utilised in [20] to significantly refine and improve the accuracy of both PD detection in early onset patients as well as the analysis of progression through the various stages of PD. We will set as the benchmark comparison the framework of [20], and we will compare our proposed EMD stochastic embedding approach combined with a purpose built version of the Likelihood Ratio test in order to make inference on disease state, which will be compared to performed from the aforementioned works SVM classifier based results. As presented in [87] it is indeed possible to compare and relate such results. The considered dataset, described in subsection 7.1, leads to a text-dependent environment where both controls (healthy subjects) and sick patients read a given text. Reasons to employ such a specific set of sentences making use of the reading text task are below clarified.

One of the features in [20] corresponds to the MGDCCs, exploiting the modified group delay function. As studied in [39], [40], [41], the instantaneous frequency (IF) is a function assigning a frequency to a given time, whereas the group delay (GD) is a function assigning a time to a given frequency and, therefore, the question of interest here is whether the two functions are inverses of each other. In practice, this is not always the case because the IF function may not be invertible. Two conditions need to be verified for the laws of the two functions to be inverse of one another: (1) the variations in time of the IF is monotonic, and (2) the bandwidth-duration (BT) product is sufficiently large. This restricts the signals of interest to be a monocomponent signal whose IF is a monotonic function of time. Furthermore, when this is the case, the laws carry an enclosed physical meaning being the IF describes the frequency modulation of the signal while the GD represents the time delay of the signal. Thus, when studying features based on such functions, a monocomponent signal is required, or the interpretability of the results might be misleading. Two of our system models strongly rely on this discussion and propose stochastic embeddings based on the IMFs, which are, by definition, monocomponent functions. Furthermore, system model 3 is built upon the IFs of the IMFs. Therefore, our final aim is to provide two models distinguishing the two families of controls and PD patients based on the IMFs and the IFs to depict ataxic speech.

6.3 Proposed Stochastic EMD Hypothesis Testing Framework for Parkinson's Detection

In this section, it is demonstrated how to use the GP stochastic models from SM1, SM2 or SM3 to develop a hypothesis testing framework that can be utilised to perform inference on the presence or absence of Parkinson's disease features in speech recorded from patients. For a given system model (SM1, SM2 or SM3), the EMD method was used

to extract IMFs from two different sampled populations of patients, those diagnosed at various stages of Parkinson's disease progression vs a second population sample of healthy patients. Given the sample speech signals from each population sample, the training stage of the inference procedure involved performing EMD method on the speech signal samples, extracting IMFs and IFs, calibrating the Fisher kernel via a generative embedding model using linear time series models for each IMF, extracting the optimal IFs time-frequency partition Π^* using CEM and then using the stochastic formulation of each system model SM1, SM2 or SM3 to train the subsequent GP models. Since the stochastic embedding of the EMD method under SM1, SM2, or SM3 are each based on GP models, we will be able to generically present the hypothesis testing framework as follows using a generic kernel $k(t,t')$, which will be replaced with the relevant kernel used to specify SM1, SM2 or SM3 as discussed in previous sections of this manuscript. The result of this process, described in more detail in the subsequent results section, will be an estimated representative stochastic EMD embedded GP population model for sick patients with Parkinson's disease (distinguished by a subscripted process $\widetilde{S}(t)_1$) and a corresponding estimated representative stochastic EMD embedded GP population model for the healthy patients (distinguished by a subscripted process $\widetilde{S}(t)_0$) in the medical study. These were then used to develop a likelihood ratio test (LRT) hypothesis testing framework that could be utilised out-of-sample to detect unclassified patients as either not presenting with any speech disorder based symptoms consistent with Parkinson's disease or presenting with speech disorder symptoms consistent with Parkinson's disease. Hence, the two models that will be compared under the LRT testing framework are given by:

$$\begin{aligned}\text{Model}_0:\ & S_0(t) \sim \mathcal{GP}\left(0, k_0(t,t')\right) \quad \forall t \in [t_1, t_N] \\ \text{Model}_1:\ & S_1(t) \sim \mathcal{GP}\left(0, k_1(t,t')\right) \quad \forall t \in [t_1, t_N]\end{aligned}$$

This results in a null and alternative hypothesis to test given as follows:

$$\begin{aligned}H_0:\ & \widetilde{S}_0(t) \overset{d}{=} \widetilde{S}_1(t) \ \ i.e.\ \ \mathcal{GP}\left(0, k_0(t,t')\right) = \mathcal{GP}\left(0, k_1(t,t')\right) \quad \forall t \in [t_1, t_N] \\ H_1:\ & \widetilde{S}_0(t) \overset{d}{\neq} \widetilde{S}_1(t) \ \ i.e.\ \ \mathcal{GP}\left(0, k_0(t,t')\right) \neq \mathcal{GP}\left(0, k_1(t,t')\right) \quad \forall t \in [t_1, t_N]\end{aligned}$$

Since a GP is also specified by its sufficient mean and covariance functions, testing for equality of distributions will be equivalent to testing for equality of the mean and covariance functions. The problem formulation in this manuscript is designed in a manner that the class of kernels utilised are restricted so that the Model_0 is nested in the Model_1, and hence these hypotheses can be tested with the Generalised Likelihood Ratio Test (GLRT). This is a GLRT formulation since the kernel hyper parameters are estimated. One can then obtain the test statistic by considering the log likelihood of each model under the GP stochastic embedding obtained from both the sick and healthy population samples for any of the system models (SM1, SM2 or SM3) given for samples $\tilde{s}(\boldsymbol{t}) = [\tilde{s}(t_1), \tilde{s}(t_2), \ldots, \tilde{s}(t_N)]$ generically by:

$$\hat{L} = -\tilde{s}(\boldsymbol{t})^{\intercal}\hat{\mathbf{K}}_0^{-1}\tilde{s}(\boldsymbol{t}) - \log\left(\det\left[\hat{\mathbf{K}}_0\right]\right) + \tilde{s}(\boldsymbol{t})^{\intercal}\hat{\mathbf{K}}_1^{-1}\tilde{s}(\boldsymbol{t}) + \log\left(\det\left[\hat{\mathbf{K}}_1\right]\right) \tag{21}$$

Defining d as the difference in dimensionality of model parameter vectors for H_0 and $H_0 \cup H_1$, one has an asymptotic distribution under the null hypothesis, for the test statistic given by

$$-2\log L \sim \mathcal{X}_d^2$$

The above tests will be carried to identify the discrimination power associated with the different IMFs stochastic embedding proposed. In this way, each embedded IMF and band limited IMFs will be individually tested.

7 Experiments

A study of Parkinson's speech samples is developed to illustrate the performances of the system models introduced for the stochastic embedding of EMD when combined with the inference procedure from Section 6.3. The reference benchmark comparison for detection of Parkinson's from speech signals will be based on the stochastic EMD model extensions we develop to generalise the framework recently introduced by [20].

We begin with an overview of the selected Parkinson's speech dataset and its experimental setup. This is proceeded by a section explaining the required pre-processing along with the procedure employed to balance the datasets since the study had an uneven number of labelled sick vs healthy patients. The construction of training and testing sets with the experimental design are then explained. We defer the interested reader to the specialised details relating to the practical pre-processing and Fisher kernel construction methods used in the provided Supplementary Materials , sections IV and V. Subsequently, the testing procedure for the validation model phase is described. Then, the description of our guideline reference model, whose method is introduced in 6.2 is provided. Finally, the results obtained through our proposed models are described. Table 1 shows the different features used, over which data and the correspondent

classifier. The classification procedure will be conducted at a patient level, providing a text-dependent and a speaker-dependent environment. Note that the python code required for the implementation of the three system models is given within this Github page https://github.com/mcampi111, where it is possible to find a repository named "EMD-Stochastic-Embedding-for-PD-Speech" containing the code.

Experiment Description			
System	**Feature**	**Data**	**Classifier**
Benchmark	MFCCs, MGDCCs	$\tilde{s}(t)$	SVM
SM1	GP	$\tilde{s}(t)$	GLRT
SM2	GP-EMD	$\gamma_1(t), \gamma_2(t)\gamma_3(t)$	GLRT per IMFs
SM3	GP-EMD	$\gamma_1(t)^{(\mathrm{BL})}, \gamma_2(t)^{(\mathrm{BL})}, \gamma_3(t)^{(\mathrm{BL})}$	GLRT per BLIMFs

Table 1: Description of the experimental set up. The selected benchmark features correspond to the traditional MFCCs and MGDCCs on the given speech signals $\tilde{s}(t)$. The configuration employed for the extraction procedure of these features are provided in subsection 7.4. Then, each system model is performed, and the GLRT is applied. Note that, when it comes to SM2 and SM3, we will consider the first three IMFs or the first three BLIMFs only since they are the ones that detect the great majority of formants required for the classification of Parkinson's disease. Both the SVM and the GLRT will be done by patient, setting up a text-dependent and a speaker-dependent environment.

7.1 Data Description and Experimental Set Up

The speech dataset considered for the analysis was provided by [2]. It contains speech recordings from two populations: healthy participants; and patients affected at various stages of Parkinson's disease progression. The recording environment uses a typical examination room for UK medical practices with dimensions of ten square meters in area and a reverberation time of approximately 500ms to perform the voice recordings. The voice recordings are performed in the realistic situation of doing a phone call and have been performed within the reverberation radius; hence, they can be considered "clean". The sampling rate is standard for speech at 44.1 kHz and a bit depth of 16 Bit (audio CD quality).

The dataset is split between two sets of recordings: in the first one, the selected participants are asked to make a phone call and then read out two tests: "The North Wind and the Sun" and "Tech. Engin. Computer applications in geography snippet". These were selected in the experimental design described in [88] since the first contains poetic structures and the second contains technical jargon, both of which are less familiar to participants' everyday text. In the second set of recordings, the participants start a spontaneous dialogue with the test executor, who asks random questions. In our case studies, we only considered the first set of recordings. Hence, the used task to assess ataxic speech in PD disease is reading a given text. The second set of recordings corresponding to spontaneous dialogue is considered highly challenging for this assessment. However, it could be employed in further research and used to study surveillance of the disease and its progression. Further details about this are given in the Supplementary Materials, in section III. The reader is referred to [88] for further detail on the collection process and experimental set-up used in the clinical setting.

We note that this database of speech signals was specifically selected given the quality of the recordings and its recording procedure. The procedure used is most aligned with the standard medical practice of relevance to telemonitoring solutions for remote PD disease detection prior to requesting the patient to travel to a hospital for further in-person testing. This is useful for pre-screening those likely to need to travel for initial diagnosis as well as for analysis of the impact on speech for disease progression analysis for those living remotely from specialist care or those unable to easily travel from their house to the hospital on a regular basis due to deterioration in their health resulting from various PD symptoms.

There are 37 participants in total, of which 21 are healthy and 16 are sick, affected by Parkinson's disease at different stage levels. Amongst the 21 healthy participants, 19 are female, while 2 are male. Of the 16 sick participants, 4 are female, and 12 are male. The dataset is therefore significantly unbalanced within both classes, i.e. healthy versus sick and male versus female. Furthermore, the Parkinson's participants are labelled according to the following scores: the HYR score, the UPDRS II-5 score and the UPDRS III-18 score introduced in 6.2. Considering the UPDRS II-5 score, the Parkinson's participants are classified in a range between 0 and 3 at maximum, particularly for the female patients, 2 are at a 0 stage level, and 2 are at a 1 stage level. In the case of the sick male patients, 5 male patients are at a 0 stage level, 4 patients at 1 stage level, 2 patients at 2 stage level and 1 patient at a 3 stage level. Hence, a further level of unbalancedness is introduced. Section III of the Supplementary Material provides a more detailed summary of the

described database. Table 2 summarises the above description. As a result, a procedure to balance the dataset and its pre-processing is presented in the following subsection.

MDVR-KCL Dataset Description										
PD Status	Healthy		Sick							
Gender	Female	Male	Female			Male				
UPDRS II-5 score	–	–	0	1	2	3	0	1	2	3
# of Speakers	19	2	2	2	–	–	5	4	2	1

Table 2: Description of the "Mobile Device Voice Recordings at King's College London (MDVR-KCL)". The number of speakers is 37, split between healthy and sick patients. Furthermore, the gender and the UPDRS II-5 score are introduced in the table. It is possible to observe how unbalanced the dataset is, particularly regarding gender and the UPDRS II-5 score. For each speaker, the dataset provides two sets of recordings. In our experiments, we use the read text and set the scenario to a text-dependent one. Moreover, we conduct our analysis by patient, and therefore we will be in a speaker-dependent setting.

7.2 Pre-Processing, Balancing the Dataset and Construction of Training and Testing Segments Sets

This subsection outlines a brief description of the pre-processing performed to obtain a balanced selection of speech records for the testing and inference tasks undertaken. The recordings taken into account are the read text for each participant. Within the recording procedure, each participant was asked to make a phone call and then read two different texts (above mentioned). Each audio file corresponds to a continuous, unsegmented recording of the read text at the sampling rate was 44.1kHz. Therefore, we will have one audio file for each patient denoted as $s(t)$. Depending on the patient, the reading order might change, and the recording lengths (due to different reading paces) vary between 73s and 203s. We removed the silence at the beginning and the end of the recordings and the initial participant's dialogue with the interlocutor.

In order to perform the EMD, the underlying signal needs to be continuous. Therefore, we fit a cubic spline with knots points placed at the sample points through each of the recordings, and we denote it as $\tilde{s}(\boldsymbol{t})$. Afterwards, we split each recording into batches of 5000 samples length for computational reasons, which approximately corresponds to 0.113 seconds (at a sample rate of 44.1kHz). Given that the audio files have different lengths, the number of resulting minibatch segments of 5000 samples for each patient differs. Figure 4 shows the number of segments for each patient divided by the scores of the UPDRS II-5 for both female (left panel) and male (right panel) patients.

As noted, one can see that the populations represented are highly unbalanced for the number of male and female patients, the different categories of the UPDRS II-5 score and the number of sick and healthy patients. To balance the representation of each patient, we compute the minimum number of segments for each patient by gender and then randomly select that minimum from each patient. We denote the minima as $N_m = 372$ and $N_f = 442$. Therefore we will have $N_m \times 14$ segments for the male patients and $N_f \times 23$ segments for the female patients.

Once we have obtained a balanced representation of each patient with respect to the number of segments, the following step consists of constructing training and testing sets of segments for our classification task, split into model estimation and model validation. Consider the female case as an example and note that an equivalent procedure is applied to the male case. To construct the training set, we firstly left one patient out for the testing set. Then from the remaining number of patients segments, i.e. $N_f \times 18$ for the healthy case and $N_f \times 3$ for the sick case, we randomly extract 80% of N_f corresponding to 354 segments. Hence, we will have 354 segments representing the class of healthy patients and 354 segments representing the class of sick patients, randomly extracted from 18 and 3 patients equally represented. For the testing set instead, we randomly select 20% of N_f from the two left out patients segments, one for the healthy and one for the sick classes, corresponding to 89 segments. Therefore, we will have 89 segments for the healthy patient left out and 89 segments for the sick patient left out. We then rotate the left out patients and repeat the procedure. In such a way, when testing a given patient, this will not be contained in the training set and over-fitting is then avoided. We will refer to $\tilde{s}(\boldsymbol{t})_0^{tr}$ and $\tilde{s}(\boldsymbol{t})_1^{tr}$ with $tr = 1, \ldots, N_{tr}$ for the training set and to $\tilde{s}(\boldsymbol{t})_0^{ts}$ and $\tilde{s}(\boldsymbol{t})_1^{ts}$ with $tr = 1, \ldots, N_{ts}$ for the testing set. Note that for the male case, $N_{tr} = 298$ and $N_{ts} = 75$.

7.3 Testing Procedure for the Model Validation Phase

The next step uses these training data sets to develop a fitting procedure which involves the construction of the generative embedding Fisher kernels from the EMD outputs as described in Section 5.1. This requires practical parts beyond the paper's main scope, detailed in the accompanying Supplementary Materials, see section IV. There are two main aspects which are relevant at this point and that the reader should consider. First, the fitting procedure aims to identify fast changes that cannot be perceived by the human ear, i.e. by a doctor. Therefore, the procedure is done on mini-batches of approximately 2.2ms, meaning that each segment will be further split into mini-batches. Each mini-batch can then be characterised by a simple model whose set of hyperparameters will be informative with respect to fast changes signalling the presence/absence of the disease. Second, it is highly likely that not all mini-batches are discriminatory for such a task. Hence, a model selection criterion is required. Once a set of best discriminatory models are identified, a rule able to describe a unique family (i.e. female sick, female healthy, male sick and male healthy) of speech signals that can then be tested is required. The steps of the fitting procedure are given as follows: (1) Split the segments into mini-batches; (2) Fit a set of ARIMA models (see Supplementary Materials for further details on this) on each mini-batch; (3) Select the best model per mini-batch and then per segment according to the Akaike Information Criterion; (4) save the obtained model hyperparameters that will then be used to derive a Fisher score employed in the testing procedure; (5) save the proportion for each winner model, i.e. how many times a specific model for the mini-batches was selected as best over its segment. In such a way, a "weighted" rule will be defined for the definition of the Fisher score in the testing procedure. Note that we will end with $N_f = 354$ best models for the female families (i.e. both sick and healthy) and $N_m = 298$ for the male families (i.e. both sick and healthy).

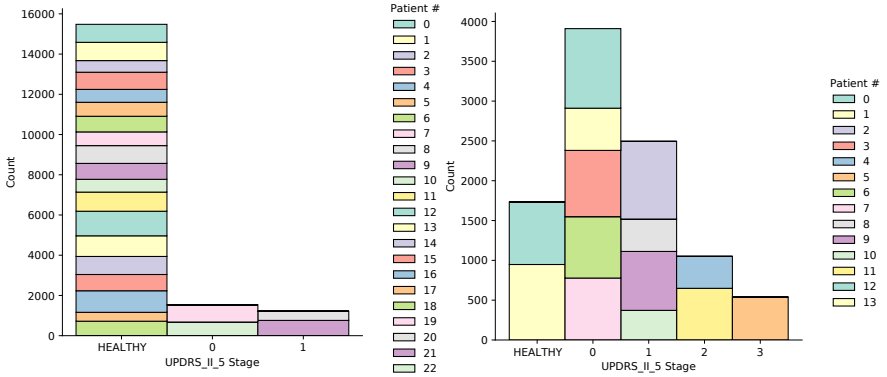


Figure 4: Barplots for the number of segments of length 5000 samples (approximately 0.113 seconds) for the female patients (left panels) and the male patients (right panels). The x-axis represents the different stages of the UPDRS II-5 where we also included the healthy patients. The y-axis represents the counts of the segments divided by patient.

The testing procedure computes the Fisher score vectors by evaluating the obtained best models on the testing data (also split by mini-batches) of each patient. By considering the healthy female case, for example, 354 models are evaluated on each mini-batch of every testing segment. In practice, one has 354 sets of hyperparameters describing one mini-batch, while the desired scenario would be having one set of hyperparameters per mini-batch. This is achieved by computing the Fisher scores for every best model per mini-batch and then aggregating them to have a unique vector testing the discriminatory power of the best models as a whole. An equivalent procedure is done for the sick female family on that same mini-batch and, therefore, one can redefine the GLRT test formulated in Eqn. 21 as

$$\begin{aligned}\hat{L} = &-(\tilde{\boldsymbol{U}}^j_{\boldsymbol{\theta}_0})\left(\tilde{\mathbf{K}}_0^{j\ S}\right)^{-1}(\tilde{\boldsymbol{U}}^j_{\boldsymbol{\theta}_0})^{\intercal} - \log\left(\det\left[\tilde{\mathbf{K}}_0^{j\ S}\right]\right)\\ &+(\tilde{\boldsymbol{U}}^j_{\boldsymbol{\theta}_1})\left(\tilde{\mathbf{K}}_1^{j\ S}\right)^{-1}(\boldsymbol{U}^j_{\boldsymbol{\theta}_1})^{\intercal} + \log\left(\det\left[\tilde{\mathbf{K}}_1^{j\ S}\right]\right)\end{aligned} \tag{22}$$

This shows that the test is done on the Fisher scores, rather than directly on the speech segments. Figure 5 shows the step of the described procedure. Furthermore, the details and derivation of such a procedure are outlined in Supplementary Materials, see section V for further details. In Eqn. (22), $\tilde{\boldsymbol{U}}^j_{\boldsymbol{\theta}_0}$ and $\boldsymbol{U}^j_{\boldsymbol{\theta}_1}$ represent the centred, weighted, aggregated Fisher scores evaluated on a testing mini-batch for healthy and sick family (of a specif gender) respectively. $\tilde{\mathbf{K}}_0^{j\ S}$ and $\tilde{\mathbf{K}}_1^{j\ S}$ represents the regularised Gram Matrices derived from such Fisher scores. Note that each Gram Matrix can be defined as

$$\tilde{\mathbf{K}}^j_{v\ (\kappa\times\kappa)} = \tilde{\boldsymbol{U}}^{j\ \intercal}_{\boldsymbol{\theta}_v}\tilde{\boldsymbol{U}}^j_{\boldsymbol{\theta}_v} \quad \text{for} \quad j = 1,\ldots,N_{f,t}$$

where $v \in \{0, 1\}$. The Gram Matrix regularisation is needed since computational instability could be encountered with the inversion of such a matrix or the log-determinant and corresponds to the covariance shrinkage estimator. Once the Gram Matrices are regularised, we added the superscript "S" for notational correctness. For further details, see the Supplementary Material. Once the GLRT has been done on each mini-batch of every segment, then the accuracy has been computed since this is a supervised learning procedure where we know in advance the labels of each segment. The results of the accuracy are provided in tables 3 and 4.

7.4 Results, Spectrogams and Formant Structures

In Tables 3 and 4, results are divided by gender and show accuracy scores achieved by benchmark and proposed models. Note that the accuracy is defined as the sum of the true positive and true negative detected examples over the sum of true positive, true negative, false positive and false negative. Each Table is split according to healthy and sick patients, ordered by their UPDRS score. In the female case, most of the patients are healthy; for the sick patients, there are only two stages, being identified as "0" and "1". In the male case, instead, there are only two healthy patients, while a great deal are instead sick patients. The UPDRS scores range between "0" and "3".

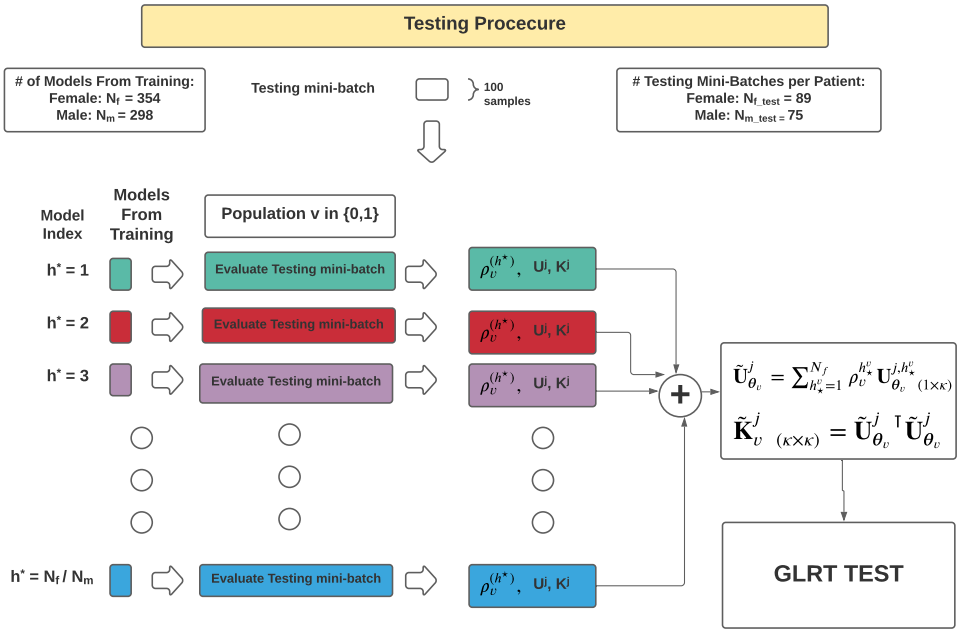


Figure 5: Figure showing a diagram for the steps required for the testing procedure of the model estimation phase. The GLRT test is computed on each mini-batch extracted by the segments of every patient. Note that each mini-batch is approximately 2.2.ms. The GLRT test is conducted on weighted and aggregated Fisher score vectors.

For comparison between the proposed solutions based on the EMD-GP models, we compare to standard reference benchemarks for speech analysis previously used in Parkinson's disease detection works. These benchmark features considered are the MFCCs, and the MGDCCs computed directly on the speech signals as proposed in [20]. In such a way, both the magnitude and the phase cepstral information are analysed. Before computing the coefficients, we first pre-emphasise and Hamming-windowed $\tilde{s}(t)$ to avoid issues of aliasing in discrete sample MFCCs or MGDCCs representations.

Each speech signal is subject to a 0.97 pre-emphasis factor. It is then segmented into frames of 25ms with 50% overlap, meaning, for a sampling frequency $f_s = 44.1$ kHz, that the total number of samples in each frame is $N_s = 1102.5$. [20] extracted both sets of coefficients on frames of the original speech signals and then averaged them across the considered frames. Hence, they end up with 12 averaged MFCCs and 11 averaged MGDCCs, for each speaker, which are then stacked in a vector for use in a kernel based classified using the Support Vector Machine optimisation maximum hyperplane classification method. In this work, the SVM is applied firstly by category, i.e. an SVM with the MFCCs and one SVM with MGDCCs only, and then stacked together within the same vector for classification.

We reproduced the same setting and ran an SVM as in their work based on features employing their configuration procedure. The radial basis function kernel has been selected, and a cross-validation procedure with 10 folds has been applied. We also provide results for SVMs (with equivalent configuration of the features and the SVM cross-validation) conducted without the averaging step of [20]. We have applied both methods, i.e. with and without the averaging of the basis coefficients. In general, [20] propose the averaging practice since it trades of accuracy for speed of computation. However, most of the discriminant power will lie in the anomalous changes of the various speech frames, and the averaging operation would smooth the energy content of the derived coefficients. The obtained results in our framework show, nevertheless, that the most significant problem of these decomposition techniques, i.e. the MFCCs or the MGCCs, is that they rely on the stationarity assumption, which is rarely achieved if not in optimal recording environments with silence and non-reverberation conditions and using a special microphone array. This is not possible in standard medical facilities or with voice recordings over wireless devices such as mobile phones. When this is not the case, such features tend to fail and provide unreliable results as what we have found. A further discussion of these challenges can be found in [16].

The results obtained using the proposed system models in this manuscript using the EMD-GP structures are then given in columns SM1, SM2 and SM3. While in the case of the benchmark features, the employed classifier was an SVM, the SMs relied on the GLRT. SM2 and SM3 results are provided for the first three IMFs and BLIMFs. As highlighted above and provided in [16] (and reference within), the first three IMFs capture most of the formants structure acting as a human fingerprint obtained from speech and provide a powerful discriminant tool for the characterisation of ataxic speech. Hence, our final goal is to detect this structure effectively to identify fast changes in its energy content in time and frequency, signalling the presence or absence of Parkinson's. In this work, results for the residual tendency and the rest of the IMFs are not presented. They have been tested, and no better results have been achieved. Another critical point is that the original IMFs $\gamma_1(t), \gamma_2(t), \gamma_3(t)$ often carry a great deal of noise. Therefore, a median-filter has been applied, providing a smoother version of such bases denoted as $\gamma_1^s(t), \gamma_2^s(t), \gamma_3^s(t)$ and results are shown based on these robust filtered IMF signals.

Once the EMD is computed, the IFs have been extracted. The following step is applying the cross-entropy method to compute the BLIMFs. We select the first three IFs for this step, i.e. $\omega_1(t), \omega_2(t), \omega_3(t)$, since the great deal of formants will be described by them. In the configuration of the CEM, we selected $M = 3$ and $D = 5$, $\rho = 0.2$, $\beta = 0.6$, $S = 100$, $N_\omega = 100$, $N_\tau = 100$ and a maximum number of CEM iteration was equal to 100. Alternatives have been considered, but similar results were obtained, and, therefore, we select the minimum number to obtain a low computational cost. Once performed, the CEM provides a set of grid points, i.e. ω_m and $s_{m,d}$ for $m = 1, \ldots, M$, $d = 1, \ldots, D$ which partition the time-frequency plane. Then the BLIMFs are derived as given in Eq. 12 and the GLRT test is applied as for SM2.

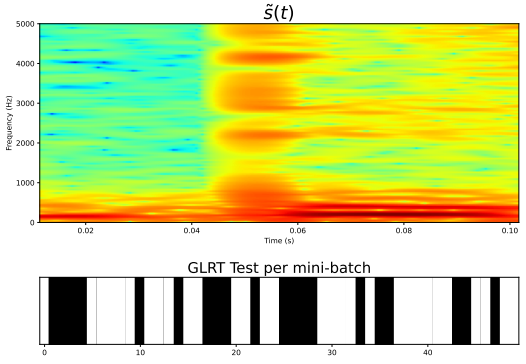
The analysis has been conducted for male and female speakers separately, primarily because it is widely known that formants differ significantly between genders, with female formants typically lying at higher frequencies than males. Therefore, any classification or inference procedure tackling speech analysis should consider gender and not pollute the classifier with resonant frequencies that are inaccurately detected since they belong to the other gender class. We provided a set of spectrograms given in Figure 6 showing speech segments of 5,000 samples for four different voices: the top left panel refers to the voice of a healthy female subject, while the top right panel represents the voice of a female sick patient.

The bottom panels are for male voices, healthy and sick, in the same order as above. We focus on the range of 0-5 kHz since the first five formants are visible. Hence, the y-axis varies within this range, while the x-axis represents time and is given in seconds (0.113 approximately). Focusing on the healthy subjects, the top left panel has an energy spectrum more spread out than the correspondent bottom one. This shows how, in general, female voices tend to have higher formants than male voices. Furthermore, F_0, also called fundamental frequency and capturing the pitch, for male voices is more pronounced and lives within 0-1kHz, while, for female voices, it often lies at higher frequencies. This is visible in the bottom panel, where the frequency content of 0-1kHz is stronger than frequencies within the rest of the spectrum. Furthermore, formants duration over time is usually more irregular for female voices than male ones; therefore, fast changes in time will be more challenging to detect for females than males. The right spectrograms refer to speech segments of sick patients.

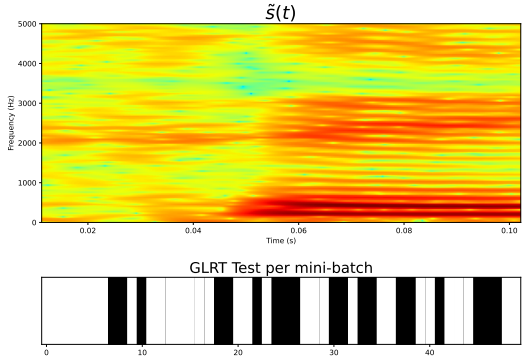
These plots demonstrate why it is possible to accurately detect Parkinson's disease with the proposed EMD-GP methods developed in this manuscript. One can observe the ataxic speech features present in sick patients compared to the non-ataxic speech of healthy patients. This manifests typically in clear spectral signatures that the EMD framework is able to accurately identify and then utilise in the EMD-GP testing framework for the GLRT test.

Furthermore, an additional effect one observes from this analysis is that the amount of energy intensity produced at various frequencies over time in the speech generated by sick patients with Parkinson's tends to be higher than the

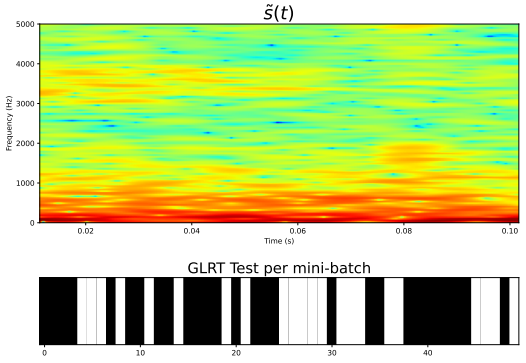
voice of a healthy subject. This is potentially indicative of lesser control of vocal structures used to modulate speech intensity in sick patients, which is also consistent with patients who tend to slur or drag words.



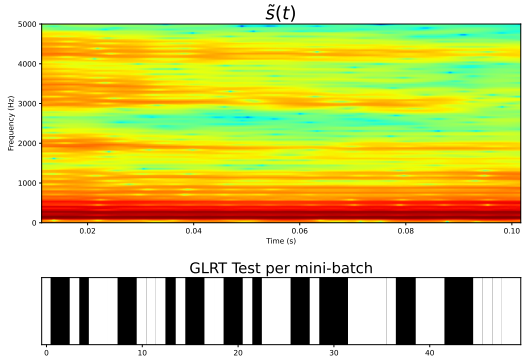
(a) Healthy female speech segment.



(b) Healthy female speech segment. UPDRS score equal to 1.



(c) Healthy male speech segment.



(d) Sick male speech segment. UPDRS score equal to 2.

Figure 6: There are two panels for every plot. The top panels are spectrograms of the original speech segments for four voices. The x-axis is time (0.113 s), given in seconds, the y-axis is frequency given in Hz (0-5000Hz). The second panel represents the results of the GLRT test conducted on every mini-batch of that segment. There are 50 mini-batches per segment. White corresponds to 0 and black to 1. 0 corresponds to equality in distribution, hence no disease detected, while 1 corresponds to the detection of Parkinson's disease.

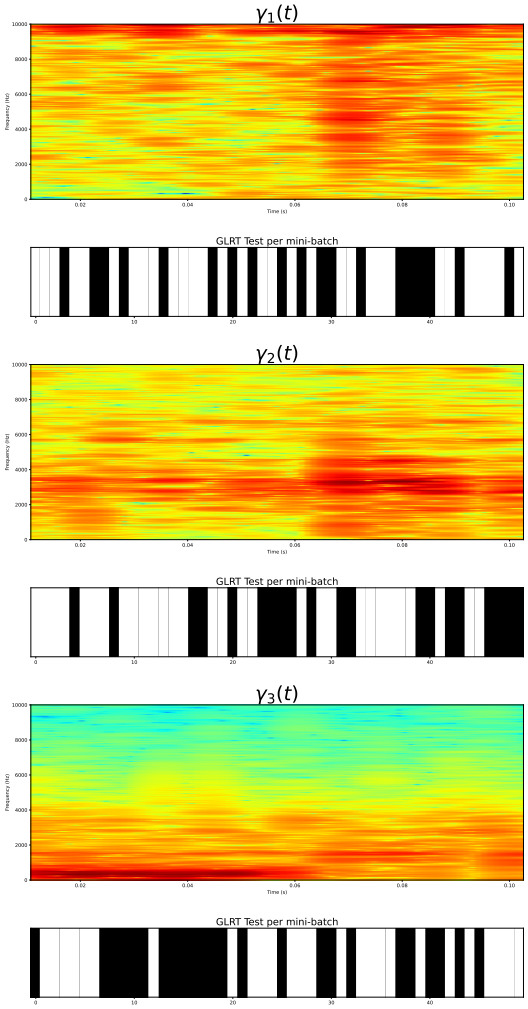
Therefore, this paper aims to construct an effective tool able to quantify such energy changes in both domains in a data-adaptive fashion. Since the location of the formants is strongly biometric for an individual, and they carry a high level of non-stationarity, the idea is first to isolate formants through basis functions that can deal with these properties and secondly to develop a statistical methodology which quantifies formants distributions that are indeed a priori unknown. Figure 6 shows four spectrograms and an extra panel per spectrogram representing the GLRT test computed on the mini-batches of that given segment. This is the same in the spectrograms given in Figure 7 where spectrograms of the IMFs and BLIMFs will be instead provided.

If we focus on Figure 7, one can observe that there are six spectrograms. The left panels are speech segments of the first three IMFs, i.e. $\gamma_1(t), \gamma_2(t), \gamma_3(t)$ extracted by the speech segment related to the sick male patient in Figure 7, (d). The right panels alternatively represent the spectrograms of the speech segments of the first three BLIMFs computed on the IMFs given in the left panels and denoted as $\gamma_1^{(\mathrm{BL})}, \gamma_2^{(\mathrm{BL})}, \gamma_3^{(\mathrm{BL})}$. This time we focused on a bigger frequency range, i.e. 0-10kHZ, to observe a broader spectrum.

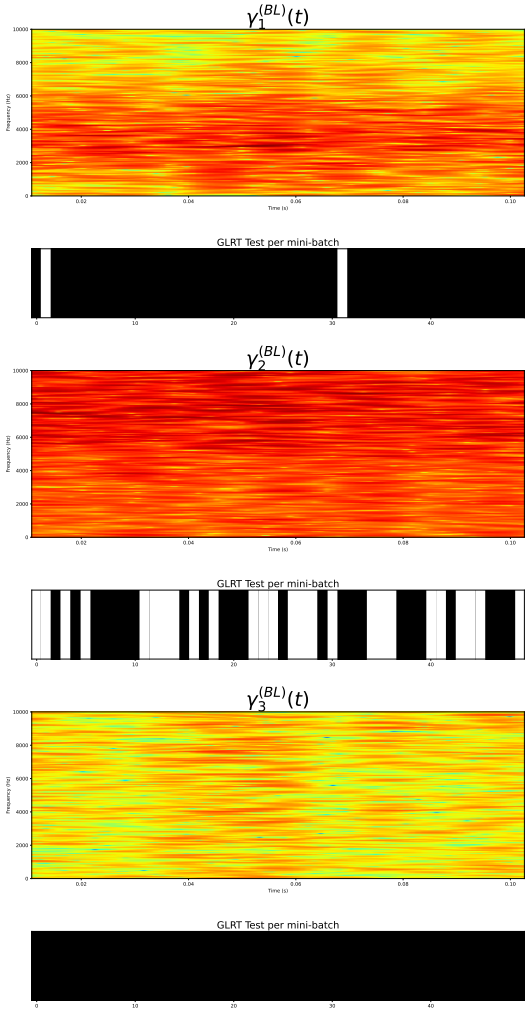
The figures clearly demonstrate that the first IMF captures the highest formants of the speech signal, the third and fourth formants. The second IMF detects the second formant and finally, the third IMF identifies the fundamental

frequency F_0. This can be observed in the left spectrograms, where the energy content decreases if one moves from the top to the bottom spectrograms.

By looking at the BLIMFs spectrograms instead, it is clear that the energy content has been reassigned within different regions since the IFs have been partitioned into an optimal partition obtained with the cross-entropy method presented in section 5.2. Indeed, $\gamma_1^{(\mathrm{BL})}$ appears to localize highest frequency content more efficiently than the basic IMF $\gamma_1(t)$. While the first IMF shows energy concentration at very high frequencies, i.e. around 9-10kHz, for most of the time, $\gamma_1^{(\mathrm{BL})}$ captures a strong energy concentration around 2kHz and 4kHz, reflecting the second and the third formants which are visible in Figure 6, (d). In the case of the IMFs, these formants are split between the second basis and third basis, which detects the fundamental frequency below 2kHz. Instead, $\gamma_2^{(\mathrm{BL})}$ presents an energy spectrum which contains a lot more energy than the correspondent second IMF.



(a) Speech segments of the first three IMFs extracted from the sick male speech segment given in Figure 6, (d).



(b) Speech segments of the first three BLIMFs computed on the IMFs of the the sick male speech segment given in Figure 6, (d).

Figure 7: There are two panels for every plot. The top panels are spectrograms of the speech segments IMFs (left) and the BLIMFs (right) obtained from the EMD of the male speech segment given in Figure 6, (d). The x-axis is time (0.113 s), given in seconds, the y-axis is frequency given in Hz (0-10000Hz). The second panel represents the results of the GLRT test conducted on every mini-batch of that IMFs or BLIMFS segment. There are 50 mini-batches per segment. White corresponds to equality in distribution, hence no disease detected, while black corresponds to the detection of Parkinson's disease.

We believe that this BLIMF isolates the noise spread across the three IMFs, and, therefore, retains information that is less useful and polluted for detecting the disease. Indeed, the spectrum looks uniform in energy concentration and recalls a spectrum of the white noise signal. The last BLIMF $\gamma_3^{(\mathrm{BL})}$ cannot localize the fundamental frequency correctly. However, this is now detecting its fast frequency changes dispersed across the entire spectrum. Therefore, the CEM can find a partition identifying basis functions that provide a more efficient decomposition in formant detection.

Figure 6, as well as Figure 7, show plots made of two subpanels. The first panel represents the spectrogram of interest, while the second one represents the GLRT test carried on the mini-batches of that considered speech segment, or, in the case of Figure 7, on the speech segment of the correspondent IMF or BLIMF. There are 50 mini-batches per segment; therefore, there is a band corresponding to 50 GLRT tests for every spectrogram. If the band of the GLRT test is coloured in white, it indicates that the GLRT test on that mini-batch found equality in distribution and, therefore, no presence of Parkinson's disease. In the opposite case, the GLRT test has detected differences in distributions, and it implies the detection of Parkinson's.

If one now considers Figure 6 which demonstrates the results for simply fitting a GP model to the speech signal (SM1), which does not use the EMD IMF or BLIMF structures. Then it is possible to observe that the GLRT based on the basic SM1 performs poorly on the original data segments. It appears to detect Parkinson's disease when there is no Parkinson's disease since the left panels refer to the segments from healthy patients and show a GLRT band with more black tests detected in the healthy patients rather than in the sick ones. This suggests that SM1 will not perform well for the given task, which is expected given that the original signal is highly non-stationary and, therefore, challenging to model with a simple covariance function for the entire signal.

If we next consider the results for the EMD-GP model using standard IMFs. Looking at the GLRT tests in Figure 7, the first two IMFs do not detect Parkinson's disease more efficiently than the raw data. This is the case since, quite often, $\gamma_1(t)$ and $\gamma_2(t)$ capture high noise levels and, therefore, are not great candidates for performing accurate inference on disease state in the patient. Regarding IMF3, the mini-batches correctly detected increase, suggesting that the fundamental frequency of male voices is a good discriminant for Parkinson's disease detection. Such facts will be reflected in the classification results provided in Tables 3 and 4.

Next, we consider the EMD-GP model using the BLIMFS. The GLRT tests of the BLIMFs perform quite differently from all the others. Particularly, the first and the third BLIMFs show perfect performances since every mini-batch (except for only two of them in $\gamma_1^{(\mathrm{BL})}(t)$) is classified correctly. Furthermore, the second BLIMF performs less effectively, suggesting that the noise affecting the formants structure can be isolated for a more discriminant decomposition. This is highly encouraging for the newly defined basis functions and will be further analysed in the discussion sections.

8 Discussion and Conclusions

We start by focusing on the female accuracy scores provided in Table 3. Across the benchmark features, the MFCCs combined with the MGDCCs were more reliable than using the individual sets of MFCC or MGDCCs separately. The combined benchmarks of MFCC+MGDCCs represent the standard to beat using the EMD-GP methods. These benchmark results produced an accuracy result around 70%. This is the case in both the averaged and non-averaged coefficients settings, suggesting that the technique undertaken in [20] provides an effective solution since saving part of the computational cost required for an SVM using all the coefficients. Equivalent results are achieved in Table 4 in the male case, showing the maximum accuracy result of 75%. The main issues encountered with these features include the following challenges. Firstly, there is a requirement for stationarity of the underlying signal, which is rarely respected, especially when the speech signal is not recorded in an ideal noise-free environment. In standard medical settings, there is significant background noise, there are non-ideal microphones used in phones or mobile devices.

Secondly, in the case of averaging the coefficients, most of the discriminant power carried by the frames describing the individual biometric formant structures will be polluted with the average operation. The final objective is indeed identifying which time-frequency regions, by gender, can discriminate ataxic speech. This is a delicate exercise per se, which should always take into account these observations and carefully consider the possibility of contamination of the classifier when reduction of complexity is in favour of the employed method. Furthermore, when it comes to health diagnostic, an accuracy score of 70% will not be considered since it is highly risky and therefore more powerful solutions need to be considered.

Next, it was demonstrated that when using a GLRT test, fitting a GP model directly to the speech signal is ineffective since the covariance function (GP kernel) function is not sufficiently flexible to capture the structure required to discriminate ataxic speech features. This is true even with the data-adaptive Fisher kernel structure; it does not provide any significant results in both sets of analysis, i.e. for males or females. Indeed, the formant behaviour of the

underlying signals carries a very complex structure affected by fast changes, which are not only due to the presence or absence of ataxic speech.

Therefore, such time-frequency fast variant modes require a refined modelling methodology, which, in this work, is represented by the stochastic embedding of the IMFs and the BLIMFs under the EMD-GP structures proposed. The next step is indeed to consider SM2 with the first three IMFs. These bases still do not show acceptable performances. It is often the case that the IMFs capture most of the data non-stationarity; therefore, their power in modelling fast changes may be reduced. However, by applying a median filter to $\gamma_1^s(t), \gamma_2^s(t), \gamma_3^s(t)$, better performances are obtained in this robust version. In the female case, the maximum achieved accuracy corresponds to 78%, while in the male case to 74%. What is important to notice at this stage is that in the former case, most of the discriminant power lies in the highest IMFs, i.e. $\gamma_1(t)$ and $\gamma_2(t)$, since females tend to have higher formants which are detected by higher frequency content of the EMD decomposition. In the male case, the third IMF shows more patients with the highest accuracy levels. Indeed, male voices tend to have formants at lower frequencies detected by $\gamma_3^s(t)$. This is particularly meaningful since it reflects the standard formants structure of female and male voices in general and provides useful interpretation to further develop such a modelling idea.

The best performing model came from the EMD-GP model structure based on using the first three BLIMFs defined previously and denoted by SM3. This outperformed all benchmarks and all other competitor models. The CEM has been applied to the first three IFs with configuration explained at the beginning of this section 7.4. The performances of this system model are outstanding compared to any other model. In the female case, $\gamma_2^{(\mathrm{BL})}(t)$ achieves levels of accuracy greater than 90% for any patient with any UPDRS score. $\gamma_3^{(\mathrm{BL})}(t)$ also provides high performances always greater than 88%, while $\gamma_1^{(\mathrm{BL})}(t)$ achieves accuracy scores of 73% at least. With the use of the CEM, the discriminatory power is shifted towards the second and third BLIMFs, rather than in IMF1 and IMF2, with significant performance gains achieved. This shows that the CEM can isolate more stationary basis functions characterised by the same frequency content and provide more powerful discrimination. As for the female case, all the BLIMFs for all patients in the male case provide high accuracy score levels. Highest performances are given by the third BLIMF, which achieves 90% for almost every patient.

While in the female case, the second BLIMF shows the best performances, in this case is $\gamma_3^{(\mathrm{BL})}(t)$ that carries most of the discriminatory power. This again reflects how males have lower formants than females and therefore detected by the third BLIMF. The second and the first BLIMFs well perform and provide high levels of accuracy. Furthermore, with the increase of the UPDRS score and hence the Parkinson's stage, the accuracy increases across all the three basis functions, which suggests the BLIMFs well detect the progression of the disease.

Two significant contributions were provided in this manuscript. The first was methodological in nature. We developed a novel technique for the stochastic embedding of the Empirical Mode Decomposition. This is lacking in the literature and introduces the definition of stochastic EMD by allowing for more powerful solutions in terms of classification or forecasting models that are based on non-stationary signal decomposition methods. As highlighted, two different stochastic EMD-GP embeddings have been presented. The first directly utilises the original IMFs in a GP compositional structure, while, the second relies on an optimal cross entropy based procedure used to define band-limited IMFs (BLIMFs), which produce distributions more consistent with stationarity properties, making the fitting of GP models in the EMD-GP based BLIMF stochastic embedding more reliable than that obtained using only the original EMD IMFs. The selection of the optimal partitions to characterise the BLIMFs utilised a novel use of the cross entropy method based on importance sampling distribution to derive the optimal time-frequency partition employed for defining the BLIMFs.

The second significant contribution produced was an essential demonstration of the utility of the stochastic embedding models for the EMD-GP frameworks, using both IMFs and BLIMFs. It was shown that the stochastic EMD-GP embedding structures could be used in a GLRT based inference testing procedure for speech signals to detect ataxic speech features. This is a critical task to solve when detecting the possibility of Parkinson's disease in patients from those who do not display standard ataxic speech features. It was demonstrated that the utilisation of the BLIMFs and GP stochastic embedding structures produced accuracies for detection of ataxic speech in Parkinson's patients with far greater accuracy than current state-of-the-art methods using SVMs and also out-performed standard GP models that did not utilise the EMD frameworks. This has been the case even when the adopted state-of-the-art kernel designs are based on a generative embedding framework for time-series kernels based on Fisher kernels. We believe that the proposed EMD-GP frameworks hold great potential for the development of other speech disorder analysis and detection of symptoms consistent with different neurological disorders, especially accurately when utilised in real-world recording environments using mobile phones in open doctors' office environments or hospitals, where background noises can be significant. We demonstrated that even in such recording environments, it was still possible to perform

diagnostics of ataxic speech accurately. This shows a significant improvement over the current state-of-the-art methods we implemented compared to the real data case study.

Female Results - Accuracy

Healthy Patients

	Benchmark (SVM)			Benchmark - not averaged (SVM)			SM1 (GLRT)	SM2 (GLRT)			SM2 (GLRT)			SM3 (GLRT)		
UPDRS	MFCCs	MGDCCs	MFCCs + MGDCCs	MFCCs	MGDCCs	MFCCs + MGDCCs	$\tilde{s}(t)$	$\gamma_1(t)$	$\gamma_2(t)$	$\gamma_3(t)$	$\gamma_1^s(t)$	$\gamma_2^s(t)$	$\gamma_3^s(t)$	$\gamma_1^{(BL)}(t)$	$\gamma_2^{(BL)}(t)$	$\gamma_3^{(BL)}(t)$
NaN	0.125	0.456	0.500	0.220	0.340	0.510	0.427	0.503	0.485	0.505	0.490	0.560	0.345	0.250	0.056	0.091
NaN	0.221	0.556	0.519	0.410	0.554	0.589	0.440	0.493	0.493	0.494	0.510	0.560	0.610	0.260	0.082	0.139
NaN	0.345	0.665	0.456	0.459	0.311	0.601	0.427	0.490	0.495	0.492	0.501	0.492	0.345	0.299	0.147	0.188
NaN	0.434	0.590	0.435	0.310	0.440	0.489	0.430	0.475	0.497	0.501	0.510	0.310	0.444	0.280	0.093	0.115
NaN	0.367	0.542	0.567	0.398	0.210	0.499	0.411	0.484	0.490	0.486	0.601	0.590	0.518	0.289	0.074	0.076
NaN	0.554	0.453	0.521	0.519	0.558	0.559	0.445	0.483	0.478	0.495	0.345	0.401	0.528	0.253	0.075	0.127
NaN	0.557	0.433	0.567	0.489	0.490	0.601	0.430	0.482	0.504	0.491	0.450	0.411	0.338	0.275	0.089	0.126
NaN	0.515	0.662	0.601	0.550	0.501	0.545	0.414	0.470	0.488	0.504	0.500	0.341	0.558	0.283	0.070	0.125
NaN	0.500	0.345	0.451	0.509	0.567	0.558	0.415	0.513	0.503	0.505	0.510	0.469	0.435	0.301	0.093	0.121
NaN	0.450	0.678	0.401	0.449	0.519	0.589	0.427	0.484	0.482	0.491	0.439	0.543	0.445	0.248	0.051	0.098
NaN	0.650	0.546	0.510	0.551	0.591	0.432	0.453	0.470	0.491	0.501	0.365	0.445	0.556	0.295	0.077	0.123
NaN	0.610	0.634	0.555	0.451	0.553	0.650	0.421	0.488	0.506	0.517	0.325	0.590	0.589	0.288	0.065	0.095
NaN	0.565	0.690	0.501	0.611	0.601	0.678	0.431	0.469	0.465	0.487	0.549	0.515	0.595	0.282	0.088	0.127
NaN	0.656	0.694	0.645	0.611	0.667	0.641	0.451	0.470	0.498	0.508	0.456	0.601	0.598	0.275	0.093	0.112
NaN	0.311	0.601	0.649	0.456	0.489	0.601	0.442	0.478	0.506	0.483	0.567	0.551	0.510	0.243	0.039	0.058
NaN	0.454	0.550	0.559	0.501	0.551	0.573	0.444	0.471	0.501	0.508	0.434	0.412	0.557	0.301	0.066	0.096
NaN	0.369	0.500	0.590	0.389	0.378	0.456	0.450	0.507	0.490	0.482	0.552	0.587	0.432	0.247	0.074	0.076
NaN	0.328	0.564	0.611	0.456	0.588	0.592	0.429	0.458	0.486	0.504	0.531	0.456	0.564	0.282	0.080	0.094
NaN	0.500	0.445	0.590	0.568	0.588	0.645	0.439	0.509	0.487	0.505	0.598	0.539	0.520	0.345	0.030	0.123

Sick Patients

	Benchmark (SVM)			Benchmark - not averaged (SVM)			SM1 (GLRT)	SM2 (GLRT)			SM2 (GLRT)			SM3 (GLRT)		
UPDRS	MFCCs	MGDCCs	MFCCs + MGDCCs	MFCCs	MGDCCs	MFCCs + MGDCCs	$\tilde{s}(t)$	$\gamma_1(t)$	$\gamma_2(t)$	$\gamma_3(t)$	$\gamma_1^s(t)$	$\gamma_2^s(t)$	$\gamma_3^s(t)$	$\gamma_1^{(BL)}(t)$	$\gamma_2^{(BL)}(t)$	$\gamma_3^{(BL)}(t)$
0	0.256	0.570	0.690	0.358	0.590	0.699	0.557	0.533	0.508	0.510	**0.701**	**0.705**	0.690	**0.736**	**0.995**	**0.895**
0	0.543	0.601	**0.701**	0.555	0.619	**0.707**	0.497	0.527	0.500	0.513	0.652	0.690	**0.711**	**0.811**	**0.959**	**0.888**
1	0.556	0.611	**0.711**	0.501	0.640	0.699	0.558	0.535	0.482	0.507	**0.710**	**0.701**	**0.700**	**0.710**	**0.935**	**0.882**
1	0.343	0.575	**0.702**	0.410	0.595	**0.710**	0.573	0.520	0.510	0.491	**0.783**	**0.711**	0.601	**0.790**	**0.950**	**0.899**

Table 3: Accuracy performance results of the female patients. Remark that the accuracy is computed as $\frac{TP+TN}{TP+TN+FP+FN}$. The columns show the UPDRS score (marked as NaN in the case of healthy patients), the benchmark measures corresponding to the MFCCs and the MGDCCs and their perfomance together run with an SVM, the results for SM1, SM2 and SM3.

Male Results - Accuracy

Healthy Patients

	Benchmark (SVM)			Benchmark - not averaged (SVM)			SM1 (GLRT)	SM2 (GLRT)			SM2 (GLRT)			SM3 (GLRT)		
UPDRS	MFCCs	MGDCCs	MFCCs + MGDCCs	MFCCs	MGDCCs	MFCCs + MGDCCs	$\tilde{s}(t)$	$\gamma_1(t)$	$\gamma_2(t)$	$\gamma_3(t)$	$\gamma_1^s(t)$	$\gamma_2^s(t)$	$\gamma_3^s(t)$	$\gamma_1^{(BL)}(t)$	$\gamma_2^{(BL)}(t)$	$\gamma_3^{(BL)}(t)$
NaN	0.410	0.515	0.519	0.500	0.511	0.558	0.379	0.513	0.445	0.463	0.500	0.567	0.490	0.225	0.235	0.059
NaN	0.643	0.590	0.571	0.519	0.576	0.598	0.379	0.488	0.449	0.462	0.531	0.450	0.441	0.225	0.250	0.026

Sick Patients

	Benchmark (SVM)			Benchmark - not averaged (SVM)			SM1 (GLRT)	SM2 (GLRT)			SM2 (GLRT)			SM3 (GLRT)		
UPDRS	MFCCs	MGDCCs	MFCCs + MGDCCs	MFCCs	MGDCCs	MFCCs + MGDCCs	$\tilde{s}(t)$	$\gamma_1(t)$	$\gamma_2(t)$	$\gamma_3(t)$	$\gamma_1^s(t)$	$\gamma_2^s(t)$	$\gamma_3^s(t)$	$\gamma_1^{(BL)}(t)$	$\gamma_2^{(BL)}(t)$	$\gamma_3^{(BL)}(t)$
0	0.520	0.650	0.611	0.551	0.656	0.678	0.627	0.499	0.528	0.521	0.690	0.611	**0.710**	**0.787**	**0.727**	**0.911**
0	0.555	0.600	0.619	0.458	0.623	0.674	0.602	0.493	0.537	0.534	0.601	0.699	**0.722**	**0.865**	**0.729**	**0.911**
0	0.390	0.588	0.690	0.553	0.598	0.593	0.597	0.502	0.527	0.549	0.610	**0.729**	**0.730**	**0.764**	**0.741**	**0.920**
0	0.430	0.590	0.699	0.441	0.489	0.563	0.635	0.480	0.522	0.523	0.673	**0.719**	**0.710**	**0.729**	**0.724**	**0.878**
0	0.551	0.500	0.652	0.428	0.649	0.693	0.610	0.485	0.548	0.549	**0.715**	0.690	**0.721**	**0.763**	**0.722**	**0.916**
1	0.439	0.595	**0.702**	0.469	0.532	0.515	0.615	0.496	0.551	0.522	**0.700**	**0.711**	**0.735**	**0.821**	**0.764**	**0.954**
1	0.312	0.610	**0.712**	0.654	0.689	**0.709**	0.626	0.502	0.548	0.545	**0.709**	**0.705**	**0.721**	**0.845**	**0.762**	**0.947**
1	0.235	0.645	**0.705**	0.613	0.601	**0.731**	0.616	0.505	0.546	0.575	**0.711**	0.610	**0.721**	**0.745**	0.690	**0.835**
1	0.611	0.650	0.675	0.689	0.673	0.678	0.595	0.510	0.534	0.534	0.687	**0.722**	**0.720**	**0.780**	**0.731**	**0.926**
2	0.387	0.611	**0.718**	0.445	0.562	0.699	0.628	0.485	0.505	0.533	**0.733**	**0.741**	**0.745**	**0.881**	**0.760**	**0.923**
2	0.654	0.674	**0.731**	0.510	0.661	**0.722**	0.581	0.492	0.543	0.550	**0.721**	**0.730**	**0.727**	**0.888**	**0.899**	**0.910**
3	0.442	0.659	**0.750**	0.567	0.698	**0.719**	0.634	0.489	0.537	0.638	**0.720**	**0.711**	**0.749**	**0.899**	**0.950**	**0.949**

Table 4: Accuracy performance results of the female patients. Remark that the accuracy is computed as $\frac{TP+TN}{TP+TN+FP+FN}$. The columns show the UPDRS score (marked as NaN in the case of healthy patients), the benchmark measures corresponding to the MFCCs and the MGDCCs and their perfomance together run with an SVM, the results for SM1, SM2 and SM3.

References

[1] Joomee Song, Ju Hwan Lee, Jungeun Choi, Mee Kyung Suh, Myung Jin Chung, Young Hun Kim, Jeongho Park, Seung Ho Choo, Ji Hyun Son, Dong Yeong Lee, et al. Detection and differentiation of ataxic and hypokinetic dysarthria in cerebellar ataxia and parkinsonian disorders via wave splitting and integrating neural networks. *PloS one*, 17(6):e0268337, 2022.

[2] Mobile device voice recordings at king's college london(mdvr-kcl) from both early and advanced parkinson's disease patients and healthy controls, 2019.

[3] Sabine Theis, Dajana Schäfer, Christina Haubrich, Christopher Brandl, Matthias Wille, Sonja A Kotz, Verena Nitsch, and Alexander Mertens. Perceived self-efficacy in parkinson's disease through mobile health monitoring. In *International Conference on Human-Computer Interaction*, pages 749–762. Springer, 2020.

[4] Frédéric Puyjarinet, Valentin Bégel, Christian Gény, Valérie Driss, Marie-Charlotte Cuartero, Sonja A Kotz, Serge Pinto, and Simone Dalla Bella. Heightened orofacial, manual, and gait variability in parkinson's disease results from a general rhythmic impairment. *npj Parkinson's Disease*, 5(1):1–7, 2019.

[5] Ian McLoughlin. *Applied speech and audio processing: with Matlab examples*. Cambridge University Press, 2009.

[6] Ian Vince McLoughlin. *Speech and Audio Processing: a MATLAB-based approach*. Cambridge University Press, 2016.

[7] Defne Abur, Rosemary A Lester-Smith, Ayoub Daliri, Ashling A Lupiani, Frank H Guenther, and Cara E Stepp. Sensorimotor adaptation of voice fundamental frequency in parkinson's disease. *PLoS One*, 13(1):e0191839, 2018.

[8] Fabiola Staróbole Juste, Fernanda Chiarion Sassi, Julia Biancalana Costa, and Claudia Regina Furquim de Andrade. Frequency of speech disruptions in parkinson's disease and developmental stuttering: A comparison among speech tasks. *Plos one*, 13(6):e0199054, 2018.

[9] Nemuel D Pah, Mohammod A Motin, Peter Kempster, and Dinesh K Kumar. Detecting effect of levodopa in parkinson's disease patients using sustained phonemes. *IEEE Journal of Translational Engineering in Health and Medicine*, 9:1–9, 2021.

[10] Christos Laganas, Dimitrios Iakovakis, Stelios Hadjidimitriou, Vasileios Charisis, Sofia B Dias, Sevasti Bostantzopoulou, Zoe Katsarou, Lisa Klingelhoefer, Heinz Reichmann, Dhaval Trivedi, et al. Parkinson's disease detection based on running speech data from phone calls. *IEEE Transactions on Biomedical Engineering*, 69(5):1573–1584, 2021.

[11] Sukhpal Kaur, Himanshu Aggarwal, and Rinkle Rani. Diagnosis of parkinson's disease using principle component analysis and deep learning. *Journal of Medical Imaging and Health Informatics*, 9(3):602–609, 2019.

[12] NP Narendra, Björn Schuller, and Paavo Alku. The detection of parkinson's disease from speech using voice source information. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1925–1936, 2021.

[13] Athanasios Tsanas, Max A Little, and Lorraine O Ramig. Remote assessment of parkinson's disease symptom severity using the simulated cellular mobile telephone network. *Ieee Access*, 9:11024–11036, 2021.

[14] Laiba Zahid, Muazzam Maqsood, Mehr Yahya Durrani, Maheen Bakhtyar, Junaid Baber, Habibullah Jamal, Irfan Mehmood, and Oh-Young Song. A spectrogram-based deep feature assisted computer-aided diagnostic system for parkinson's disease. *IEEE Access*, 8:35482–35495, 2020.

[15] Laureano Moro-Velazquez, Jesus Villalba, and Najim Dehak. Using x-vectors to automatically detect parkinson's disease from speech. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1155–1159. IEEE, 2020.

[16] Marta Campi, Gareth W Peters, Nourddine Azzaoui, and Tomoko Matsui. Machine learning mitigants for speech based cyber risk. *IEEE Access*, 9:136831–136860, 2021.

[17] Betul Erdogdu Sakar, M Erdem Isenkul, C Okan Sakar, Ahmet Sertbas, Fikret Gurgen, Sakir Delil, Hulya Apaydin, and Olcay Kursun. Collection and analysis of a parkinson speech dataset with multiple types of sound recordings. *IEEE Journal of Biomedical and Health Informatics*, 17(4):828–834, 2013.

[18] María Luisa Barragán Pulido, Jesús Bernardino Alonso Hernández, Miguel Ángel Ferrer Ballester, Carlos Manuel Travieso González, Jiří Mekyska, and Zdeněk Smékal. Alzheimer's disease and automatic speech analysis: a review. *Expert systems with applications*, 150:113213, 2020.

[19] Nicholas Cummins, Alice Baird, and Bjoern W Schuller. Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods*, 151:41–54, 2018.

[20] Bipasha Kashyap, Pubudu N Pathirana, Malcolm Horne, Laura Power, and David Szmulewicz. Quantitative assessment of speech in cerebellar ataxia using magnitude and phase based cepstrum. *Annals of biomedical engineering*, 48(4):1322–1336, 2020.

[21] Hermann Ackermann and Ingo Hertrich. Speech rate and rhythm in cerebellar dysarthria: An acoustic analysis of syllabic timing. *Folia phoniatrica et logopaedica*, 46(2):70–78, 1994.

[22] Bettina Brendel, Matthis Synofzik, Hermann Ackermann, Tobias Lindig, Theresa Schölderle, Ludger Schöls, and Wolfram Ziegler. Comparing speech characteristics in spinocerebellar ataxias type 3 and type 6 with friedreich ataxia. *Journal of neurology*, 262(1):21–26, 2015.

[23] Ray D Kent, Jane Finley Kent, Joe R Duffy, Jack E Thomas, Gary Weismer, and Sarah Stuntebeck. Ataxic dysarthria. *Journal of Speech, Language, and Hearing Research*, 43(5):1275–1289, 2000.

[24] Aileen K Ho, Robert Iansek, Caterina Marigliani, John L Bradshaw, and Sandra Gates. Speech impairment in a large sample of patients with parkinson's disease. *Behavioural neurology*, 11(3):131–137, 1998.

[25] Anthony E Lang and Andres M Lozano. Parkinson's disease. *New England Journal of Medicine*, 339(16):1130–1143, 1998.

[26] Anna Pompili, Rubén Solera-Urena, Alberto Abad, Rita Cardoso, Isabel Guimaraes, Margherita Fabbri, Isabel P Martins, and Joaquim Ferreira. Assessment of parkinson's disease medication state through automatic speech analysis. *arXiv preprint arXiv:2005.14647*, 2020.

[27] Tobias Bocklet, Stefan Steidl, Elmar Nöth, and Sabine Skodda. Automatic evaluation of parkinson's speech-acoustic, prosodic and voice related cues. In *Interspeech*, pages 1149–1153, 2013.

[28] Brian Harel, Michael Cannizzaro, and Peter J Snyder. Variability in fundamental frequency during speech in prodromal and incipient parkinson's disease: A longitudinal case study. *Brain and cognition*, 56(1):24–29, 2004.

[29] Lena Hartelius and Per Svensson. Speech and swallowing symptoms associated with parkinson's disease and multiple sclerosis: a survey. *Folia phoniatrica et logopaedica*, 46(1):9–17, 1994.

[30] Sabine Skodda, Heiko Rinsche, and Uwe Schlegel. Progression of dysprosody in parkinson's disease over time—a longitudinal study. *Movement disorders: official journal of the Movement Disorder Society*, 24(5):716–722, 2009.

[31] Shimon Sapir, A Pawlas, L Ramig, S Countryman, C O'BRIEN, M Hoehn, and LA Thompson. Speech and voice abnormalities in parkinson disease: relation to severity of motor impairment, duration of disease, medication, depression, gender and age. *NCVS Status and Progress Report*, 14:149–161, 1999.

[32] Jeri A Logemann, Hilda B Fisher, Benjamin Boshes, and E Richard Blonsky. Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of parkinson patients. *Journal of Speech and hearing Disorders*, 43(1):47–57, 1978.

[33] Neha Singh, Viness Pillay, and Yahya E Choonara. Advances in the treatment of parkinson's disease. *Progress in neurobiology*, 81(1):29–44, 2007.

[34] Athanasios Tsanas, Max Little, Patrick McSharry, and Lorraine Ramig. Accurate telemonitoring of parkinson's disease progression by non-invasive speech tests. *Nature Precedings*, pages 1–1, 2009.

[35] Leon Cohen. *Time-frequency analysis*, volume 778. Prentice hall New Jersey, 1995.

[36] Shie Qian and Dapang Chen. *Joint time-frequency analysis: methods and applications*. Prentice-Hall, Inc., 1996.

[37] Norden E Huang, Zheng Shen, Steven R Long, Manli C Wu, Hsing H Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H Liu. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 454(1971):903–995, 1998.

[38] T Adrián de Pérez, J Restrepo, and LM Díaz. Optimum time-frequency representations of monocomponent signal combinations. *Signal processing*, 38(2):187–195, 1994.

[39] Boualem Boashash. Estimating and interpreting the instantaneous frequency of a signal. i. fundamentals. *Proceedings of the IEEE*, 80(4):520–538, 1992.

[40] Boualem Boashash and Graeme Jones. Instantaneous frequency and time-frequency distributions. Longman Cheshire, 1992.

[41] Boualem Boashash. *Time-frequency signal analysis and processing: a comprehensive reference*. Academic Press, 2015.

[42] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[43] Grace Wahba. *Spline models for observational data*. SIAM, 1990.

[44] Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268, 2011.

[45] Francis Bach. Exploring large feature spaces with hierarchical multiple kernel learning. *arXiv preprint arXiv:0809.1493*, 2008.

[46] Pratik Jawanpuria, Jagarlapudi Saketha Nath, and Ganesh Ramakrishnan. Generalized hierarchical kernel learning. *Journal of Machine Learning Research*, 16(20):617–652, 2015.

[47] Felipe Tobar, Thang D Bui, and Richard E Turner. Learning stationary time series using gaussian processes with nonparametric kernels. *Advances in Neural Information Processing Systems*, 28:3501–3509, 2015.

[48] Miguel Lázaro-Gredilla, Joaquin Quiñonero-Candela, Carl Edward Rasmussen, and Aníbal R Figueiras-Vidal. Sparse spectrum gaussian process regression. *The Journal of Machine Learning Research*, 11:1865–1881, 2010.

[49] Tommi S Jaakkola, David Haussler, et al. Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*, pages 487–493, 1999.

[50] Tommi S Jaakkola, Mark Diekhans, and David Haussler. Using the fisher kernel method to detect remote protein homologies. In *ISMB*, volume 99, pages 149–158, 1999.

[51] Pedro J Moreno and Ryan Rifkin. Using the fisher kernel method for web audio classification. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 4, pages 2417–2420. IEEE, 2000.

[52] Nathan Smith and Mahesan Niranjan. Data-dependent kernels in svm classification of speech patterns. In *Sixth International Conference on Spoken Language Processing*, 2000.

[53] Dirk P Kroese, Reuven Y Rubinstein, Izack Cohen, Sergey Porotsky, and Thomas Taimre. Cross-entropy method'. *European Journal of Operational Research*, 31:276–283, 2011.

[54] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005.

[55] Eric Deléchelle, Jacques Lemoine, and Oumar Niang. Empirical mode decomposition: an analytical approach for sifting process. *IEEE Signal Processing Letters*, 12(11):764–767, 2005.

[56] Mina B. Abd el Malek and Samer S. Hanna. The hilbert transform of cubic splines. *Communications in Nonlinear Science and Numerical Simulation*, 80:104983, 2020.

[57] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[58] Saburou Saitoh. Theory of reproducing kernels and its applications. *Longman Scientific & Technical*, 1988.

[59] Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[60] Andreas Argyriou, Charles A Micchelli, and Massimiliano Pontil. When is there a representer theorem? vector versus matrix regularizers. *The Journal of Machine Learning Research*, 10:2507–2529, 2009.

[61] George Kimeldorf and Grace Wahba. Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971.

[62] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International conference on artificial neural networks*, pages 583–588. Springer, 1997.

[63] Carl Edward Rasmussen. Gaussian processes to speed up hybrid monte carlo for expensive bayesian integrals. In *Seventh Valencia international meeting, dedicated to Dennis V. Lindley*, pages 651–659. Oxford University Press, 2003.

[64] Grace Wahba. Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(3):364–372, 1978.

[65] Jaakko Riihimäki and Aki Vehtari. Gaussian processes with monotonicity information. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 645–652. JMLR Workshop and Conference Proceedings, 2010.

[66] Johannes S Maritz and T Lwin. *Empirical bayes methods*. Chapman and Hall/CRC, 2018.

[67] Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007.

[68] Stefan Rüping. Svm kernels for time series analysis. Technical report, Technical report, 2001.

[69] Shai Fine, Jiri Navratil, and Ramesh A Gopinath. A hybrid gmm/svm approach to speaker identification. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, volume 1, pages 417–420. IEEE, 2001.

[70] N. Smith and M. Gales. Speech recognition using svms. In *NIPS*, 2001.

[71] Reuven Y Rubinstein and Dirk P Kroese. *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning*, volume 133. Springer, 2004.

[72] Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease. The unified parkinson's disease rating scale (updrs): status and recommendations. *Movement Disorders*, 18(7):738–750, 2003.

[73] P1 Martínez-Martín, A Gil-Nagel, L Morlán Gracia, J Balseiro Gómez, J Martinez-Sarries, F Bermejo, and Cooperative Multicentric Group. Unified parkinson's disease rating scale characteristics and structure. *Movement disorders*, 9(1):76–83, 1994.

[74] N. Zheng, T. Lee, and P. C. Ching. Integration of complementary acoustic features for speaker recognition. *IEEE Signal Processing Letters*, 14(3):181–184, 2007.

[75] Robert McAulay and Thomas Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(4):744–754, 1986.

[76] Rajib Sharma, Leandro Vignolo, Gastón Schlotthauer, Marcelo A Colominas, H Leonardo Rufiner, and SRM Prasanna. Empirical mode decomposition for adaptive am-fm analysis of speech: A review. *Speech Communication*, 88:39–64, 2017.

[77] T Ananthapadmanabha and B Yegnanarayana. Epoch extraction from linear prediction residual. In *ICASSP'78. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 8–11. IEEE, 1978.

[78] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

[79] Stephen Jannetts and Anja Lowit. Cepstral analysis of hypokinetic and ataxic voices: correlations with perceptual and other acoustic measures. *Journal of Voice*, 28(6):673–680, 2014.

[80] Sophia Luna-Webb. Comparison of acoustic measures in discriminating between those with friedreich's ataxia and neurologically normal peers. 2015.

[81] Mikko-Ville Laitinen, Sascha Disch, and Ville Pulkki. Sensitivity of human hearing to changes in phase spectrum. *Journal of the Audio Engineering Society*, 61(11):860–877, 2013.

[82] Kuldip K Paliwal and Leigh Alsteris. Usefulness of phase spectrum in human speech perception. In *Eighth European Conference on Speech Communication and Technology*, 2003.

[83] Manfred R Schroeder. New results concerning monaural phase sensitivity. *The Journal of the Acoustical Society of America*, 31(11):1579–1579, 1959.

[84] Rajesh M Hegde, Hema A Murthy, and Venkata Ramana Rao Gadde. Significance of the modified group delay feature in speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):190–202, 2007.

[85] R Frail, JI Godino-Llorente, N Saenz-Lechon, Vıctor Osma-Ruiz, and Corinne Fredouille. Mfcc-based remote pathology detection on speech transmitted through the telephone channel. *Proc Biosignals*, 2009.

[86] CM Vikram and K Umarani. Pathological voice analysis to detect neurological disorders using mfcc and svm. *Int. J. Adv. Electr. Electron. Eng*, 2(4):87–91, 2013.

[87] Dayu Huang, Jayakrishnan Unnikrishnan, Sean Meyn, Venugopal Veeravalli, and Amit Surana. Statistical svms for robust detection, supervised learning, and universal classification. In *2009 IEEE Information Theory Workshop on Networking and Information Theory*, pages 62–66. IEEE, 2009.

[88] Higini Arau-Puchades and Umberto Berardi. The reverberation radius in an enclosure with asymmetrical absorption distribution. In *Proceedings of Meetings on Acoustics ICA2013*, volume 19, page 015141. Acoustical Society of America, 2013.

A KL Divergence formulation

This corresponds to the derivation of KL divergence in Eqn. (15):

$$
\begin{aligned}
KL(\hat{\pi}, \pi; \psi) &= \sum_{m=1}^{M}\sum_{d=1}^{d} \pi\left(x=(m,d)\right) \log\left(\frac{\pi\left(x=(m,d)\right)}{\hat{\pi}\left(x=(m,d)\right)}\right) \\
&= \sum_{m=1}^{M}\sum_{d=1}^{d} \Big\{ \pi\left(x=(m,d)\right)\Big(\log \pi\left(x=(m,d)\right) - \log \hat{\pi}\left(x=(m,d)\right)\Big)\Big\} = \sum_{m=1}^{M}\sum_{d=1}^{d} \Big\{ \frac{|\Pi_{m,d}|}{|\Pi|}\Big(\log \frac{|\Pi_{m,d}|}{|\Pi|} - \log \frac{|\mathcal{P}_{m,d}|}{KN}\Big)\Big\} \\
&= \frac{1}{|\Pi|}\sum_{m=1}^{M}\sum_{d=1}^{d} \Big\{ |\Pi_{m,d}|\Big(\log|\Pi_{m,d}| - \log|\Pi| - \log|\mathcal{P}_{m,d}| + \log KN\Big)\Big\} \\
&= \frac{1}{|\Pi|}\sum_{m=1}^{M}\sum_{d=1}^{d} \Big\{ |\Pi_{m,d}|\Big(\log|\Pi_{m,d}| - \log|\mathcal{P}_{m,d}|\Big)\Big\} + \frac{1}{|\Pi|}\sum_{m=1}^{M}\sum_{d=1}^{d} \Big\{ |\Pi_{m,d}|\Big(\log KN - \log|\Pi|\Big)\Big\} \\
&= \frac{1}{|\Pi|}\sum_{m=1}^{M}\sum_{d=1}^{d} \Big\{ |\Pi_{m,d}|\Big(\log|\Pi_{m,d}| - \log|\mathcal{P}_{m,d}|\Big)\Big\} + \Big(\log KN - \log|\Pi|\Big)\frac{1}{|\Pi|}\sum_{m=1}^{M}\sum_{d=1}^{d} |\Pi_{m,d}| \\
&= \log KN - \log|\Pi| + \frac{1}{|\Pi|}\sum_{m=1}^{M}\sum_{d=1}^{d} \Big\{ |\Pi_{m,d}|\Big(\log|\Pi_{m,d}| - \log|\mathcal{P}_{m,d}|\Big)\Big\}.
\end{aligned}
$$

B Kernel Density Estimator for KL Divergence formulation

The KL divergence considering the kernel density estimator provided in subsection 5.2.2 whose solution is given in Eqn. 17 is derived in this section

$$
\begin{aligned}
KL(\hat{\pi}^e, \pi; \psi) &= \int_{x\in\mathcal{X}} \pi(x) \log\left(\frac{\pi(x)}{\hat{\pi}^e\left(x;k,h\right)}\right) dx = \sum_{m=1}^{M}\sum_{d=1}^{d} \pi\left(x=(m,d)\right) \log\left(\frac{\pi\left(x=(m,d)\right)}{\hat{\pi}^e(x=(m,d);k,h)}\right) \\
&= \sum_{m=1}^{M}\sum_{d=1}^{d} \Big\{ \frac{|\Pi_{m,d}|}{|\Pi|}\Big(\log|\Pi_{m,d}| - \log|\Pi| - \log\hat{\pi}^e(x=(m,d);k,h)\Big)\Big\} \qquad (23) \\
&= \frac{1}{|\Pi|}\sum_{m=1}^{M}\sum_{d=1}^{d} \Big\{ |\Pi_{m,d}|\Big(\log|\Pi_{m,d}| - \log\hat{\pi}^e(x=(m,d);k,h)\Big)\Big\} - \frac{\log|\Pi|}{|\Pi|}\sum_{m=1}^{M}\sum_{d=1}^{D} |\Pi_{m,d}| \qquad (24) \\
&= -\log|\Pi| + \frac{1}{|\Pi|}\sum_{m=1}^{M}\sum_{d=1}^{d} |\Pi_{m,d}|\Big(\log|\Pi_{m,d}| - \log\hat{\pi}^e(x=(m,d);k,h)\Big) \qquad (25)
\end{aligned}
$$

and with a numerical trick, for $C > 0$ being a very small number, ie $C = 10^{-100}$

$$
\begin{aligned}
KL(\hat{\pi}^e, \pi; \psi) &= \int_{x\in\mathcal{X}} \pi(x) \log\left(\frac{\pi(x)}{\hat{\pi}^e\left(x;k,h\right)}\right) dx = \sum_{m=1}^{M}\sum_{d=1}^{d} \pi\left(x=(m,d)\right) \log\left(\frac{\pi\left(x=(m,d)\right)}{\hat{\pi}^e(x=(m,d);k,h)}\right) \\
&= \sum_{m=1}^{M}\sum_{d=1}^{d} \Big\{ \frac{|\Pi_{m,d}|}{|\Pi|}\Big(\log|\Pi_{m,d}| - \log|\Pi| - \log C\frac{\hat{\pi}^e(x=(m,d);k,h)}{C}\Big)\Big\} \qquad (26) \\
&= \sum_{m=1}^{M}\sum_{d=1}^{d} \Big\{ \frac{|\Pi_{m,d}|}{|\Pi|}\Big(\log|\Pi_{m,d}| - \log|\Pi| - \log C - \log\frac{\hat{\pi}^e(x=(m,d);k,h)}{C}\Big)\Big\} \qquad (27) \\
&= \frac{1}{|\Pi|}\sum_{m=1}^{M}\sum_{d=1}^{d} \Big\{ |\Pi_{m,d}|\Big(\log|\Pi_{m,d}| - \log\frac{\hat{\pi}^e(x=(m,d);k,h)}{C}\Big)\Big\} - \frac{\log|\Pi| + \log C}{|\Pi|}\sum_{m=1}^{M}\sum_{d=1}^{D} |\Pi_{m,d}| \qquad (28)
\end{aligned}
$$

$$= -\log|\Pi| - \log C + \frac{1}{|\Pi|}\sum_{m=1}^{M}\sum_{d=1}^{d}|\Pi_{m,d}|\Big(\log|\Pi_{m,d}| - \log\frac{\hat{\pi}^{e}(x=(m,d);k,h)}{C}\Big) \tag{29}$$

C Cross Entropy Algorithm for Optimal Time-Frequency Partition Estimation

The following algorithm presents the random partition cross-entropy discrete optimisation solution.

Algorithm 1: Random Partition via CEM for Discrete Optimisation

Input: Set $M, D, S > 0,\ N_\omega \geq M,\ N_\tau \geq D$;
Input: Set hyperparameters: $\rho > 0$,$\beta > 0$,
Input: Set initial parameters $\mathbf{p}^{[0]}, \mathbf{p}_1^{,[0]}, \ldots, \mathbf{p}_M^{,[0]}$
for $i > 0$ **do**

1. Generate S sets of realisations $[\mathbf{X}^{(s)[i]}, \mathbf{X}_1^{,(s)[i]}, \ldots, \mathbf{X}_M^{,(s)[i]}], \mathbf{X}^{(s)[i]} \sim \pi\left((\mathbf{x}, \mathbf{p}^{[i]}\right),\ \mathbf{X}_m^{,(s)[i]} \sim \pi\left(\mathbf{x}'_m, \mathbf{p}^{,[i-1]}\right),$;
2. Calculate $\psi^{(s)[i]} = \left[\omega_1^{(s)[i]}, \ldots, \omega_{M-1}^{(s)[i]}, s_{1,1}^{(s)[i]}, \ldots, s_{M,D-1}^{(s)[i]}\right]\ \omega_m^{(s)[i]} = \omega_0 + \Delta_\omega \sum_{m'=1}^{m} x_{m'}^{(s)[i]},\ s_{m,d}^{(s)[i]} = t_0 + \Delta_\tau \sum_{d'=1}^{d} x_{m,d'}^{,(s)[i]};$
3. Calculate $KL(\hat{\pi}, \pi; \psi^{(s)[i]})$ for $s = 1, \ldots, S$ and specify $\gamma^{[i]}$ being $1-\rho$ empirical quantile of their values;
4. Calculate

$$\hat{p}_m = \frac{\sum_{s=1}^{S}\mathbf{1}_{\left\{KL(\hat{\pi},\pi;\psi^{(s)[i]})\leq\gamma^{[i]}\right\}}\frac{x_m^{(s)[i]}}{N_\omega}}{\sum_{s=1}^{S}\mathbf{1}_{\left\{KL(\hat{\pi},\pi;\psi^{(s)[i]})\leq\gamma^{[i]}\right\}}},\ \hat{p}'_{m,d} = \frac{\sum_{s=1}^{S}\mathbf{1}_{\left\{KL(\hat{\pi},\pi;\psi^{(s)[i]})\leq\gamma^{[i]}\right\}}\frac{x_{m,d}^{'(s)[i]}}{N_\tau}}{\sum_{s=1}^{S}\mathbf{1}_{\left\{KL(\hat{\pi},\pi;\psi^{(s)[i]})\leq\gamma^{[i]}\right\}}}$$

5. Smooth update of the parameters

$$p_m^{[i]} = \beta p_m^{[i-1]} + (1-\beta)\hat{p}_m,\ p_{m,d}^{,[i]} = \beta p_{m,d}^{,[i-1]} + (1-\beta)\hat{p}'_{m,d};$$

$i = i + 1$

until a convergence criterion is satisfied

After convergence, specify points of partition $\omega_m = \omega_0 + |\mathcal{I}|\sum_{m'=1}^{m} p_{m'}^{[i]},\ s_{m,d} = t_0 + |\mathcal{T}|\sum_{d'=1}^{d} p_{m,d'}^{,[i]};$.