# Discriminative Wavelet Packet Filter Bank Selection for Pattern Recognition

Jorge Silva, *Student Member, IEEE*, and Shrikanth S. Narayanan, *Fellow, IEEE*

***Abstract*—This paper addresses the problem of discriminative wavelet packet (WP) filter bank selection for pattern recognition. The problem is formulated as a complexity regularized optimization criterion, where the tree-indexed structure of the WP bases is explored to find conditions for reducing this criterion to a type of minimum cost tree pruning, a method well understood in regression and classification trees (CART). For estimating the conditional mutual information, adopted to compute the fidelity criterion of the minimum cost tree-pruning problem, a nonparametric approach based on product adaptive partitions is proposed, extending the Darbellay–Vajda data-dependent partition algorithm. Finally, experimental evaluation within an automatic speech recognition (ASR) task shows that proposed solutions for the WP decomposition problem are consistent with well understood empirically determined acoustic features, and the derived feature representations yield competitive performances with respect to standard feature extraction techniques.**

***Index Terms*—Automatic speech recognition, Bayes' decision approach, complexity regularization, data-dependent partitions, filter bank selection, minimum cost tree pruning, minimum probability of error signal representation, mutual information, mutual information estimation, tree-structured bases and wavelet packets (WPs).**

## I. Introduction

WAVELET PACKETs (WPs) and general multirate filter banks [1]–[3] have emerged as important signal representation schemes for compression, detection and classification [4]–[8]. This basis family is particularly appealing for the analysis of pseudo-stationary time series processes and quasi-periodic random fields, such as the acoustic speech signals, and texture image sources [9]–[11], where a filter bank analysis has shown to be suitable for decorrelating the process into its basic innovation components. In pattern recognition (PR), filter bank structures have been the basic signal analysis block for several acoustic and image classification tasks, notably including automatic speech recognition (ASR) and texture classification. In this domain, an interesting problem is to determine the optimal filter bank structure for a given classification task [10], [11], or the equivalent optimal basis selection (BS) problem [12], [13].

In pattern recognition (PR), the optimal signal representation problem can be associated with the feature extraction (FE). It is well known that if the joint class observation distribution is available, the Bayes' decision provides a means of minimizing the risk [14]. However, in practice, the joint distribution is typically not available, and in the Bayes' decision approach this distribution is estimated from a finite amount of training data [14]–[16]. It is also well understood that the accuracy of this estimation is affected by the dimensionality of the observation space—the curse of dimensionality [7], [14], [16]–[18]. Hence, an integral part of FE is to address the problem of optimal dimensionality reduction, particularly necessary in scenarios where the original raw-observation measurements lie in a high dimensional space, and a limited amount of training data is available, such as in most speech classification [19], image classification [5] and hyperspectral classification scenarios [20], [21].

Toward addressing of this problem, Vasconcelos [7] has formalized the minimum probability of error signal representation (MPE-SR) principle. Under certain conditions, this work presents a tradeoff between the quality of the signal space (or approximation error quantity) and an information theoretic indicator for the estimation error across a sequence of embedded feature representations of increasing dimensionality, and connects this result with the notion of optimal signal representation for PR. In [22], these results were extended to a more general theoretical setting, introducing the idea of family of consistent distributions associated with an embedded sequence of feature representations. Furthermore, [22] approximated the MPE-SR problem as solution of an operational cost-fidelity problem using mutual information (MI) as a discriminative fidelity criterion [23] and dimensionality as the cost term.

The focus of this study is to extend the MPE-SR formulation for the important family of filter bank feature representations induced by the wavelet packets (WPs) [1], [3]. The idea is to take advantage of the WP tree structure to characterize sufficient conditions that guarantee algorithmic solutions for the cost-fidelity problem. This approach was motivated by algorithmic solutions obtained for the case of tree-structured vector quantization (TSVQ) in lossy compression [6], [24] and TSVQ for nonparametric classification and regression problems [15], [25], [26].

Discriminative basis selection (BS) problems for tree-structured bases family have been proposed independently in [5] and [13]. Saito *et al.* [13], extending the idea of BS for signal

Manuscript received March 04, 2008; accepted December 08, 2008. First published January 23, 2009; current version published April 15, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gerald Schuller. This work was supported in part by grants from the Office of Naval Research, the Army, the National Science Foundation, and the Department of Homeland Security. The work of J. Silva was supported by funding from Fondecyt Grant 1090138, CONICYT-Chile.

J. Silva is with the Electrical Engineering Department, University of Chile, Santiago 412-3, Chile (e-mail: josilva@ing.uchile.cl).

S. S. Narayanan is with the Department of Electrical Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles 90089-2564 USA (e-mail: shri@sipi.usc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2009.2013898

representation in [12], proposed a fidelity criterion that measures interclass discrimination, in the Kullback–Leibler divergence (KLD) sense [27], considering the average energy of the transform coefficients for every basis. Etemad *et al.* [5] used an empirical fidelity criterion based on Fisher's class separability metric [16]. Both these efforts used the tree structure of the WPs for designing local pruning-growing algorithms for addressing the BS of their respective optimality criteria. The approximation-estimation error tradeoff was not formally considered in these BS algorithms, while dimensionality reduction is addressed in a postprocessing stage.

This work is related to the aforementioned discriminative BS formulations, however is distinct, in terms of both the set of feature representations obtained from the WP bases, and the optimality criterion used to formulate the BS problem. For feature representation, we consider an analysis-measurement framework that projects the signal into different filter bank subspace decompositions and then compute measurements for the resulting subspaces as a way to obtain a sequence of successively refined features. The filter bank energy measurement is the focus in this paper, motivated by its used in several acoustic [11], [28] and image classification problems [5], [10], [13], [29]. In this way, a family of tree-embedded filter bank energy features is obtained. Concerning the BS, the approximation-estimation error tradeoff is explicitly considered as the objective function in terms of a complexity regularization formulation, where the embedded structure of the WP feature family is used to study conditions that guarantee algorithmic solutions—minimum cost tree-pruning algorithms. In the context of WP filter bank selection, Chang *et al.* [10] proposed a growing algorithm considering the subband energy concentration as the local splitting criterion.

### A. Organization and Contribution

This paper is organized in two parts. In the first part, WP tree-structured feature representations are characterized in terms of an analysis-measurement framework, where the notion of dimensionally embedded feature representations is introduced. Then, sufficient conditions are studied in the adopted fidelity criterion, with respect to the tree-structure of the WP basis family, which allow for implementing the cost-fidelity problem using dynamic programing (DP) techniques. Those conditions are based on a conditional independent structure of the family of random variables induced from the analysis-measurement process, where the cost-fidelity problem reduces to a minimum cost tree-pruning problem [25]. Finally, theoretical results and algorithms, with polynomial complexity in the size of the WP tree are presented, extending ideas both from the context of regression and classification trees [15], [26] and the general single and family-pruning problems recently presented by Scott [25].

In the second part, we address implementation issues and provide some experimental results. First, a nonparametric data-driven approach is derived for estimating the conditional mutual information (CMI) [23], [30]. This is a necessary building block for computing the adopted fidelity criterion—empirical mutual information (MI)—given the aforementioned conditional independence assumptions. In this context, we extend the Darbellay–Vajda tree-structured data-dependent partition [31], originally formulated for estimating MI between two continuous random variables, into our scenario for the CMI. We consider a product partition structure for the proposed CMI estimator, which satisfies desirable asymptotic properties (weak consistency). For experimental evaluation, solutions for the optimal WP decomposition problem are evaluated on a speech phonetic classification task, where the solutions of the proposed optimal filter bank decomposition are evaluated and compared with some standard feature extraction techniques.

The rest of the paper is organized as follows. Section II provides basic notations and summarizes the complexity regularized formulation adopted for this learning problem. Section III presents the WP bases family and its indexing in terms of a tree-structured feature representation. Section IV addresses the cost-fidelity problem for the WP indexed family in terms of minimum cost tree pruning. Section V is devoted to the nonparametric CMI estimation. Finally, Section VI presents experimental evaluations and Section VII provides final remarks. Proofs are provided in the Appendix.

## II. Preliminaries

We adopt standard notation for random variables [32], where some background in probability theory and information theory is assumed, in particular concerning definitions and properties of information theoretic quantities [23], [30]. Let $X:(\Omega,\mathcal{F},\mathbb{P}) \rightarrow (\mathcal{X},\mathcal{F}_{\mathcal{X}})$ denote the observation random variable (r.v.) with values in $\mathcal{X}=\mathbb{R}^K$ (for some $K\in\mathbb{N}$), and $Y:(\Omega,\mathcal{F},\mathbb{P})\rightarrow(\mathcal{Y},\mathcal{F}_{\mathcal{Y}})$ be the class r.v. with values in a finite alphabet set $\mathcal{Y}$, where $(\Omega,\mathcal{F},\mathbb{P})$ refers to the underlying probability space.[1] Considering $\mathcal{M}(\mathcal{X},\mathcal{Y})$ the set of measurable transformations from $(\mathcal{X},\mathcal{F}_{\mathcal{X}})$ to $(\mathcal{Y},\mathcal{F}_{\mathcal{Y}})$, the pattern recognition (PR) problem chooses the decision with the minimum risk, given by $\arg\min_{g\in\mathcal{M}(\mathcal{X},\mathcal{Y})}\mathbb{E}_{X,Y}[l(g(X),Y)]$, where $l(y_1,y_2)$ represents the penalization of labeling an observation with the value $y_1$, when its true label is given by $y_2$. This optimal solution is known as the Bayes' rule, where for the emblematic 0-1 cost function [14], it reduces to the *maximum a posterior* (MAP) decision, $g_{P_{X,Y}}(\bar{x}) = \arg\max_{y\in\mathcal{Y}} P_{X,Y}(\bar{x},y)$, $\forall\bar{x}\in\mathcal{X}$, with the corresponding Bayes' error given by $L_{\mathcal{X}}=\mathbb{P}(\{u\in\Omega: g_{P_{X,Y}}(X(u))\neq Y(u)\})$ [14].

In practice, the joint distribution $P_{X,Y}$ is unknown, and a set of independent and identically distributed (i.i.d.) realizations of the pair $(X,Y)$, denoted by $D_N=\{(x_i,y_i):i\in\{1,\ldots,N\}\}$, is assumed. In the Bayes' decision approach, the supervised data is used to obtain an empirical observation-class distribution $\hat{P}_{X,Y}$ that is in turn used to derive the empirical Bayes' rule, $\hat{g}_{\hat{P}_{X,Y}}(\cdot)=\arg\max_{y\in\mathcal{Y}}\hat{P}_{X,Y}(\cdot,y)$. Important examples for probability estimation are the rich family of $L_1$-consistent kernel-based density estimators and the widely adopted family of Gaussian mixture models (GMMs) density estimators [14], [16], this last one our choice for experimental evaluation in Section VI. It is well known that the risk of the empirical Bayes' rule deviates from $L_{\mathcal{X}}$ as a consequence of estimation

[1] A natural choice for $\mathcal{F}_{\mathcal{X}}$ is the Borel sigma field $\mathcal{B}(\mathbb{R}^K)$ [33], and for $\mathcal{F}_{\mathcal{Y}}$ the power set of $\mathcal{Y}$.

errors [7], [18], [20], [22]. This implies a strong dependency between the number of training examples and the complexity of the observation space, that justifies dimensionality reduction as a fundamental part of feature extraction (FE). For addressing this FE problem in the Bayes' setting, we revisit the minimum probability of error signal representation (MPE-SR) criterion [7], [22].

### A. Minimum Probability of Error Signal Representation and Approximations

Let $\mathbb{D}$ be a dictionary of feature transformations, where any $f \in \mathbb{D}$ is a mapping from the original signal space $\mathcal{X}$ to a transform space $\mathcal{X}_f$, equipped with joint empirical distribution $\hat{P}_{X_f,Y}$ on $(\mathcal{X}_f \times \mathcal{Y}, \sigma(\mathcal{F}_f \times \mathcal{F}_\mathcal{Y}))$ obtained from $D_N$ and an implicit probability estimation approach. Consequently, we have a collection of empirical Bayes' rules that we denote by $\{\hat{g}_f(\cdot) = \arg\max_{y\in\mathcal{Y}} \hat{P}_{X_f,Y}(\cdot, y) : f \in \mathbb{D}\}$. The oracle MPE-SR problem [22] is given by

$$\mathbf{f}^* = \arg\min_{f\in\mathbb{D}} P_{X,Y}\left(\{(x,y) \in \mathcal{X}\times\mathcal{Y} : \hat{g}_f(f(x)) \neq y\}\right) \quad (1)$$

where $P_{X,Y}$ refers to the true joint distribution on $(\mathcal{X} \times \mathcal{Y}, \sigma(\mathcal{F}_\mathcal{X} \times \mathcal{F}_\mathcal{Y}))$. Note that $\forall f \in \mathbb{D}$, $P_{X,Y}(\{(x,y) \in \mathcal{X}\times\mathcal{Y} : \hat{g}_f(f(x)) \neq y\}) \geq L_{\mathcal{X}_f} \geq L_\mathcal{X}$ [14], where $L_{\mathcal{X}_f}$ denotes de Bayes' error of the transform space $\mathcal{X}_f$. Then, this ideal criterion chooses the function whose risk is the closest to $L_\mathcal{X}$. Using the following upper bound $P_{X,Y}(\{(x,y) \in \mathcal{X}\times\mathcal{Y} : \hat{g}_f(f(x)) \neq y\}) \leq \Delta(\hat{P}_{X_f,Y}, P_{X_f,Y}) + L_{\mathcal{X}_f}$ proposed by Vasconcelos [7], where $\Delta(\hat{P}_{X_f,Y}, P_{X_f,Y})$ quantifies the estimation error and is a nondecreasing function of the KLD [27] between the true conditional class probabilities and their empirical counterparts [7], [22], the objective criterion in (1) can be approximated by this bound resulting in [22]

$$\tilde{\mathbf{f}}^* = \arg\min_{f\in\mathbb{D}} \left[L_{\mathcal{X}_f} - L_\mathcal{X}\right] + \Delta\left(\hat{P}_{X_f,Y}, P_{X_f,Y}\right). \quad (2)$$

This last criterion, as desired, makes explicit that the minimum risk decision needs to find the best tradeoff between signal representation quality (approximation error) and learning complexity (estimation error). In practice, neither terms in (2) are available since they require the knowledge of the true distributions. To address this problem from observed data $D_N$, [22] proposes the use of MI [23], [30] as a discriminative indicator to approximate the Bayes' error,[2] and a function proportional to the dimensionality of $\mathcal{X}_f$ for the estimation error term.[3] Consequently, the following complexity regularized selection criterion is adopted:

$$\hat{\mathbf{f}}^*(\lambda) = \arg\min_{f\in\mathbb{D}} -\hat{I}(f(X);Y) + \lambda\cdot\Phi(R(\mathbf{f})) \quad (3)$$

for some $\lambda > 0$, and where $R(\mathbf{f})$ denotes the dimensionality of $\mathcal{X}_f$, $\Phi(\cdot)$ is a strictly increasing real function and $\hat{I}(f(X);Y)$ denotes the empirical mutual information between $f(X)$ and $Y$ estimated from the empirical data. Note that independent of $\Phi(\cdot)$, the domain of solutions of (3), i.e., $\{\hat{\mathbf{f}}^*(\lambda) : \lambda > 0\}$, resides in a sequence of feature transformations solution of the following cost-fidelity problem [22]:

$$\hat{\mathbf{f}}^{k*} = \arg\max_{\substack{f\in\mathbb{D}\\ R(f)\leq k}} \hat{I}(f(X);Y) \quad \forall k \in K(\mathbb{D}) \quad (4)$$

with $K(\mathbb{D}) = \{R(f) : f \in \mathbb{D}\}$. Equation (3) is an approximation of the oracle complexity regularization criterion in (2). In particular, tightness between the Bayes' error and mutual information is not guaranteed. Then, it is not possible to rigorously claim that (3) addresses the ideal minimum risk decision in (1). However, this criterion models with practical terms the mentioned estimation and approximation error quantities and their tradeoff in this learning problem, supporting its adoption as feature selection criterion. Furthermore, it has been implicitly used in some emblematic dimensionality reduction techniques [22], [34]. Finally, the solution for the approximated MPE-SR in (3) requires to choose an appropriate complexity-fidelity weight represented by $\lambda$. As mentioned in more details in Section IV, this again needs to be obtained from the data by evaluating the empirical risk for the family of cost-fidelity empirical Bayes' rules $\{\hat{g}_{\hat{\mathbf{f}}^{k*}}(\cdot) : k \in K(\mathbb{D})\}$ in an independent test set or by cross-validation [7], [15].

Next we particularize this learning-decision framework to our case of interest, the family of filter bank representations induced by WPs. First, we show how the alphabet of feature transformations is created using an analysis-measurement process, and second, how the tree structure of the WPs is used to index this dictionary of feature transformations. This abstraction will be crucial to address the cost-fidelity problem algorithmically, as presented in Section IV.

## III. Tree-Indexed Filter Bank Representations: The Wavelet Packets

WPs allow decomposing the observation space into subspaces associated with different frequency bands [1]. This basis family is characterized by a tree structure induced by its filter bank implementation, that recursively iterates a two channel orthonormal filter bank. In the process of cascading this basic block of analysis, it is possible to generate a rich collection of orthonormal bases for $\mathcal{L}_2(\mathbb{Z})$ [35]—the space of finite energy sequences—associated with different time-scale signal properties [1], [3]. Emblematic examples for these bases include the wavelet basis, which recursively iterates the low frequency band generating a multiresolution type of analysis [2], and the short-time Fourier transform (STFT) with a balanced filter bank structure [3], [6], illustrated in Fig. 1. For a comprehensive treatment of WPs, we refer to the excellent expositions in [1], [6], and [12].

### A. Tree-Indexed Basis Collections and Subspace Decomposition

Here, as considered in [12] and [13], we use the WP two-channel filter bank implementation to hierarchically index the WP bases and its subspace decomposition. Let $\mathcal{X} = \mathbb{R}^K$ again be our finite-dimensional raw observation space. Then, the application of the basic block of analysis—two channel filter bank and down-sampling by 2 [1]—decomposes $\mathcal{X}$

[2]Fano's inequality [23, Ch. 2.11] characterizes a lower bound for the probability of error of any decision framework $g(\cdot) : \mathcal{X} \to \mathcal{Y}$ that tries to infer $Y$ as a function of $X$ and offers the tightest lower bound for the Bayes' rule.

[3]Supporting this choice, [22, Th. 2] shows that the estimation error is monotonically increasing with the dimensionality of the space under some general dimensionally embedded consistency assumptions.

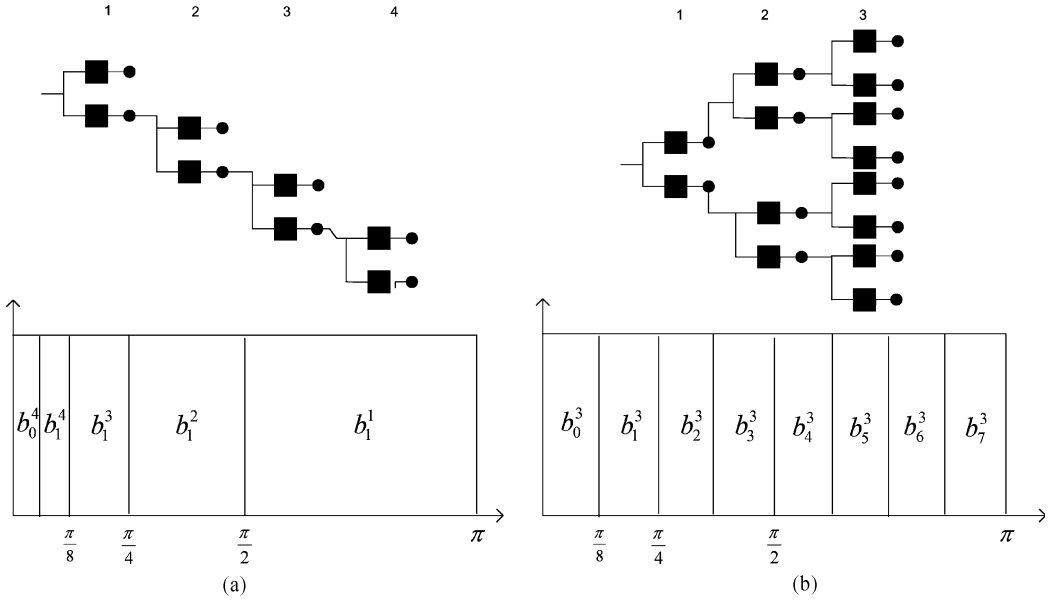


Fig. 1. Filter bank decomposition given the tree-structured of wavelet packet bases for the case of the ideal Sinc half-band two channel filter bank. **Case A**: Octave-band filter bank characterizing a Wavelet type of basis representation. **Case B**: Short-time Fourier transform (STFT) type of basis representation.

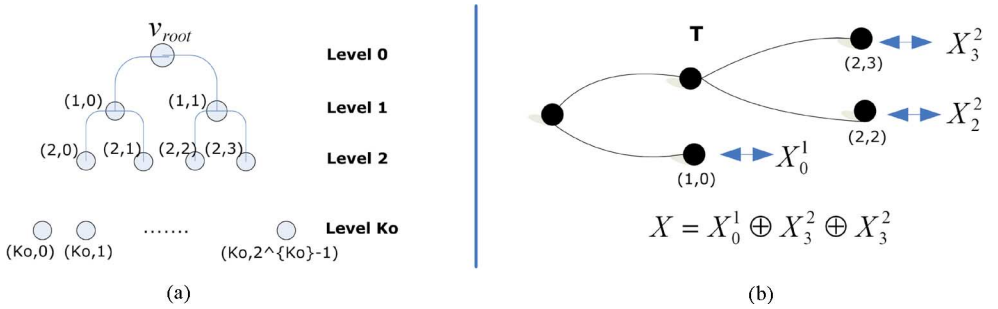


Fig. 2. Topology of the full rooted binary tree $\mathbf{T}_{\text{full}}$ and representation of the tree indexed subspace WP decomposition.

into two subspaces $\mathcal{X}_0^1$ and $\mathcal{X}_1^1$, respectively associated with its low and high frequency content. This process can be represented as an indexed-orthonormal basis that we denote by $\mathcal{B} = \{\psi_{0,k1}^1, \psi_{1,k2}^1 : k_1 \in \mathcal{A}_0^1, k_2 \in \mathcal{A}_1^1\}$, where $\mathcal{X} = \mathcal{X}_0^1 \bigoplus \mathcal{X}_1^1$, being $\mathcal{X}_i^1 \equiv span\{\psi_{i,k}^1 : k \in \mathcal{A}_i^1\}, i \in \{0,1\}$. The indexed structure of $\mathcal{B}$ is represented by the way its basis elements are dichotomized in terms of the filter bank index sets $\mathcal{A}_0^1$ and $\mathcal{A}_1^1$, which are responsible for the subspace decomposition. In any of the resulting subband spaces, $\mathcal{X}_0^1$ and $\mathcal{X}_1^1$, we can reapply the basic block of analysis to generate a new indexed basis. By iterating this process, it is possible to construct a binary tree-structured collection of indexed bases for $\mathcal{X}$. For instance, by iterating this decomposition recursively $l$-times from one step to another in every subband space, we can generate the indexed basis $\bigcup_{j\in\{0,\ldots,2^l-1\}}\{\psi_{j,k}^l : k \in \mathcal{A}_j^l\}$, where $\mathcal{X} = \bigoplus_{j\in\{0,\ldots,2^l-1\}} \mathcal{X}_j^l$, and $\mathcal{X}_j^l \equiv \text{span}\{\psi_{j,k}^l : k \in \mathcal{A}_j^l\}, \forall j \in \{0,\ldots,2^l-1\}$. It is important to mention that this construction ensures that in any iteration we have the following relationship: $\mathcal{X}_j^l = \mathcal{X}_{2j}^{l+1} \bigoplus \mathcal{X}_{2j+1}^{l+1}$, $\forall l \in \{0,\ldots,K_o-1\}, \forall j \in \{0,\ldots,2^l-1\}$ and, hence $\{\psi_{2j,k_1}^{l+1}, \psi_{2j+1,k_2}^{l+1} : k_1 \in \mathcal{A}_{2j}^{l+1}, k_2 \in \mathcal{A}_{2j+1}^{l+1}\}$ is an indexed basis for the subspace $\mathcal{X}_j^l$. Finally from this construction, it is clear that there is a one-to-one mapping between a family of trees in a certain graph and the family of WP bases, which we formalize next.

### B. Rooted Binary Tree Representation

We represent the generative process of producing a particular indexed basis in the WP family by a rooted binary tree [25]. Let $K_o = \lfloor \log_2(K) \rfloor$ be the maximum number of iterations of this subband decomposition process, given our finite dimensional setting.[4] Let $G = (V, E)$ be a graph with $V = \{(0,0),(1,0),(1,1),\cdots,(K_o,0)\cdots(K_o,2^{K_o}-1)\}$, and $E$ the collection of arcs on $V \times V$ that characterizes a full rooted binary tree with root $v_{\text{root}} \equiv (0,0)$, as illustrated in Fig. 2(a). Instead of representing the tree as a collection of arcs in $G$, we use the convention proposed by Breiman *et al.* [15], where subgraphs are represented by subset of nodes of the full graph. In this context, any pruned version of the full rooted binary tree represents a particular way of iterating the basic two channel block analysis of the WP.

Before continuing with the exposition, let us introduce some basic terminology. We use the basic concepts of *child*, *parent*, *path*, *leaf* and *root* used in graph theory [36]. We

[4]Without loss of generality, we consider $K = 2^{K_o}$ for the rest of the paper to simplify the exposition.

describe a rooted binary tree $T = \{v_0, v_1, \ldots\} \subset V$ as a collection of nodes with only one with degree 2, the root node, and the remaining nodes with degree 3 (internal nodes) and 1 (leaf nodes).[5] We define $\mathcal{L}(T)$ as the set of leaves of $T$ and $\mathcal{I}(T)$ as the set of internal nodes $T$, consequently, $\mathcal{L}(T) \cup \mathcal{I}(T) = T$. We say that a rooted binary tree $S$ is a subtree of $T$ if $S \subset T$. In the previous definition, if the roots of $S$ and $T$ are the same then $S$ is a pruned subtree of $T$, denoted by $S \ll T$. In addition if the root of $S$ is an internal node of $T$ then $S$ is called a branch of $T$. In particular, we denote the largest branch of $T$ rooted at $v \in T$ as $T_v$. We define the size of the tree $T$ as the number of terminal nodes, i.e., the cardinality of $\mathcal{L}(T)$, and denote it by $|T|$. Finally, let $\mathbf{T}_{\text{full}} = V$ denote the full binary tree illustrated in Fig. 2(a).

The WP bases can be indexed by $\{T : T \ll \mathbf{T}_{\text{full}}\}$. More precisely, for any $T \ll \mathbf{T}_{\text{full}}$ the induced tree indexed basis is given by $\mathcal{B}_T = \bigcup_{(l,j)\in\mathcal{L}(T)} \{\psi^l_{j,k} : k \in \mathcal{A}^l_j\}$ and its filter bank subspace decomposition by $\{\mathcal{X}^l_j : (l,j) \in \mathcal{L}(T)\}$.

### C. Analysis-Measurement Process

In association with the subspace decomposition, we consider a final measurement step for feature extraction. Let $\mathcal{B}_T$ be a basis element in the collection, the analysis-measurement mapping is given by

$$m_T(x) = \left(M^l_j(x)\right)_{(l,j)\in\mathcal{L}(T)}, \quad \forall x \in \mathbb{R}^K \tag{5}$$

where $M^l_j(x)$ represents a measurement of the signal components in the subspace $\mathcal{X}^l_j$. In particular, we consider $M^l_j(x) = \mathbf{F}((\langle x, \psi^l_{j,k}\rangle)_{k\in\mathcal{A}^l_j})$ with $\mathbf{F}(\cdot)$ representing the measurement function. While in the development of this work we focus on the subspace energy as the measurement function, the formulation and results presented in the next sections could be extended for more general feature transformations.

## IV. Approximated MPE-SR for WP: The Minimum Cost Tree-Pruning Problem

Let $X$ and $Y$ be the observation and class label random variables, respectively. For the MPE-SR formulation, we consider $\mathbb{D}_{K_o} = \{m_T(\cdot) : T \ll \mathbf{T}_{\text{full}}\}$ the dictionary of transformations with $K_o$ levels of decomposition, where from Section II-A the approximated MPE-SR reduces to solve the following cost-fidelity problem:

$$\mathbf{T}^{k*} = \arg \max_{\substack{T \ll \mathbf{T}_{\text{full}} \\ |T| \le k}} \hat{I}\left(m_T(X); Y\right) \tag{6}$$

$\forall k \in \{1, \ldots, |\mathbf{T}_{\text{full}}| = 2^{K_o}\}$, where $\hat{I}(m_T(X); Y)$ is the empirical MI between $m_T(X)$ and $Y$.

The solution of this problem turns to finding the subband decomposition of $\mathcal{X}$ that maximizes MI for a given number of frequency bands, a discriminative band allocation problem. Note that without some additive property on the tree functionals involved in (6), in particular the mutual information, an exhaustive search is needed for solving it, which grows exponentially with the size of the problem. The next subsections derive general sufficient conditions to address this problem using DP techniques and provide some theoretical results and connections with known minimum cost tree-pruning algorithms [15], [24], [25]. To simplify this analysis, MI will be considered as a general function of a joint probability distribution [23], [30] (the empirical MI will not be mentioned explicitly). Then, Section V will address the specific problem of MI estimation based on empirical data, which at the end is needed for solving (6).

[5]The degree of a node is the number of arcs connecting the node with its neighbors.

### A. Tree-Embedded Feature Representation Results

To simplify notation let $\rho(T) \equiv I(m_T(X); Y)$ denote our target MI tree functional, and $X^l_j \equiv M^l_j(X)$ denote the random measurement of $X$ in the subspace $\mathcal{X}^l_j$, then $m_T(X) = (X^l_j)_{(l,j)\in\mathcal{L}(T)}$. The following propositions state some basic tree-embedded properties of our dictionary of feature representations.

*Proposition 1:* The collection $\{m_T(X) : T \ll \mathbf{T}_{\text{full}}\}$ is embedded in the sense that

$$H\left(X^l_j | X^{l+1}_{2j}, X^{l+1}_{2j+1}\right) = 0 \quad \forall (l,j) \in \mathcal{I}(\mathbf{T}_{\text{full}}) \tag{7}$$

where $H(\cdot|\cdot)$ refers to the conditional differential entropy [23]. Furthermore, for any sequence of rooted binary trees $T_1 \ll \cdots \ll T_m$, $\{m_{T_1}(X), \cdots, m_{T_m}(X)\}$ is embedded in the sense that

$$H\left(m_{T_i}(X) | m_{T_k}(X)\right) = 0 \quad \forall i,k\ 1 \le i < k \le m. \tag{8}$$

The proof is presented in Appendix A.

*Proposition 2:* Let us consider $T \ll \bar{T}$, then from Proposition 1 we have that $\rho(T) \le \rho(\bar{T})$. Furthermore, this MI difference can be expressed by

$$\rho(\bar{T}) - \rho(T) = I\left(\left(X^l_j\right)_{(l,j)\in\bar{T}\setminus T}; Y | m_T(X)\right), \tag{9}$$

where $I(\cdot;\cdot|\cdot)$ denotes the CMI [23].

This result says that this MI gain can be expressed by the conditional mutual information of the energy features of the branches of $\bar{T}$ emerging from $T$ condition to $m_T(X)$. This results is a consequence of the tree-embedded structure of $\{m_T(X) : T \ll \mathbf{T}_{\text{full}}\}$ in (8), the proof is given in Appendix B. Note that from Proposition 2, there exists a solution for (6), such that $|\mathbf{T}^{k*}| = k$, $\forall k \in \{1, \ldots, |\mathbf{T}_{\text{full}}|\}$.

### B. Studying Additive Properties for the Mutual Information Tree Functional

We begin studying the MI gain by iterating the two channel filter bank of WP in a particular scale-frequency band. More precisely, let $T \ne \mathbf{T}_{\text{full}}$ be a rooted binary tree and let $T^+_{l,j}$ denote the tree induced by splitting an admissible leaf node $(l,j) \in \mathcal{L}(T) \setminus \mathcal{L}(\mathbf{T}_{\text{full}})$. From Proposition 2, the MI gain $\rho(T^+_{l,j}) - \rho(T)$ is equal to $I(X^{l+1}_{2j}, X^{l+1}_{2j+1}; Y | (X^{l'}_{j'})_{(l',j')\in\mathcal{L}(T)})$, which is not a local function of $(l,j)$, but a function of the statistical dependency of the complete holding tree structure $(X^{l'}_{j'})_{(l',j')\in\mathcal{L}(T)}$. The following result presents sufficient conditions to simplify this dependency, which requires the introduction of a Markov

tree assumption on the conditional independence structure of $\{X_j^l : (l,j) \in \mathbf{T}_{\text{full}}\}$.

*Proposition 3:* Let $\{X_j^l : (l,j) \in \mathbf{T}_{\text{full}}\}$ be the filter bank energy measurements and let $\mathbf{T}_{(l,j)}$ denote the largest branch of $\mathbf{T}_{\text{full}}$ rooted at $(l,j)$. If $\forall (l,j) \in \mathcal{I}(\mathbf{T}_{\text{full}})$, $\{X_{\bar{j}}^{\bar{l}} : (\bar{l},\bar{j}) \in \mathbf{T}_{(l,j)} \setminus \{(l,j)\}\}$ and $\{X_{\bar{j}}^{\bar{l}} : (\bar{l},\bar{j}) \in \mathbf{T}_{\text{full}} \setminus \mathbf{T}_{(l,j)}\}$ are conditionally independent given $X_j^l$ and given both $X_j^l$ and $Y$, then, $\forall T \ll \mathbf{T}_{\text{full}}, \forall (l,j) \in \mathcal{L}(T) \cap \mathcal{I}(\mathbf{T}_{\text{full}})$,

$$\rho\left(T_{l,j}^{+}\right) - \rho(T) = I\left(X_{2j}^{l+1}, X_{2j+1}^{l+1}; Y | X_j^l\right). \tag{10}$$

The proof is direct from the definition of the conditional mutual information [23].

The condition stated in Proposition 3 is a Markov property with respect to the tree ordering of the family of filter bank energy random variables. This Markov tree property depends on the goodness of the index bases family to decompose the observation process into conditional independent components. Given that we are working with the WP bases, their frequency band decomposition provides good decorrelation for wide-sense stationary random processes, and independent components for stationary Gaussian processes [32], [37] under the ideal Sinc half-band two channel filter bank [1], scenario that shows the mentioned Markov tree property. Working under this Markov tree assumption will be the focus for the rest of this exposition, and consequently the following algorithmic solutions and results (Sections IV-C and IV-D) are restricted to this condition. This Markov tree property can be considered a reasonable approximation, assuming good frequency selectivity in the two channel filter bank, as it has been empirically shown to be an important design consideration for time-series phonetic classification [9], and that the observation source has a stationary behavior, as it has been considered to model the short-term scale behavior of the acoustic speech process.

Before continuing, let us introduce some short-hand notations. We denote the local CMI gain in (10) by $\Delta\rho(l,j) \equiv I(X_{2j}^{l+1}, X_{2j+1}^{l+1}; Y | X_j^l)$, well defined $\forall (l,j) \in \mathcal{I}(\mathbf{T}_{\text{full}})$. Let $T$ be a nontrivial tree (i.e., $|T| > 1$) and $(l,j) \in \mathcal{I}(T)$, then $\rho(T_{(l,j)}) \equiv I((X_{\hat{j}}^{\hat{l}})_{(\hat{l},\hat{j}) \in \mathcal{L}(T_{(l,j)})}; Y)$ denotes the MI between the energy features associated to the branch $m_{T_{(l,j)}}(X)$ and $Y$. Finally, for $T$ nontrival and $(l,j) \in \mathcal{I}(T)$, let us define

$$\rho_T(l,j) \equiv I\left(m_{T_{(l,j)}}(X); Y | X_j^l\right) = \rho\left(T_{(l,j)}\right) - I\left(X_j^l; Y\right) \tag{11}$$

as the MI between the energy measurements $m_{T_{(l,j)}}(X)$ and $Y$ condition to the root random variable $X_j^l$.

Under the Markov tree property of Proposition 3, $\rho(T)$ can be expressed as a function of the local CMIs $\{\Delta\rho(l,j) : (l,j) \in \mathcal{I}(\mathbf{T}_{\text{full}})\}$ and the MI of the root node, i.e., $I(X_0^0; Y)$. The following results formalize this point and the general additive property of our MI tree functional.

*Theorem 1:* Let $T, \bar{T}$ be binary trees such that $T \ll \bar{T}$. Then the following results hold:

$$\rho(\bar{T}) = \rho(T) + \sum_{(l,j) \in \mathcal{I}(\bar{T}) \setminus \mathcal{I}(T)} \Delta\rho(l,j) \tag{12}$$

$$= \rho(T) + \sum_{(l,j) \in \mathcal{L}(T)} \rho_{\bar{T}}(l,j). \tag{13}$$

In particular from (12), $\forall T \ll \mathbf{T}_{\text{full}}$ nontrivial (i.e., $|T| > 1$) we have that

$$\rho(T) = I\left(X_0^0; Y\right) + \sum_{(l,j) \in \mathcal{I}(T)} \Delta\rho(l,j). \tag{14}$$

The proof is presented in Appendix C.

The following proposition presents the important pseudo-additive property of $\rho(\cdot)$ when the tree argument of the functional is partitioned in terms of its primary left and right branches.

*Proposition 4:* Let $T \ll \mathbf{T}_{\text{full}}$ be a nontrivial tree ($|T| > 1$). Then, for all $(l,j) \in \mathcal{I}(T)$, we have that

$$\rho_T(l,j) = \Delta\rho(l,j) + \rho_T(l+1,2j) + \rho_T(l+1,2j+1) \tag{15}$$

while for $(l,j) \in \mathcal{L}(T)$ by definition $\rho_T(l,j) = 0$. The proof is presented in Appendix D.

From (12), we observe that $\rho(\cdot)$ is additive with respect to the internal nodes of the tree, which implies that $\rho(\cdot)$ is an affine tree functional [24].[6] Moreover, by definition (11)

$$\rho(T) = I\left(X_0^0; Y\right) + \rho_T(v_{\text{root}}) \quad \forall T \ll \mathbf{T}_{\text{full}} \tag{16}$$

then from (15), we have a way of characterizing $\rho(T)$ as an additive combination of a root dependent term and $\rho(\cdot)$ evaluated in its primary left and right branches. Next, we present a DP solution for the cost-fidelity problem in (6) using the additive properties of our fidelity indicator presented in Theorem 1 and Proposition 4.

### C. Minimum Cost Tree-Pruning Problem

The cost-fidelity problem in (6) can be formalized as a minimum cost tree-pruning problem [15], [24], [25]. Adopting the new short-hand notation for the MI tree functionals in (11), we need to solve

$$\mathbf{T}^{k*} = \arg \max_{\substack{T \ll \mathbf{T}_{\text{full}} \\ |T| = k}} \rho(T) = \arg \max_{\substack{T \ll \mathbf{T}_{\text{full}} \\ |T| = k}} \rho_T(v_{\text{root}}) \tag{17}$$

$\forall k \in \{1, \ldots, |\mathbf{T}_{\text{full}}| = 2^{Ko}\}$. Let $\mathbf{T}_v$ be the largest branch of $\mathbf{T}_{\text{full}}$ rooted at $v \in \mathbf{T}_{\text{full}}$ and let $\mathbf{T}_v^{k*} (\ll \mathbf{T}_v)$ denote the solution of the more general branch dependent optimal tree-pruning problem

$$\mathbf{T}_v^{k*} = \arg \max_{\substack{T \ll \mathbf{T}_v \\ |T| = k}} \rho_T(v) \quad \forall k \in \{1, \ldots, |\mathbf{T}_v|\}. \tag{18}$$

Then, we can state the following result.

*Theorem 2:* Let us consider an arbitrary internal node $v \in \mathcal{I}(\mathbf{T}_{\text{full}})$ and denote its left and right children by $l(v)$ and $r(v)$, respectively. Assuming that we know the solution of (18) for the child nodes $l(v)$ and $r(v)$, i.e., we know $\{\mathbf{T}_{l(v)}^{k_1 *}, \mathbf{T}_{r(v)}^{k_2 *} : k_1 = 1, \ldots, |\mathbf{T}_{l(v)}| k_2 = 1, \ldots, |\mathbf{T}_{r(v)}|\}$, the solution of (18) for the parent node is given by

[6] A tree functional $\rho(\cdot)$ is affine if, for any $T, S$ rooted binary trees such that $S \ll T$, then $\rho(T) = \rho(S) + \sum_{v \in \mathcal{L}(S)} [\rho(T_v) - \rho(\{v\})]$, where $\{v\}$ represents a trivial binary tree. For our MI tree functional, this property is obtained from (13).

$\mathbf{T}_v^{k*} = [v, T_{l(v)}^{\hat{k}_1*}, T_{r(v)}^{\hat{k}_2*}]$, where[7] $(\hat{k}_1, \hat{k}_2)$ is given in (19) at the bottom of the page, $\forall k \in \{1, \ldots, |\mathbf{T}_v|\}$. In particular, when $v$ is equal to the root of $\mathbf{T}_{\text{full}}$ the solution for the optimal pruning problem in (17), is given by $\mathbf{T}^{k*} = [v_{\text{root}}, \mathbf{T}_{l(v_{\text{root}})}^{\hat{k}_1*}, \mathbf{T}_{r(v_{\text{root}})}^{\hat{k}_2*}]$. The proof is presented in Appendix E.

This DP solution is a direct consequence of solving (18) for the parent node as a function of the solutions of the same problem for its direct descendants. In particular, if we index all the nodes from top to bottom, such that $\text{index}(v) > \max\{\text{index}(l(v)), \text{index}(r(v))\}$, then we can solve an ordered sequence of optimal tree-pruning problems, from the terminal nodes of $\mathbf{T}_{\text{full}}$—where the solution is trivial—to the root. The algorithm presented by Scott [25] for the minimum cost tree pruning with additive fidelity tree functional can be extended directly to this problem. Bohanec *et al.* [38] showed that the computational complexity of this algorithm is $O(|\mathbf{T}_{\text{full}}|^2)$ for balanced trees, which is our case.

The next subsection goes one step back to revisit our main complexity regularized problem in (3) and provides further connections with the minimum cost trees $\{\mathbf{T}^{k*} : k = 1, \ldots, 2^{K_o}\}$ presented here. Furthermore, the next subsection shows that under additional conditions in the penalization term, the problem in (3) reduces to finding a more restrictive sequence of optimal tree-pruned representations.

### *D. Connections With the Family-Pruning Problem With General Size-Based Penalty*

In our filter bank selection scenario, the approximated MPE-SR problem in (3) can be equivalently expressed as the following "single tree-pruning problem" with generalized size-based penalty [25]

$$\mathbf{T}^*(\lambda) = \arg \min_{T \ll \mathbf{T}_{\text{full}}} -\rho(T) + \lambda \Phi(|T|) \tag{20}$$

with $\Phi : \mathbb{N} \to \mathbb{R}^+$ a nondecreasing function and $\lambda \in \mathbb{R}^+$ the relative weight between the fidelity and cost terms. In this context, $-\rho(T)$ can be seen as the MI loss for having a coarse representation of the raw observation, and $\Phi(|T|)$ is the regularization term that penalizes dimensionality. Proposition 1 in [25] shows that when $\Phi$ is strictly increasing, then there exists $\lambda_0 = 0 < \lambda_1 < \cdots \lambda_m = \infty$ and a sequence of pruned trees $R_1, \ldots, R_m$ (with $|R_1| > |R_2| \ldots > |R_m| = 1$), such that $\forall i \in \{1, \ldots, m\}$,

$$\mathbf{T}^*(\lambda) = R_i \quad \forall \lambda \in [\lambda_{i-1}, \lambda_i). \tag{21}$$

This result characterizes the full range of solutions for (20) (or the achievable cost-fidelity boundary [24]). The problem of finding $\{\lambda_0, , \ldots \lambda_{m-1}\}$ and the associated solutions $\{R_1, \ldots, R_m\}$ was coined as the "family-pruning problem" under general size-based penalties [25]. It is not difficult to see that $R_j$ is an admissible solution of the minimum cost tree pruning in (17) for $k_j = |R_j|$, i.e., $R_j = \mathbf{T}^{k_j*}$. Consequently, we can consider that $\{R_1, \ldots, R_m\} \subset \{\mathbf{T}^{k*} : k = 1, \ldots, 2^{K_o}\}$. Interestingly, if the cost function is additive, the following result can be stated.

*Theorem 3 (Chou et al. [24]):* If $\Phi(|T|) = |T|$, then the solution of the family-pruning problem admits an embedded structure, i.e., $R_m \ll R_{m-1} \ll \cdots \ll R_1$.[8]

We derived a clean algebraic proof for this result based on Breiman *et al.*'s derivations [15, Ch. 10.2], which is not reported here for space considerations. By Theorem 3, we have that the family-pruning problem admits a nested solution. Consequently, a simpler algorithm can be used for finding $\{R_1, \ldots, R_m\}$. The algorithm is presented in [24] and has complexity $O(|\mathbf{T}_{\text{full}}| \log(|\mathbf{T}_{\text{full}}|))$ for the case of balanced trees. Furthermore, Theorem 3 and its algorithm can be extended to a more general family of subadditive penalties—functionals dominated by an additive cost, as presented by Scott in [25, Th. 2].

Finally, as in the CART pruning algorithm [15], [25], [26], the true value of $\lambda$ that reflects the right weight between the fidelity and cost term of the problem is unknown. The problem then reduces to finding the optimal $\lambda^*$ and consequently $\mathbf{T}^*(\lambda^*)$. In practice the empirical data $D_N$ has to be used for this final decision as well, for instance using an independent test set or by cross-validation, depending of how much data is available. In our Bayes' setting, this is done by considering the empirical risk minimization (ERM) criterion across the set of empirical Bayes' rules defined for every member of $\{R_1, \ldots, R_m\}$, or to the more complete family of minimum cost trees $\{\mathbf{T}^{k*} : k = 1, \ldots, 2^{K_o}\}$. This is the step where the set of empirical Bayes' rules come into play and where feature extraction and classification are optimized jointly for the task. Considering that additive assumption for the cost term is difficult to be rigorously justified in our Bayes' decision setting (and consequently the more efficient solution to find $\{R_1, \ldots, R_m\}$ from Theorem 3), and that re-sampling is used as the final decision step, it is reasonable to consider the full minimum cost tree family as the domain for this final empirical decision.

What we have not addressed so far and was taken for granted to obtain the minimum cost tree solutions in this section, is how to estimate the fidelity functional in (17) and (20) based on empirical data. The adopted approach is based on a nonparametric techniques, which is the focus of the next section.

[7]Using Scott's nomenclature [25], the notation $[v, T_1, T_2]$ represents a binary tree $T$ with root $v$, $T_{l(v)} = T_1$ and $T_{r(v)} = T_2$.

[8]The proof of this theorem can be obtained from the fact that this set of solutions characterizes an operational rate-distortion region associated with two monotone affine tree functionals. See the argument of Chou *et al.* [24, Lemma 1] for details.

$$(\hat{k}^1, \hat{k}^2) = \arg \max_{\substack{(k_1,k_2) \in \{1,\ldots,|\mathbf{T}_{l(v)}|\} \times \{1,\ldots,|\mathbf{T}_{r(v)}|\} \\ k_1+k_2=k}} \left[ \rho_{\mathbf{T}_{l(v)}^{k_1*}}(l(v)) + \rho_{\mathbf{T}_{r(v)}^{k_2*}}(r(v)) \right] \tag{19}$$

## V. Nonparametric Estimation of the CMI Gains

The solutions of the minimum cost tree-pruning problems in (17) and (18) require the estimation of the conditional mutual information (CMI) quantities $\{I(X_j^{l+1}, X_{j+1}^{l+1}; Y|X_j^l) : (l,j) \in \mathcal{L}(\mathbf{T}_{\text{full}})\}$, by Theorem 1. To solve this problem a nonparametric approach is adopted based on vector quantization (VQ) [31]. In this section, we propose a quantized CMI construction, state its asymptotic desirable properties and finally introduce the role of data-dependent VQ for the problem, where an algorithm is presented based on Darbellay–Vajda tree-structured data-dependent partition [14], [31].

### A. Quantized CMI Construction

Our basic problem is to estimate $\Delta\rho = I(X_1, X_2; Y|X_3)$ based on i.i.d. realizations of the joint phenomenon. Without loss of generality let $X_1$, $X_2$, and $X_3$ be three continuous random variables in $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $Y$ the finite alphabet class random in $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$. We denote by $P_{X_i}$ the probability of $X_i$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, and we assume it has a probability density function (pdf) given by $p_{X_i}$. The same is assumed for the joint probability of $(X_1, X_2, X_3)$, with pdf $p_{X_1,X_2,X_3}$ defined on $(\mathbb{R}^3, \mathcal{B}(\mathbb{R}^3))$ and for the class conditional probabilities denoted by $P_{X_1,X_2,X_3|Y}(\cdot|y)$ with corresponding pdfs given by $p_{X_1,X_2,X_3|Y}(\cdot|y)$ $\forall y \in \mathcal{Y}$. Our CMI construction follows Darbellay *et al.* [31], by using quantized versions of $X_1$, $X_2$, and $X_3$ by the following type of product partition, $Q_{1,2\times 3} \equiv Q_{1,2} \times Q_3 = \{R_i^{1,2} \times R_j^3 : i = 1, \ldots, n; j = 1, \ldots, n\}$, where $Q_{1,2} = \{R_i^{1,2} : i = 1, \ldots, n\}$ and $Q_3 = \{R_j^3 : j = 1, \ldots, n\}$ are measurable partitions of $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ and $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, respectively. Based on this product partition our quantized CMI is given by

$$\Delta\rho(Q_{1,2\times 3}) = I^{Q_{1,2\times 3}}(X_1, X_2, X_3; Y) - I^{Q_3}(X_3; Y) \quad (22)$$

where for any arbitrary continuous random variable $X$ in $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ and partition $Q$ of $\mathbb{R}^k$, $I^Q(X;Y)$ refers to[9] $\sum_{y\in\mathcal{Y}} \sum_{A\in Q} P_{X,Y}(A \times \{y\}) \log(P_{X,Y}(A \times \{y\})/P_X(A)P_Y(\{y\}))$. It is well known that quantization reduces the magnitude of information quantities [30], [31], which is also the case for our quantized CMI construction, i.e., $\Delta\rho(Q_{1,2\times 3}) \leq \Delta\rho = I(X_1, X_2; Y|X_3)$. Then it is interesting to study the approximation properties of the proposed product CMI construction. In other words, the goal is to find if this suggested construction can achieve $\Delta\rho$ by systematically increasing the resolution of a sequence of product quantizers—a notion of asymptotic sufficient partitions for the CMI estimation. In this direction, we have extended the work of Darbellay *et al.* [31] showing general sufficient conditions in the asymptotic structure of a sequence of nested product partitions for approximating the CMI. This important result justifies our choice of product partition in the asymptotic regime. The proof of this result is not in the main scope of this paper and not reported here for space considerations.

In practice we have a collection of i.i.d. samples and hence empirical distributions will be used to estimate $\Delta\rho(Q_{1,2\times 3})$ in (22). More precisely, let $\{(x_1^i, x_2^i, x_3^i, y^i) : i = 1, \ldots, N\}$ be our empirical data and $Q_{1,2\times 3}$ an arbitrary product measurable partition. The empirical joint distribution of the quantized observation random variable $((X_1, X_2)^{Q_{1,2}}, X_3^{Q_3})$ and class random variable $Y$, using the maximum-likelihood (ML) criterion, is given by $\hat{P}^N_{(X_1,X_2)^{Q_{1,2}}, X_3^{Q_3}, Y}(A_{1,2} \times A_3 \times \{y\}) = (1/N) \sum_{i=1}^N \mathbb{I}_{A_{1,2}\times A_3 \times \{y\}}((x_1^i, x_2^i), x_3^i, y^i)$, $\forall A_{1,2} \in Q_{1,2}$, $\forall A_3 \in Q_3$ and $\forall y \in \mathcal{Y}$. The associated marginal empirical distributions are computed accordingly. Hence, we can obtain the empirical MIs using the following formula[10]:

$$\hat{I}_N^{Q_{1,2\times 3}}(X_1, X_2, X_3; Y) = \frac{1}{N} \sum_{i=1}^N \log \frac{\hat{P}^N\left(A_{1,2}^i \times A_3^i \times \{y\}\right)}{\hat{P}^N\left(A_{1,2}^i \times A_3^i\right) \cdot \hat{P}^N\left(\{y\}\right)} \quad (23)$$

$$\hat{I}_N^{Q_3}(X_3; Y) = \frac{1}{N} \sum_{i=1}^N \log \frac{\hat{P}^N\left(A_3^i \times \{y\}\right)}{\hat{P}^N\left(A_3^i\right) \cdot \hat{P}^N\left(\{y\}\right)} \quad (24)$$

and consequently the empirical CMI by $\hat{\Delta\rho}_N(Q_{1,2\times 3}) = \hat{I}_N^{Q_{1,2\times 3}}(X_1, X_2, X_3; Y) - \hat{I}_N^{Q_3}(X_3; Y)$.

Considering the product sufficient partition sequence for the CMI and sufficient number of samples points, from the weak law of large numbers [39], [40] it is simple to show that $\hat{\Delta\rho}_N(Q_{1,2\times 3})$ can be arbitrarily close to $\Delta\rho$ in probability, which is a desired weak consistency result. However, in practice we need to deal with the nonasymptotic case of having a finite amount of training data. In this context, the problem of finding a good estimation for $\Delta\rho$ across a sequence of nested partitions needs to consider an approximation and estimation error tradeoff, as with any other statistical learning problem [14]. To address this issue, we follow the data-dependent partition framework proposed by Darbellay *et al.* [31].

### B. Darbellay–Vajda Data-Dependent Partition

The Darbellay-Vajda algorithm partitions the observation space, by iterating a splitting rule that generates a sequence of tree-indexed nested partitions [31]. To illustrate the idea, let $X$ and $Y$ be continuous scalar random variables and let us consider the problem of estimating $I(X;Y)$. In addition, let $\{(x^i, y^i) : i = 1, \ldots, N\}$ denote the training data and $\hat{P}^N$ be the empirical probability with distribution function denoted by $\hat{F}_X(x) = \hat{P}_X^N((-\infty, x])$.[11] The algorithm starts with a partition that considers the full space. In the $k$-phase of this algorithm the criterion checks every atom $A$ of the current partition $Q^k$ by evaluating the empirical MI gain obtained by partitioning $A$ with a product structure adaptively generated with the marginal distribution of the training points in $A$, denoted by $Q(A)$.[12] If this gain is above a critical threshold the algorithm splits the atom to upgrade the partition by $Q^{k+1} = (Q^k \setminus \{A\}) \cup Q(A)$ and continues in this region applying recursively the aforementioned

[9] $I^Q(X;Y)$ can be seen as the MI between the quantized random variable. $X^Q = \sum_{A\in Q} \mathbb{I}_A(X) \cdot f(A)$—$f(\cdot)$ being a general injective function from $Q$ to $\mathbb{R}^k$—and $Y$.

[10] The subscript indexes on the probabilities are omitted to simplify notation in (23) and (24).

[11] We consider $X$ as a scalar random variable, however the construction extents naturally for the finite dimensional scenario.

[12] The marginal MI gain can be expressed by $\hat{P}_X^N(A) \cdot \hat{I}_N^{Q^2(A)}(X; Y|X \in A)$.

**Phase 0**: (Initialization: Statistically Equivalent Blocks)
$Q^{(1)} = \{(-\infty, a_1], (a_1, a_2], ..., (a_{r-1}, \infty)\}$, where $a_j \equiv \hat{F}_X^{-1}(\frac{j}{r})$, $j \in \{1, .., r-1\}$.
**Phase 1**: Step $k \rightarrow k+1$ (Main recursion)
$Q^{(k+1)} = \varnothing$, (*initialization*)
**foreach**, $A \in Q^{(k)}$
**if** ($\hat{P}_X^N(A) > \frac{N_c}{N}$) (*critical number of samples per bin*)
- compute: $\hat{F}_A(x) = \frac{\hat{P}_X^N((-\infty, x] \cap A)}{\hat{P}_X^N(A)}$
- construct: $Q_r^{(k+1)}(A) = \{(-\infty, a_1^r], (a_1^r, a_2^r], ..., (a_{r-1}^r, \infty)\} \cap A$ and $Q_s^{(k+1)}(A) = \{(-\infty, a_1^s], (a_1^s, a_2^s], ..., (a_{s-1}^s, \infty)\} \cap A$ where:

$$a_j^r \equiv \hat{F}_A^{-1}(\frac{j}{r}), \quad j \in \{1, .., r-1\}$$

$$a_j^s \equiv \hat{F}_A^{-1}(\frac{j}{s}), \quad j \in \{1, .., s-1\}$$

- compute: $\Delta I(A) = \hat{P}_X^N(A) \cdot \hat{I}_N^{Q_s^{(k+1)}(A)}(X; Y | X \in A)$
**if** ($\Delta I(A) > \delta$) (*critical MI gain*)
$Q^{(k+1)} = Q^{(k+1)} \bigcup Q_r^{(k+1)}(A)$
**else**
$Q^{(k+1)} = Q^{(k+1)} \bigcup \{A\}$
**end, if**
**end, if**
**end, foreach**
**Phase 3**: (Termination)
**if** ($Q^{k+1}$ equal to $Q^k$)
**done**,
**else**
$k = k+1$, goto **Phase 1**,
**end, if**

Fig. 3. Darbellay–Vajda data-dependent partition algorithm for estimating the conditional mutual information.

splitting criterion. But in the negative case, the algorithm stops the refinement of this region under the assumption that condition to the event $X \in A$, $X$ and $Y$ can be considered almost independent, i.e., $\hat{I}_N^{Q^2(A)}(X; Y | X \in A) < \epsilon \Rightarrow I(X; Y | X \in A) \approx 0$. Furthermore to control estimation error, we introduce a threshold in the splitting rule to control the minimum number of training points associated with $A$, for having a good representation of the joint distribution between $X$ and $Y$ in this target region. The pseudocode is presented in Fig. 3, which considers the following set of parameters:

- $(s, r) \in \mathbb{N}^2$, $s > r$: number of splits per coordinate to partition the space in statistically equivalent sets;
- $\delta > 0$: threshold for the MI gain;
- $N_c \in \mathbb{N}$: minimum number of sample points for refinement.

Finally, in our problem we have $X_1$, $X_2$, $X_3$, and $Y$, and we need to estimate $\Delta\rho = I(X_1, X_2; Y | X_3)$, with the i.i.d. samples $\{(x_1^i, x_2^i, x_3^i, y^i) : i = 1, \ldots, N\}$. The nonparametric estimation is as follows.

1) Use the Darbellay–Vajda algorithm to construct partition $Q_{1,2}^N$ for $(X_1, X_2)$ using $\{x_1^i, x_2^i, y^i : i = 1, \ldots, N\}$.
2) Use the Darbellay–Vajda algorithm to construct partition $Q_3^N$ for $X_3$ using $\{x_3^i, y^i : i = 1, \ldots, N\}$.
3) Consider the product adaptive partition $Q_{1,2\times 3}^N$ to:
   - compute empirical joint distribution $\hat{P}_{X_1, X_2, X_3, Y}^N$ for every event in $\sigma(Q_{1,2\times 3}^N) \times \mathcal{F}_{\mathcal{Y}}$;
   - compute empirical MI indicators $\hat{I}_N^{Q_{1,2\times 3}^N}(X_1, X_2, X_3; Y)$ and $\hat{I}_N^{Q_3^N}(X_3; Y)$;
   - finally, compute the CMI estimate $\hat{\Delta}\rho_N(Q_{1,2\times 3}^N)$.

## VI. Experiments

In this section, we report experiments to evaluate: the nonparametric CMI estimator across the different scale-frequency values of the WP basis family; the solutions of the minimum cost tree pruning in terms of the expected frequency band decompositions, and the classification performance of the resulting feature descriptions in comparison with some standard feature representations.

### A. Frame-Level Phone Classification From Speech Signal

We consider an automatic speech recognition scenario, where filter banks have been widely used for feature representations and, furthermore, concrete ideas for the optimal frequency band decompositions are well understood based on perceptual studies of the human auditory system. The corpus used was collected in our group at USC and comprises about 1.5 h of spontaneous conversational speech from a male English speaker, sampled at 16 kHz. A standard frame-by-frame analysis was performed on those acoustic signals where, every 10 ms (frame rate), a segment of the acoustic signal of 64 ms around a time center position was extracted. Word-level transcriptions were used for generating phone level time segmentations on the acoustic signals by using automatic forced Viterbi alignment techniques. Using the phone-level time segmentations, the collection of those acoustic frame vectors, dimension $K = 1024$, with their corresponding phone class information (47 classes) was created, where we considered one session of the data comprising $N = 14\,979$ supervised sample points. Finally, for creating the set of feature representations, we use the Daubechies' maximally flat filter (db4) for the WP basis family [1], [41], and the energy on the resulting bands. We first present some analysis of the minimum cost tree pruning in terms of topology of those solutions (the optimal filter bank decomposition problem), and then we evaluate performances associated with those solutions.

### B. Analysis of the MI Gain and Optimal Tree Pruning

We estimated the CMI gains in (10), using the algorithm presented in Section V. We considered $s = 8$, $r = 2$ for generating the product refinement (associated with the MI gain obtained by refining the product partition), following the general recommendations suggested in [31]. We tried different configurations for $\delta$ and $N_c$, which strongly govern the tradeoff between approximation and estimation error. We conducted an exhaustive analysis of the CMI estimation obtained across those configurations observing marginal discrepancies on the relative differences of CMI estimated values across scale and frequency bands. In this respect, it is important to point out that

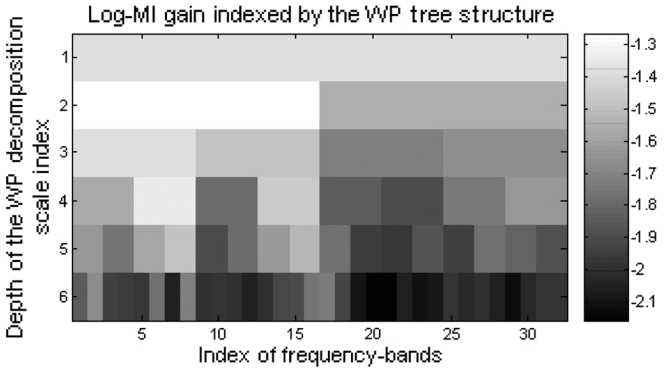


Fig. 4. Graphical representation of the CMI magnitudes, $\{\Delta\rho(l,j) : \forall l \in 1,\ldots,6, j \in 0,\ldots,2^l-1\}$, obtained by splitting the two channel block of analysis of the wavelet packet bases. The CMI magnitudes are organized across scale (level of decomposition, vertical axes) and frequency bands (horizontal axes) in the WP decomposition.

the relative differences among the CMI values $\{\Delta\rho(l,j) : \forall l \in 1,\ldots,6, j \in 0,\ldots,2^l-1\}$ fully characterize the topology of solutions of the minimum cost tree-pruning problem. This behavior can be explained because the implicit overestimation (because of estimation error) and underestimation (because of quantization) uniformly affect all the CMI estimations across scales and bands (same dimension for involved random variables and same number of samples points). For this setting, we have chosen a conservative configuration, $\delta = 1.0^{-200}$ and $N_c = 200$, to have a reasonable estimation of the class-observation distributions during the quantization process and consequently bias to an underestimation of the real CMI values.

Fig. 4 represents the CMI estimations (or MI gains) across scales and frequency bands for the WP decomposition. The global trend presented in Fig. 4 is expected, in the sense that the iteration of lower frequency bands provides more phone discrimination information than the iteration on higher frequency bands across almost all the scales of the analysis. This fact is consistent with studies of the human auditory system showing that overall there is higher discrimination for lower frequency regions than higher frequency regions in the auditory range of 55 Hz–15 KHz [28]. This global trend was also observed for all the other sessions of the corpus (not reported here), supporting the generality of results obtained from the mutual information decomposition across bands of the acoustic signals. Based on this trend the general solution of the optimal tree-pruning problem follows the expected tendency, where for a given number of bands, more level of decompositions are allocated in lower frequency components of the acoustic space. Interestingly, exact Wavelet type of filter bank solutions (the type of filter bank structure obtained from human perceptual studies, MEL scale [42]) were obtained for solutions associated with small dimensions. It is important to mention the same analysis was conducted in a synthetic setting to evaluate CMI trends across scale-frequency and solutions of the optimal filter bank decomposition. Expected trends and decompositions were obtained in terms of the discrimination of the different frequency bands of the signals, designed during the synthesis part. Results are not reported here for space considerations.

### C. Frame-Level Phone Recognition

The solutions of the cost-fidelity were used as feature representations for frame level phone classification. In particular, we evaluated solutions associated with the following dimensions: 4, 7, 10, 13, 19, 25, 31, 37, 43, 49, 55, and 61. GMMs were used for estimating class-conditional densities in the Bayes' decision setting, which is the standard parametric model adopted for this type of frame level phone classification [9], and a tenfold cross validation was used for performance evaluation. 32 mixture components per class were considered and the EM-algorithm was used for ML parameter estimation. As a reference, we consider the standard 13-dimensional Mel-Cepstrum (MFCCs) plus delta and acceleration coefficients using the same frame rate (10 ms) and window length (64 ms)—39-feature vector associated with a total window length of 100 ms, where the correct phone classification rate (mean and standard deviation) obtained was 53.01%(1.01). The performances for the minimum cost tree-pruning family using the proposed nonparametric CMI as fidelity indicator, as well as the energy considered in [10], are reported in Table I. Table I also reports performances of two widely used dimensionality reduction techniques acting on the raw time domain data: linear discriminant analysis (LDA) and nonparametric discriminant analysis (NDA). These two techniques were only used as feature extraction where the same GMM classifier setting was adopted for the performance evaluation.

LDA and NDA present relatively poor performances compared to using filter bank representations of the acoustic process. This can be attributed to two reasons: first these methods are constrained to the family of linear transformations on the raw data, and second, there is an implicit Gaussianity assumption in considering the between-within class scatter matrices ratio as the optimality criterion on both techniques [22], [34], which is not guaranteed to be valid in this particular high dimensional setting. When comparing filter bank energy solutions, in particular the minimum cost tree pruning using the proposed empirical MI and energy as the fidelity criterion in Table I, the former as expected shows consistently better performance, demonstrating the effectiveness of the empirical MI as an indicator of discrimination information. As a final corroboration of the goodness of the filter bank WP family and correctness of proposed optimality criterion, for the range of dimensions [31–43] our data-driven minimum cost tree-pruning family provide competitive performances with respect to the largely adopted 39-MFCCs. Note that 39-MFCC features are used as a benchmark because they consider higher contextual information—about 150 ms of window context, and consequently they are not directly comparable with our filter bank solutions.

In conclusion, these experiments show the importance of having on the signal processing side, a good target family of feature representations, ratifying the approximation quality of filter bank energy features for the analysis of pseudo-stationary stochastic phenomena, and on the learning side, an optimality criterion that reflects the estimation-approximation error tradeoff presented in the learning problem of pattern recognition.

TABLE I
CORRECT PHONE CLASSIFICATION (CPC) RATES (MEAN AND STANDARD DEVIATION) FOR THE MINIMUM COST TREE-PRUNING (MCTP) SOLUTIONS USING THE PROPOSED EMPIRICAL MUTUAL INFORMATION (MI) AND SOLUTIONS USING ENERGY AS FIDELITY CRITERION FOR SUBBAND SPLITTING (WP-ENERGY-DECOM). AS A REFERENCE, PERFORMANCES ARE PROVIDED FOR LINEAR DISCRIMINANT ANALYSIS (LDA) AND NONPARAMETRIC DISCRIMINANT ANALYSIS (NDA). PERFORMANCES OBTAINED USING TENFOLD CROSS VALIDATION AND A GMM-BASED CLASSIFIER

| Dimension | MCTP-MI mean (std. dev.) | WP-energy-decom mean (std. dev.) | LDA mean (std. dev.) | NDA mean (std. dev.) |
|---|---|---|---|---|
| 4 | 28.53(0.65) | 28.04(1.03) | 12.58(0.53) | 6.73(0.55) |
| 7 | 37.95(1.32) | 35.54(0.98) | 17.95(1.19) | 7.48(0.99) |
| 10 | 40.34(1.31) | 39.39(1.30) | 21.55(1.02) | 8.03(0.71) |
| 13 | 44.29(1.24) | 40.25(1.58) | 25.20(0.99) | 8.57(0.67) |
| 19 | 46.61(1.10) | 44.09(0.97) | 30.34(1.32) | 8.73(0.88) |
| 25 | 48.27(1.22) | 46.21(1.55) | 31.72(1.31) | 9.25(0.73) |
| 31 | 49.64(1.03) | 46.58(1.93) | **31.84**(1.04) | 9.37(0.76) |
| 37 | 51.11(0.97) | 47.64(1.51) | 31.62(1.03) | 9.32(0.56) |
| 43 | 52.10(1.43) | 47.25(0.99) | 31.28(1.16) | 9.50(0.55) |
| 49 | **52.98(1.52)** | **47.47(1.72)** | 29.61(1.52) | 9.80(0.74) |
| 55 | 52.87(1.23) | 47.07(1.33) | 27.60(0.93) | **10.12(0.72)** |
| 61 | 52.44(1.41) | 39.33(1.67) | 25.42(1.03) | 10.08(1.12) |

## VII. DISCUSSION AND FUTURE WORK

It is important to remind the reader that although the presented formulation is theoretically motivated by the MPE-SR, this optimization problem is practically intractable and requires the introduction of approximations, in particular concerning the Bayes' error. In this paper empirical MI is adopted for that purpose. This choice has some theoretical justification in terms of information theoretic inequalities and monotonic behavior of the indicator across sequence of embedded transformation of the data [23]; however, tightness is not guaranteed. In that respect, the presented formulation is open to considering alternative fidelity criteria. The empirical risk (ER) is a natural candidate with a strong theoretical support [14], [43]; however, the optimization problem requires an exhaustive evaluation in our alphabet of feature transformations, which for reasonable dimensions of the problem becomes impractical. Another attractive alternative is the family of Ali–Silvey distance measures, used to evaluate the effect of vector quantization in hypothesis testing problems [44], [45], or even indicators like Fisher like scatter ratios [5]. This is an interesting direction for future research, where as presented in this work additivity property of these indicators, with respect to structure of WP bases, can be studied to extend algorithmic solutions, or alternatively, greedy algorithms can be proposed and empirically evaluated, when the resulting optimal BS problem does not admit polynomial time algorithmic solutions.

Concerning the presented phone classification experiments, the proposed data-driven feature extraction offers promising results, however a systematic study of the problem still remains to be conducted to explore the full potentiality of the proposed formulation. This may include a careful design of the two-channel filter bank evaluating its impact in classification performances [9], the use of other tree-structured bases families, as well as experimental validation under more general acoustic conditions and considering a state-of-the-art time-series classification task.

## APPENDIX

### A. Proof of Proposition 1

Equation (7) is just a consequence of the Parseval relationship [1], [41] and the fact that by construction if $T \ll \bar{T}$, then $\mathcal{B}_{\bar{T}}$ is a subspace refinement of $\mathcal{B}_T$.[13] Concerning the second result, without loss of generality let us consider $T_1 \ll T_2 \ll \mathbf{T}_{\text{full}}$, where we need to show that $H(m_{T_1}(X)|m_{T_2}(X)) = 0$. Before going to the actual proof we will use the following result.

*Lemma 1:* Let us consider $T \ll \mathbf{T}_{\text{full}}$, then we have that

$$H\left(\left(X_j^l\right)_{(l,j)\in T}\right) = H\left(\left(X_j^l\right)_{(l,j)\in\mathcal{L}(T)}\right). \tag{25}$$

*Proof:* We use the fact that $H(X_j^l|X_{2j}^{l+1}, X_{2j+1}^{l+1}) = 0$, for all $(l,j) \in \mathcal{I}(\mathbf{T}_{\text{full}})$, from (7). The idea is to partition the set of nodes as a function of its depth with respect to the root $v_{\text{root}} = (0,0)$, and use the chain rule [23], [30]. Let $T = \{v_{\text{root}}\} \cup T^1 \cup \cdots \cup T^{K_o}$, where $T^i$ is the collection of nodes in $T$ with depth $i$, and $K_o$ the maximum depth of the tree (see Fig. 5(a)). In addition, let us define $\tilde{T}^i \equiv T^i \cap \mathcal{L}(T)$ and $\hat{T}^i \equiv T^i \cap \mathcal{I}(T)$ the set of terminal and internal nodes of depth $i$ of the tree, respectively. See Fig. 5(a). Note that $T^{K_o} = \tilde{T}^{K_o}$ and that $\mathcal{L}(T) = \bigcup_{k=1}^{K_o} \tilde{T}^k$. By the tree structure, $\forall k \in \{1,\ldots,K_o-1\}$, and $\forall (k,j) \in \hat{T}^k$, we have that $(k+1,2j)$ and $(k+1,2j+1)$ belong to $T^{k+1} = \tilde{T}^{k+1} \cup \hat{T}^{k+1}$. We will use this node depth dependent partition of $T$ for the following derivations. In particular, considering that $T = (\cup_{k=1}^{K_o-1} \hat{T}^k \cup \{v_{\text{root}}\}) \bigcup (\cup_{k=1}^{K_o} \tilde{T}^i)$ we have that

$$H\left(\left(X_j^l\right)_{(l,j)\in T}\right) = H\left(\left(X_j^l\right)_{(l,j)\in\cup_{k=1}^{K_o}\tilde{T}^k}\right) + H\left(\left(X_j^l\right)_{(l,j)\in\cup_{k=1}^{K_o-1}\hat{T}^k\cup\{v_{\text{root}}\}} \,|\, \left(X_j^l\right)_{(l,j)\in\cup_{k=1}^{K_o}\tilde{T}^k}\right). \tag{26}$$

[13] $\mathcal{B}_{\mathcal{T}}$ is a subspace refinement of $\mathcal{B}_T$, in the sense that for any subspace $\mathcal{X}_j^l$, $(l,j) \in \mathcal{L}(T)$, $\exists \hat{\mathcal{L}} \subset \mathcal{L}(\bar{T})$ such that $\mathcal{X}_j^l = \bigoplus_{(l,j)\in\hat{\mathcal{L}}} \mathcal{X}_j^{\hat{l}}$.

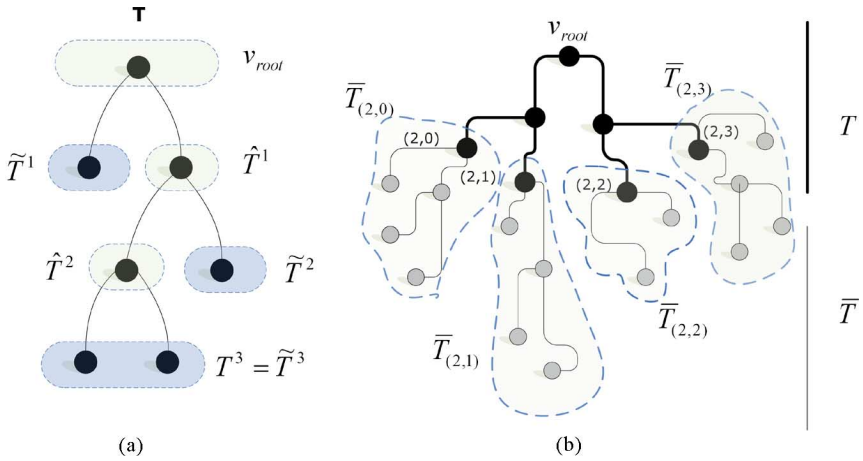


Fig. 5. Example of the notation and topology of a tree-indexed WP representation.

Hence, for proving (25), we only need to show that the last right term on (26) is equal to zero. Using the chain rule, we have that

$$\begin{aligned} H&\left(\left(X_j^l\right)_{(l,j)\in\cup_{k=1}^{K_o-1}\hat{T}^k\cup\{v_{\rm root}\}}\,|\,\left(X_j^l\right)_{(l,j)\in\cup_{k=1}^{K_o}\tilde{T}^k}\right)\\ &= H\left(\left(X_j^l\right)_{(l,j)\in\hat{T}^{K_o-1}}\,|\,\left(X_j^l\right)_{(l,j)\in\cup_{k=1}^{K_o}\tilde{T}^k}\right)\\ &\quad + H\left(\left(X_j^l\right)_{(l,j)\in\hat{T}^{K_o-2}}\,|\,\left(X_j^l\right)_{(l,j)\in\cup_{k=1}^{K_o}\tilde{T}^k},\right.\\ &\qquad\left.\left(X_j^l\right)_{(l,j)\in\hat{T}^{K_o-1}}\right)+\cdots\\ &\quad + H\left(\left(X_j^l\right)_{(l,j)\in\{v_{\rm root}\}}\,|\,\left(X_j^l\right)_{(l,j)\in\cup_{k=1}^{K_o}\tilde{T}^k},\right.\\ &\qquad\left.\left(X_j^l\right)_{(l,j)\in\cup_{k=1}^{K_o-1}\hat{T}^k}\right). \end{aligned} \tag{27}$$

Let us analyze one of the generic terms of (27), say $i\in\{1,\ldots,K_o-2\}$. By chain rule, we have that

$$\begin{aligned} H&\left(\left(X_j^l\right)_{(l,j)\in\hat{T}^i}\,|\,\left(X_j^l\right)_{(l,j)\in\cup_{k=1}^{K_o}\tilde{T}^k},\left(X_j^l\right)_{(l,j)\in\cup_{k=i+1}^{K_o-1}\hat{T}^k}\right)\\ &\le H\left(\left(X_j^l\right)_{(l,j)\in\hat{T}^i}\,|\,\left(X_j^l\right)_{(l,j)\in\tilde{T}^{i+1}\cup\hat{T}^{i+1}}\right). \end{aligned} \tag{28}$$

Enumerating $\left(X_j^l\right)_{(l,j)\in\hat{T}^i}$ by the sequence $(X_{j1}^i,X_{j2}^i,\ldots,X_{jN_i}^i)$, where $\{j1,j2,\ldots,jN_i\}\subset\{0,\ldots,2^i-1\}$ and considering the notation $\bar{X}_j^{l+1}=(X_{2j}^{l+1},X_{2j+1}^{l+1})$, the inequality in (28) is equivalent to

$$\begin{aligned} &H\left(\left(X_{j1}^i,X_{j2}^i,\ldots,X_{jN_i}^i\right)\,|\,\left(X_j^l\right)_{(l,j)\in\tilde{T}^{i+1}\cup\hat{T}^{i+1}}\right)\\ &\le H\left(\left(X_{j1}^i,X_{j2}^i,\ldots,X_{jN_i}^i\right)|\left(\bar{X}_{j1}^{i+1},\bar{X}_{j2}^{i+1},\ldots,\bar{X}_{jN_i}^{i+1}\right)\right)\\ &\le \sum_{j\in\{j1,j2,\ldots,jN_i\}} H\left(X_j^i|\bar{X}_j^{i+1}\right)=0. \end{aligned} \tag{29}$$

The first inequality is because of the fact that $\{(i+1,2\cdot j1),(i+1,2\cdot j1+1),\ldots,(i+1,2\cdot jN_i+1)\}\subset(\tilde{T}^{i+1}\cup\hat{T}^{i+1})$ and the chain rule, and the last equality by the hypothesis. The same derivations can be extended for all the terms on (27), which proves the lemma. □

Returning to our problem, let us consider

$$\begin{aligned} H&\left(\left(X_j^l\right)_{(l,j)\in\mathcal{L}(T_1)\cup\mathcal{L}(T_2)}\right)=H\left(\left(X_j^l\right)_{(l,j)\in\mathcal{L}(T_2)}\right)\\ &\quad+H\left(\left(X_j^l\right)_{(l,j)\in\mathcal{L}(T_1)}\,|\,\left(X_j^l\right)_{(l,j)\in\mathcal{L}(T_2)}\right) \end{aligned} \tag{30}$$

where given that $\mathcal{L}(T_1)\cup\mathcal{L}(T_2)\subset T_2$, by Lemma 1 we have that, $H((X_j^l)_{(l,j)\in\mathcal{L}(T_1)\cup\mathcal{L}(T_2)})\le H((X_j^l)_{(l,j)\in T_2})=H((X_j^l)_{(l,j)\in\mathcal{L}(T_2)})$. This last inequality in conjunction with (30) proves the result. □

### B. Proof of Proposition 2

*Proof:* Let us start by considering $T\ll T_{lo,jo}^+\ll\mathbf{T}_{\rm full}$, where $T_{lo,jo}^+$ denotes the tree induced from $T$ by splitting one of its terminal nodes, $(lo,jo)\in\mathcal{L}(T)$. By definition we have that $m_T(X)=(X_j^l)_{(l,j)\in\mathcal{L}(T)}$ and $m_{T_{lo,jo}^+}(X)=(X_j^l)_{(l,j)\in\mathcal{L}(T_{lo,jo}^+)}$ with $\mathcal{L}(T_{lo,jo}^+)=\{(lo+1,2jo),(lo+1,2jo+1)\}\cup\mathcal{L}(T)\setminus\{(lo,jo)\}$. By multiple application of the chain rule, it follows that

$$\begin{aligned} &I\left(m_{T_{lo,jo}^+}(X);Y\right)-I\left(m_T(X);Y\right)\\ &\quad=I\left(\left(X_j^l\right)_{(l,j)\in\mathcal{L}\left(T_{lo,jo}^+\right)};Y\right)-I\left(\left(X_j^l\right)_{(l,j)\in\mathcal{L}(T)};Y\right)\\ &\quad=I\left(\left(X_j^l\right)_{(l,j)\in\mathcal{L}\left(T_{lo,jo}^+\right)\cup\{(lo,jo)\}};Y\right)\\ &\qquad-I\left(X_{jo}^{lo};Y|\left(X_j^l\right)_{(l,j)\in\mathcal{L}\left(T_{lo,jo}^+\right)}\right)\\ &\qquad-I\left(\left(X_j^l\right)_{(l,j)\in\mathcal{L}(T)};Y\right)\\ &\quad=I\left(\left(X_j^l\right)_{(l,j)\in\mathcal{L}(T)\cup\{(lo+1,2jo),(lo+1,2jo+1)\}};Y\right)\\ &\qquad-I\left(\left(X_j^l\right)_{(l,j)\in\mathcal{L}(T)};Y\right)\\ &\qquad-I\left(X_{jo}^{lo};Y|\left(X_j^l\right)_{(l,j)\in\mathcal{L}\left(T_{lo,jo}^+\right)}\right)\\ &\quad=I\left(\bar{X}_{jo}^{lo+1};Y|\left(X_j^l\right)_{(l,j)\in\mathcal{L}(T)}\right)\\ &\qquad-I\left(X_{jo}^{lo};Y|\left(X_j^l\right)_{(l,j)\in\mathcal{L}\left(T_{lo,jo}^+\right)}\right). \end{aligned} \tag{31}$$

Finally noting that $0 \leq I(X_{jo}^{lo};Y|(X_j^l)_{(l,j)\in\mathcal{L}(T^+_{lo,jo})}) \leq H(X_{jo}^{lo}|(X_j^l)_{(l,j)\in\mathcal{L}(T^+_{lo,jo})}) \leq H(X_{jo}^{lo}|\bar{X}_{jo}^{lo+1}) = 0$, by definition of the CMI and the chain rule, we get that

$$I\left(m_{T^+_{lo,jo}}(X);Y\right) - I\left(m_T(X);Y\right) = I\left(\bar{X}_{jo}^{lo+1};Y|\left(X_j^l\right)_{(l,j)\in\mathcal{L}(T)}\right) \quad (32)$$

which proves the result for this particular case. For the general case $T \ll \bar{T}$ we can consider one of the possible sequence of internal nodes $\{(l1,j1),\ldots,(ln,jn)\}$, which needs to be split to go from $T$ to $\bar{T}$. More precisely, we can consider the sequence of embedded trees $T = T_0 \ll T_1 \ll \cdots \ll T_n = \bar{T}$, such that $T_i = (T_{i-1})^+_{li,ji}$ $\forall i \in \{1,\ldots,n\}$. Using telescope series expansion and (32)

$$\begin{aligned}
&I\left(m_{T_n}(X);Y\right) - I\left(m_{T_0}(X);Y\right) \\
&= \sum_{i=1}^{n} I\left(m_{T_i}(X);Y\right) - I\left(m_{T_{i-1}}(X);Y\right) \\
&= \sum_{i=1}^{n} I\left(\bar{X}_{ji}^{li+1};Y|\left(X_j^l\right)_{(l,j)\in\mathcal{L}(T_{i-1})}\right) \quad (33)\\
&= \sum_{i=1}^{n} I\left(\bar{X}_{ji}^{li+1};Y|\left(X_j^l\right)_{(l,j)\in T_{i-1}}\right) \quad (34)\\
&= I\left(\bar{X}_{j1}^{l1+1},\cdots,\bar{X}_{jn}^{ln+1};Y|\left(X_j^l\right)_{(l,j)\in T_0}\right) \\
&= I\left(\bar{X}_{j1}^{l1+1},\cdots,\bar{X}_{jn}^{ln+1};Y|\left(X_j^l\right)_{(l,j)\in\mathcal{L}(T)}\right) \quad (35)\\
&= I\left(\left(X_j^l\right)_{(l,j)\in\bar{T}\setminus T};Y|\left(X_j^l\right)_{(l,j)\in\mathcal{L}(T)}\right) \\
&= I\left(\left(X_j^l\right)_{(l,j)\in\bar{T}\setminus T};Y|m_T(X)\right). \quad (36)
\end{aligned}$$

Equation (33) is because of the chain rule, and (36) by construction, where we have that $\bar{T}\setminus T = \{(l1+1,2\cdot j1),(l1+1,2\cdot j1+1),\ldots,(ln+1,2\cdot jn+1)\}$. The equalities involving interchanging $(X_j^l)_{(l,j)\in T}$ by $(X_j^l)_{(l,j)\in\mathcal{L}(T)}$ in (34) and (35) are a direct consequence of Lemma 1. □

### C. Additive Property of the Mutual Information Tree Functional $\rho(\cdot)$: Theorem 1

*Proof:* We have that $T \ll \bar{T}$. As in the proof presented in Appendix B, we can consider a sequence of internal nodes $\{(l1,j1),\ldots,(ln,jn)\}$, and the sequence of embedded trees $T = T_0 \ll T_1 \ll \cdots \ll T_n = \bar{T}$, such that $T_i = (T_{i-1})^+_{li,ji}$, $\forall i \in \{1,\ldots,n\}$. From the first equality of (34) we have that

$$\begin{aligned}
\rho(\bar{T}) - \rho(T) &= \sum_{i=1}^{n} I\left(\bar{X}_{ji}^{li+1};Y|\left(X_j^l\right)_{(l,j)\in\mathcal{L}(T_{i-1})}\right) \\
&= \sum_{i=1}^{n} I\left(\bar{X}_{ji}^{li+1};Y|X_{ji}^{li}\right) = \sum_{i=1}^{n}\Delta\rho(li,ji) \quad (37)
\end{aligned}$$

where the equalities in (37) is the result of the Markov tree property. Using the fact that $\{(l1,j1),\ldots,(ln,jn)\} = \mathcal{I}(\bar{T})\setminus\mathcal{I}(T)$, (37) shows the first results of the theorem in (12). For proving the next expression in (13), we start with the result presented in Proposition 2, where $\rho(\bar{T}) - \rho(T) = I((X_j^l)_{(l,j)\in\bar{T}\setminus T};Y|m_T(X))$. Using that $m_T(X) = (X_j^l)_{(l,j)\in\mathcal{L}(T)}$ and the chain rule, it directs to show that

$$\begin{aligned}
\rho(\bar{T}) - \rho(T) &= I\left(\left(X_j^l\right)_{(l,j)\in\bar{T}\setminus T\cup\mathcal{L}(T)};Y|m_T(X)\right) \\
&= I\left(\left(X_j^l\right)_{(l,j)\in\bar{T}\setminus\mathcal{I}(T)};Y|m_T(X)\right) \quad (38)
\end{aligned}$$

where noting that $\bar{T}\setminus\mathcal{I}(T) = \bigcup_{(\bar{l},\bar{j})\in\mathcal{L}(T)}\bar{T}_{(\bar{l},\bar{j})}$ [see Fig. 5(b)], we get that

$$\rho(\bar{T}) - \rho(T) = I\left(\left(X_j^l\right)_{(l,j)\in\bigcup_{(\bar{l},\bar{j})\in\mathcal{L}(T)}\bar{T}_{(\bar{l},\bar{j})}};Y|\left(X_j^l\right)_{(l,j)\in\mathcal{L}(T)}\right). \quad (39)$$

Finally from (39) by using the chain rule for CMI, and the conditional independence assumption stated in Proposition 3, it is simple to show that

$$\begin{aligned}
\rho(\bar{T}) - \rho(T) &= \sum_{(\bar{l},\bar{j})\in\mathcal{L}(T)} I\left(\left(X_j^l\right)_{(l,j)\in\bar{T}_{(\bar{l},\bar{j})}};Y|X_{\bar{j}}^{\bar{l}}\right) \quad (40)\\
&= \sum_{(\bar{l},\bar{j})\in\mathcal{L}(T)} I\left(\left(X_j^l\right)_{(l,j)\in\mathcal{L}\left(\bar{T}_{(\bar{l},\bar{j})}\right)};Y|X_{\bar{j}}^{\bar{l}}\right) \\
&= \sum_{(\bar{l},\bar{j})\in\mathcal{L}(T)} I\left(m_{\bar{T}_{(\bar{l},\bar{j})}}(X);Y|X_{\bar{j}}^{\bar{l}}\right) \quad (41)
\end{aligned}$$

which from the definition of $\rho_T(l,j)$ proves (13). Finally for proving the last expression in (14), we just need to consider the trivial tree $\{(0,0)\}$ and $T \ll \mathbf{T}_{\text{full}}$. It is clear that $\{(0,0)\} \ll T$ and from (37)

$$\rho(T) = \rho\left(\{(0,0)\}\right) + \sum_{(l,j)\in\mathcal{I}(T)\setminus\mathcal{I}(\{(0,0)\})}\Delta\rho(l,j) \quad (42)$$

where given that $\mathcal{I}(\{(0,0)\}) = \phi$ and that $\rho(\{(0,0)\}) = I(X_0^0;Y)$, we get the result.

### D. Proof of Proposition 4

*Proof:* For proving (15) by definition

$$\begin{aligned}
\rho_T(l,j) &= I(m_{T_{(l,j)}}(X);Y) - I(X_j^l;Y) \\
&= I((X_{\bar{j}}^{\bar{l}})_{(\bar{l},\bar{j})\in T_{(l,j)}\setminus\{(l,j)\}};Y|X_j^l)
\end{aligned}$$

where the second equality is because of Proposition 2. Using the binary structure of $T$, it follows that $T_{(l,j)}\setminus\{(l,j)\} =$

$$\begin{aligned}\max_{\substack{T\ll\mathbf{T}_v\\|T|=k}}\rho_T(v)&=\Delta\rho(v)+\max_{\substack{T=\left[v,T_{l(v)},T_{r(v)}\right]T_{l(v)}\ll\mathbf{T}_{l(v)},T_{r(v)}\ll\mathbf{T}_{r(v)}\\|T_{l(v)}|+|T_{r(v)}|=k}}\left[\rho_T\left(l(v)\right)+\rho_T\left(r(v)\right)\right]\\&=\Delta\rho(v)+\max_{\substack{k_1\geq1,k_2\geq1\\k_1+k_2=k}}\left[\max_{\substack{T\ll\mathbf{T}_{l(v)}\\|T|=k_1}}\left[\rho_T\left(l(v)\right)\right]+\max_{\substack{T\ll\mathbf{T}_{r(v)}\\|T|=k_2}}\left[\rho_T\left(r(v)\right)\right]\right]\\&=\Delta\rho(v)+\max_{\substack{|\mathbf{T}_{l(v)}|\geq k_1\geq1,|\mathbf{T}_{r(v)}|\geq k_2\geq1\\k_1+k_2=k}}\left[\rho_{\mathbf{T}_{l(v)}^{k_1*}}\left(l(v)\right)+\rho_{\mathbf{T}_{r(v)}^{k_2*}}\left(r(v)\right)\right].\end{aligned}\tag{44}$$

$T_{(l+1,2j)}\cup T_{(l+1,2j+1)}$. Hence, considering the notation $\bar{X}_j^{l+1}=(X_{2j}^{l+1},X_{2j+1}^{l+1})$ we have that

$$\begin{aligned}\rho_T(l,j)&=I\left(\left(X_{\bar{j}}^{\bar{l}}\right)_{(\bar{l},\bar{j})\in T_{(l+1,2j)}\cup T_{(l+1,2j+1)}};Y|X_j^l\right)\\&=I\left(\bar{X}_j^l;Y|X_j^l\right)\\&\quad+I\left(\left(X_{\bar{j}}^{\bar{l}}\right)_{(\bar{l},\bar{j})\in T_{(l+1,2j)}};Y|X_j^l,\bar{X}_j^l\right)\\&\quad+I\left(\left(X_{\bar{j}}^{\bar{l}}\right)_{(\bar{l},\bar{j})\in T_{(l+1,2j+1)}};\right.\\&\qquad\left.Y|X_j^l,\bar{X}_j^l,\left(X_{\bar{j}}^{\bar{l}}\right)_{(\bar{l},\bar{j})\in T_{(l+1,2j)}}\right)\\&=\Delta\rho(l,j)+I\left(\left(X_{\bar{j}}^{\bar{l}}\right)_{(\bar{l},\bar{j})\in T_{(l+1,2j)}};Y|X_{2j}^{l+1}\right)\\&\quad+I\left(\left(X_{\bar{j}}^{\bar{l}}\right)_{(\bar{l},\bar{j})\in T_{(l+1,2j+1)}};Y|X_{2j+1}^{l+1}\right)\\&=\Delta\rho(l,j)+I\left(m_{T_{(l+1,2j)}}(X);Y|X_{2j}^{l+1}\right)\\&\quad+I\left(m_{T_{(l+1,2j+1)}}(X);Y|X_{2j+1}^{l+1}\right).\end{aligned}$$

The second equality is because of the chain rule for the CMI [23], the third by the Markov tree property, and the last a direct consequence of Proposition 1 (see Lemma 1 in Appendix A for details), which proves the result.

### *E. Dynamic Programming Solution for the Optimal Tree-Pruning Problem: Theorem 2*

*Proof:* Let us consider $v\in\mathcal{I}(\mathbf{T}_{\text{full}})$, we want to find the solution for

$$\mathbf{T}_v^{k*}=\arg\max_{\substack{T\ll\mathbf{T}_v\\|T|=k}}\rho_T(v)\tag{43}$$

as a function of solutions of its direct descendants—$l(v)$ and $r(v)$—which are assumed to be known, $\{\mathbf{T}_{l(v)}^{k_1*},\mathbf{T}_{r(v)}^{k_2*}:k_1=1,\ldots,|\mathbf{T}_{l(v)}|\ k_2=1,\ldots,|\mathbf{T}_{r(v)}|\}$. Let us consider the nontrivial case $k>1$, and an arbitrary tree $T\ll\mathbf{T}_v$ such that $|T|=k$. Then, $l(v)\in T$ and $r(v)\in T$ and by Proposition 4 it follows that, $\rho_T(v)=\Delta\rho(v)+\rho_T(l(v))+\rho_T(r(v))$, where in addition if we denote $T$ by $[v,T_{l(v)},T_{r(v)}]$, then by definition $|T|=|T_{l(v)}|+|T_{r(v)}|$, and $T\ll\mathbf{T}_v$ is equivalent to $T_{l(v)}\ll\mathbf{T}_{l(v)}$ and $T_{r(v)}\ll\mathbf{T}_{r(v)}$. Consequently analyzing (43), it follows that [see (44) shown at the top of the page]. The last equality is direct from the definition of the optimal pruning tree, (43). Finally, from (44) we have that $\mathbf{T}_v^{k*}$ can be represented by $[v,\mathbf{T}_{l(v)}^{\hat{k}_1*},\mathbf{T}_{r(v)}^{\hat{k}_2*}]$, being $(\hat{k}_1,\hat{k}_2)$ the solution of

$$\max_{\substack{(k_1,k_2)\in\{1,\ldots,|\mathbf{T}_{l(v)}|\}\times\{1,\ldots,|\mathbf{T}_{r(v)}|\}\\k_1+k_2=k}}\rho_{\mathbf{T}_{l(v)}^{k_1*}}\left(l(v)\right)+\rho_{\mathbf{T}_{r(v)}^{k_2*}}\left(r(v)\right).$$

□

## REFERENCES


[1] M. Vetterli and J. Kovacevic, *Wavelet and Subband Coding*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[2] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, pp. 674–693, Jul. 1989.

[3] R. Coifman, Y. Meyer, S. Quake, and V. Wickerhauser, "Signal processing and compression with wavelet packets," Numerical Algorithms Research Group, Yale Univ., New Haven, CT, Tech. Rep., 1990.

[4] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 46, no. 4, pp. 886–902, Apr. 1998.

[5] K. Etemad and R. Chellapa, "Separability-based multiscale basis selection and feature extraction for signal and image classification," *IEEE Trans. Image Process.*, vol. 7, no. 10, pp. 1453–1465, Oct. 1998.

[6] K. Ramchandran, M. Vetterli, and C. Herley, "Wavelet, subband coding, and best bases," *Proc. IEEE*, vol. 84, no. 4, pp. 541–560, Apr. 1996.

[7] N. Vasconcelos, "Minimum probability of error image retrieval," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2322–2336, Aug. 2004.

[8] A. S. Willsky, "Multiresolution Markov models for signal and image processing," *Proc. IEEE*, vol. 90, no. 8, pp. 1396–1458, Aug. 2002.

[9] G. F. Choueiter and J. R. Glass, "An implementation of rational wavelets and filter design for phonetic classification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 939–948, Mar. 2007.

[10] T. Chang and C. J. Kuo, "Texture analysis and classification with tree-structured wavelet transform," *IEEE Trans. Image Process.*, vol. 2, no. 4, pp. 429–441, 1993.

[11] R. E. Learned, W. Karl, and A. S. Willsky, "Wavelet packet based transient signal classification," in *Proc. IEEE Conf. Time Scale Time Frequency Analysis*, 1992, pp. 109–112.

[12] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithm for best basis selection," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 713–718, Mar. 1992.

[13] N. Saito and R. R. Coifman, "Local discriminant basis," in *Proc. SPIE 2303, Mathematical Imaging: Wavelet Applications Signal Image Processing*, Jul. 1994, pp. 2–14.

[14] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.

[15] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.

[16] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1983.

[17] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, pp. 252–264, Mar. 1991.

[18] N. A. Schmid and J. A. O'Sullivan, "Thresholding method for dimensionality reduction in recognition system," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 2903–2920, Nov. 2001.

[19] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[20] L. O. Jimenez and D. A. Landgrebe, "Hyperspectral data analysis and supervised feature reduction via projection pursuit," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 6, pp. 2653–2667, Nov. 1999.
[21] S. Kumar, J. Ghosh, and M. M. Crawford, "Best-bases feature extraction algorithms for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1368–1379, Jul. 2001.
[22] J. Silva and S. Narayanan, "Minimum probability of error signal representation," in *Proc. IEEE Int. Workshop Machine Learning for Signal Processing*, Thessaloniki, Greece, 2007, pp. 348–353.
[23] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley Interscience, 1991.
[24] P. Chou, T. Lookabaugh, and R. Gray, "Optimal pruning with applications to tree-structure source coding and modeling," *IEEE Trans. Inf. Theory*, vol. 35, no. 2, pp. 299–315, 1989.
[25] C. Scott, "Tree pruning with subadditive penalties," *IEEE Trans. Signal Process.*, vol. 53, no. 12, pp. 4518–4525, Dec. 2005.
[26] A. B. Nobel, "Analysis of a complexity-based pruning scheme for classification tree," *IEEE Trans. Inf. Theory*, vol. 48, no. 8, pp. 2362–2368, Aug. 2002.
[27] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1958.
[28] T. F. Quatieri, *Discrete-Time Speech Signal Processing Principles and Practice*. Englewood Cliffs, NJ: Prentice-Hall, 2002.
[29] A. Laine and J. Fan, "Texture classification by wavelet packet signatures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 11, pp. 1186–1191, 1993.
[30] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1990.
[31] G. A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partition of the observation space," *IEEE Trans. Inf. Theory*, vol. 45, no. 4, pp. 1315–1321, 1999.
[32] R. Gray and L. D. Davisson, *Introduction to Statistical Signal Processing*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[33] P. R. Halmos, *Measure Theory*. New York: Van Nostrand, 1950.
[34] M. Padmanabhan and S. Dharanipragada, "Maximizing information content in feature extraction," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 512–519, Jul. 2005.
[35] A. K. Soman and P. P. Vaidyanathan, "On orthonormal wavelet and paraunitary filter banks," *IEEE Trans. Signal Process.*, vol. 41, no. 3, pp. 1170–1183, Mar. 1993.
[36] T. Cormen, C. Leiserson, and R. L. Rivest, *Introduction to Algorithms*. Cambridge, MA: MIT Press, 1990.
[37] D. L. Donoho, M. Vetterli, R. A. DeVore, and I. Daubechies, "Data compression and harmonic analysis," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2435–2476, 1998.
[38] M. Bohanec and I. Bratko, "Trading accuracy for simplicity in decision trees," *Mach. Learn.*, vol. 15, pp. 223–250, 1994.
[39] S. Varadhan, *Probability Theory*. Providence, RI: Amer. Math. Soc., 2001.
[40] L. Breiman, *Probability*. Reading, MA: Addison-Wesley, 1968.
[41] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA: SIAM, 1992.
[42] X. Yang, K. Wang, and S. A. Shamma, "Auditory representation of acoustic signals," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 824–839, Mar. 1992.
[43] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1999.
[44] H. V. Poor and J. B. Thomas, "Applications of ali-silvey distance measures in the design of generalized quantizers for binary decision problems," *IEEE Trans. Commun.*, vol. COM-25, no. 9, pp. 893–900, 1977.
[45] A. Jain, P. Moulin, M. I. Miller, and K. Ramchandran, "Information-theoretic bounds on target recognition performances based on degraded image data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1153–1166, 2002.

**Jorge Silva** (S'06) received the Master's of Science degree and Ph.D. degree in electrical engineering from the University of Southern California (USC) in 2005 and 2008, respectively.

He is an Assistant Professor at the Electrical Engineering Department, University of Chile.

He was a Research Assistant at the Signal Analysis and Interpretation Laboratory (SAIL) at USC from 2003 to 2008 and was also research intern at the Speech Research Group, Microsoft Corporation, Redmond, WA, during summer 2005. His current research interests include: optimal signal representation for pattern recognition; speech recognition; vector quantization for lossy compression and statistical learning; tree-structured representations (Wavelet Packets) for inference and decision.

Dr. Silva is a member of the IEEE Signal Processing and Information Theory societies and he has participated as a reviewer in various IEEE publications on signal processing. He is recipient of the Viterbi Doctoral Fellowship 2007–2008 and the Simon Ramo Scholarship 2007–2008 at USC.

**Shrikanth S. Narayanan** (F'09) received the Ph.D. degree in electrical engineering from the University of California at Los Angeles (UCLA) in 1995.

He was previously with AT&T Bell Labs and AT&T Research, first as a Senior Member, and later as a Principal member, of its Technical Staff from 1995 to 2000. He is currently Andrew J. Viterbi Professor of Engineering at the University of Southern California (USC), Los Angeles, and holds appointments as Professor of Electrical Engineering and jointly as Professor in Computer Science, Linguistics and Psychology. He is a member of the Signal and Image Processing Institute and directs the Speech Analysis and Interpretation Laboratory. He has published over 300 papers and has 15 granted/pending U.S. patents.

Dr. Narayanan has been an Editor for the *Computer Speech and Language Journal* since 2007. He is an Associate Editor for the *IEEE Signal Processing Magazine* and the IEEE TRANSACTIONS ON MULTIMEDIA. He was also an Associate Editor of the IEEE TRANSACTIONS OF SPEECH AND AUDIO PROCESSING from 2000 to 2004. He served on the Speech Processing technical committee from 2003 to 2007 and Multimedia Signal Processing technical committees from 2004 to 2008 of the IEEE Signal Processing Society and has served on the Speech Communication committee of the Acoustical Society of America since 2003 and the Advisory Council of the International Speech Communication Association. He has served on several program committees and is a Technical Program Chair for the upcoming 2009 NAACL HLT and 2009 IEEE ASRU. He is a Fellow of the Acoustical Society of America, and a member of Tau Beta Pi, Phi Kappa Phi, and Eta Kappa Nu. He is a recipient of an NSF CAREER award, USC Engineering Junior Research Award, USC Electrical Engineering Northrop Grumman Research Award, a Provost fellowship from the USC Center for Interdisciplinary research, a Mellon Award for Excellence in Mentoring, an IBM Faculty award, an Okawa Research award, and a 2005 Best Paper award from the IEEE Signal Processing society (with A. Potamianos). Papers he has coauthored with his students have won best paper awards at ICSLP'02, ICASSP'05, MMSP'06, and MMSP'07.