OPTIMAL WAVELET PACKETS DECOMPOSITION BASED ON A RATE-DISTORTION OPTIMALITY CRITERION

Jorge Silva and Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory, *http://sail.usc.edu*

University of Southern California, Viterbi School of Engineering

jorgesil@usc.edu, shri@sipi.usc.edu

ABSTRACT

We address the problem of optimal decomposition of Wavelet Packets (WPs) for pattern recognition based on the minimum probability of error signal representation (MPE-SR) principle. The problem is formulated as a complexity regularized optimization, where the tree-indexed structure of the WP family is used to reduce it to a type of minimum cost tree pruning problem used in regression and classification trees (CART). MPE-SR solutions are obtained for a frame level phone recognition task showing promising performance results.

Index Terms— Signal representation for classification, basis selection, complexity regularization, Wavelet packets, minimum cost tree pruning.

1. INTRODUCTION

Optimal signal representation is a fundamental problem in signal processing, and has been addressed from different conceptual point of views and in multiple research areas. In the context of pattern recognition, signal representation issues are naturally associated with feature extraction (FE). In contrast to compression and denoising scenarios, in recognition we are looking for representations that capture an unobserved phenomena that need to be inferred from the observed signal — decision rule — where typically a criterion reflecting the average risk of taking the mentioned decision is optimized

In this direction, Vasconcelos [1] has recently formalized the *minimum probability of error signal representation* (MPE-SR) principle. Under certain conditions, [1] formalizes a tradeoff between the *Bayes error bound* (quality of the representation space) and an information theoretic indicator for the estimation error across a sequence of embedded representations of increasing dimensionality, and finally connects this result with the notion of optimal signal representation for pattern recognition. In [4] these results were extended for a more general theoretical setting introducing the important notion of family of *consistent distributions* associated to an embedded sequence of representations. Furthermore [4] addresses the MPE-SR problem as the solution of an equivalent operational rate-distortion problem, motivated by a similar tradeoff presented in the problem of lossy compression involving a fidelity criterion [5].

In this work we extend the rate-distortion optimality criterion for the family of feature representations induced by the filter-bank structure of the Wavelet packets (WPs). The solution of this problem reduces to finding the minimum probability of error (MPE) filter bank decomposition for the observation phenomenon, which implicitly provides a way for finding the optimal time-frequency or space-frequency resolution for a given classification task. The rest of the paper is organized as follows. We begin by introducing the family of signal representations induced by the WPs and their tree-indexed representation. Section 3 formally presents the MPE-SR problem and how this problem can be addressed as the solution of a *minimum cost tree pruning problem*. Finally Section 4 presents experimental evaluation of MPE-SR solutions in a frame level phone classification task.

2. TREE-INDEXED FILTER BANK REPRESENTATION: WAVELET PACKETS

WPs allow decomposing the observation space into subspaces associated with different frequency bands [2], which has been shown to be an attractive analysis scheme for pseudo-stationary time series phenomena, such as the acoustic speech process. This bases family has a strong hierarchical tree-indexed structure induced by its filter bank implementation, that recursively iterates a two channel orthonormal filter bank (high and low frequency filters) to generate a family of orthonormal bases for $\mathcal{L}_2(\mathbb{R})$, the family of WPs [2].

Let $\mathcal{X} = \mathbb{R}^{\mathbb{Z}}$ be the raw sequence observation space, then the application of the basic two channel filter bank is equivalent to decomposing $\mathcal{X}$ into two subspaces $\mathcal{X}_0^1$ and $\mathcal{X}_1^1$ associated with two frequency bands of $\mathcal{X}$. This process induces an indexed-orthonormal basis $\mathcal{B} = \left\{ \psi_{0,k1}^1, \psi_{1,k2}^1 : k_1 \in \mathcal{A}_0^1, k_2 \in \mathcal{A}_1^1 \right\}$, where we have that

$$\mathcal{X} = \mathcal{X}_0^1 \bigoplus \mathcal{X}_1^1, \tag{1}$$

being $\mathcal{X}_i^1 = span\left\{ \psi_{i,k}^1 : k \in \mathcal{A}_i^1 \right\}$, $i \in \{0,1\}$. Associ-

This material is based upon work supported by awards from National Science Foundation, ONR-MURI, DARPA and the U.S. Army.

ated with the subspace decomposition of the indexed basis $\mathcal{B}$, we consider a measurement phase applied in every subspace. This reflects the feature extraction part (for instance, the subspace energy). In any of the subspaces, $\mathcal{X}_0^1$ and $\mathcal{X}_1^1$, we can reapply the basic analysis block to generate a new index basis. By iterating this process, it is possible to construct a tree-indexed collection of bases and their measurements for $\mathcal{X}$, see Fig. 1.B for notation. Consequently, there is a one-to-one mapping between the family of *rooted binary trees* in a certain graph $G = (E, V)$, Fig. 1.A, and the family of bases induced by the WPs.

In this context, instead of representing the tree as a collection of arcs in G, we use the convention proposed by Breiman *et al.* [3], where sub-graphs are represented just by subset of nodes of the full graph. We consider a rooted binary tree $T_{v_0} = \{v_0, v_1, ...\}$ as collection of nodes rooted at v_0. This tree has only one node with degree 2, the root node, and the rest with degree 3 and 1, for the internal node and leaf nodes, respectively. We define $\mathcal{L}(T)$ as the leafs of T and $\mathcal{I}(T)$ the internal nodes T. We say that a rooted binary tree S is a *subtree* of T if $S \subset T$. In the previous definition, if the root of S and T are the same and $\mathcal{L}(S) \neq \mathcal{L}(T)$, 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, and $\mathcal{L}(S) \subset \mathcal{L}(T)$, then S is called a *branch* of T. In particular, we denote T_v the branch of T rooted at $v \in T$. We define the size of the tree T as the cardinality of $\mathcal{L}(T)$, the number of terminal nodes, and denote it by $|T|$. Finally any WP decomposition can be represented by the set of pruned binary trees of $\mathbf{T}_{full}$ (the full tree), $\{T \subset E : T \ll \mathbf{T}_{full}\}$, corresponding to the set of rooted binary trees in G rooted at v_{root} (the root of $\mathbf{T}_{full}$).

Let $X(u)$ and $Y(u)$ be the random observation vector and the class random variable, respectively, defined on the underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$[1] and with values on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Applying the analysis-measurement to $X(u)$ for all WP decompositions, we induce a family of observation representations $\{X_T(u) \equiv m_T(X(u)) : T \ll \mathbf{T}_{full}\}$[2] taking values in $\left\{(\mathbb{R}^{|T|}, \mathcal{B}(\mathbb{R}^{|T|})) : T \ll \mathbf{T}_{full}\right\}$. This will be the family of lossy representations that will be considered as observation evidences in the classification problem. The next section summarizes some results for addressing the problem of optimal signal representation for classification for this particular family of tree-indexed feature representations.

3. MINIMUM PROBABILITY OF ERROR SIGNAL REPRESENTATION (MPE-SR)

Let us consider $\mathcal{D}_N = \{(x_i, y_i) : i = 1, .., N\}$ iid realizations of $(X(u), Y(u))$ and the family of feature representations by $\mathbb{D} = \{X_T(u) : T \ll \mathbf{T}_{full}\}$ with their respective empirical distributions estimated with $\mathcal{D}_N$. The minimum probability of error signal representation (MPE-SR) is given by [4]:

$$\mathbf{T}^* \equiv \arg \min_{\mathbf{T} \ll \mathbf{T}_{full}} \mathbb{E}\left(\mathbb{I}_{\{u \in \Omega : \hat{g}_T(X_T(u)) \neq Y(u)\}}(u)\right), \quad (2)$$

where $\hat{g}_T(\cdot)$ denotes the *empirical Bayes decision rule* — from $m_T(\mathcal{X})$ to $\mathcal{Y}$ — and the expected value is taken with respect to the true underlying distribution $\mathbb{P}$. Vasconcelos [1] has shown that the probability of error of an empirical Bayes decision rule, in Eq.(2), is affected by two sources of perturbations. One, the *Bayes error bound* which is associated with the intrinsic discrimination power of a given feature representation, and the other, the *estimation error*, which is the consequence of using an empirical class-observation probability measure for implementing the Bayes decision rule [1, 4].

Based on results originally presented in [1] and extended in [4], Eq.(2) can be addressed as a complexity regularized optimization problem with a fidelity indicator, *mutual information* (MI) $I(T) \equiv I(X_T(u), Y(u))$[3], reflecting the Bayes error bound and a penalization, dimensionality of the feature space, given by $|T|$, reflecting the estimation error [4],

$$\mathbf{T}^*(\lambda) = \arg \min_{T \ll \mathbf{T}_{full}} \Psi(I(T)) + \lambda \cdot \Phi(|T|). \quad (3)$$

Considering the tendency of our fidelity-cost indicators, $\Psi(\cdot)$ and $\Phi(\cdot)$ should be a strictly decreasing and increasing functions, respectively. The real dependency between the Bayes error bound and estimation error in terms of our new fidelity complexity values, $I(T)$ and $|T|$, is hidden and, furthermore, problem dependent. As a result, Ψ, Φ and λ in Eq.(3) provide degrees of freedom for approaching the solution of the MPE-SR problem, presented in Eq.(2). However, independent of those degrees of freedom, the MPE-SR solution $\mathbf{T}^*(\lambda)$ resides in a sequence of representations $\left\{\mathbf{T}^{k*} : k \in K(\mathbb{D})\right\}$, which are the solution of the following type of rate-distortion problem [4],

$$\mathbf{T}^{k*} \equiv \arg \max_{\substack{T \ll \mathbf{T}_{full} \\ |T| \leq k}} I\left(X_T; Y\right) \quad (4)$$

$\forall k \in K(\mathbb{D})$, with $K(\mathbb{D}) = \{|T| : T \ll \mathbf{T}_{full}\}$. Finally, the *empirical risk minimization* criterion using cross validation can be adopted as the final criterion for solving Eq.(2) [4].

3.1. Minimum Cost Tree Pruning Problem

Eq.(4) addresses the problem of finding the sub-band decomposition of the observation space that maximizes the MI, constraining it to a specific number of frequency bands. If we do not have some additive property on the tree functional involved, in particular $I(T)$, an exhaustive search needs to be conducted for solving Eq.(4),which grows exponentially with the size of the problem. For addressing this issue, let us characterize the analysis-measurement process as a function of a

[1] $\mathcal{F}$ and $\mathbb{P}$ denote the sigma field and probability measure, respectively.

[2] $m_T(X(u))$ represents the analysis-measurement process for T.

[3] Because of Fano's Inequality [5].

[4] Equivalent to finding the optimal $\lambda \in \mathbb{R}^+$ in Eq.(3) and consequently the optimal $k \in K(\mathbb{D})$ associated with $\mathbf{T}^{k*}$ for solving the MPE-SR problem.

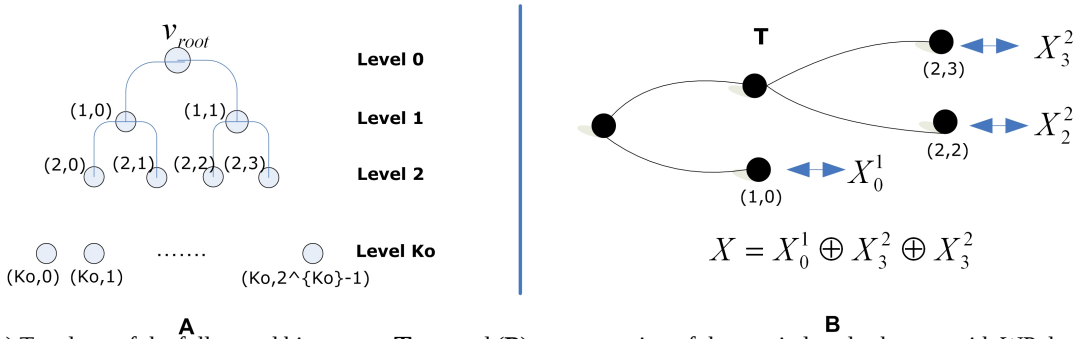

Fig. 1. **(A)** Topology of the full rooted binary tree $\mathbf{T}_{full}$ and **(B)** representation of the tree indexed subspace with WP decomposition.

family of rvs. indexed by the full tree, $\mathbf{T}_{full}$. More precisely, $m_T(X)(u) = (X_j^l(u))_{(l,j)\in\mathcal{L}(T)}$, where $X_j^l(u)$ represents the lossy measurement associated with the subspace $\mathcal{X}_j^l$ induced by the WP family, see Fig. 1.B. We have proved that $I(T)$ is an affine tree functional [6]. In particular, considering $\{v_{root}\} \ll T \ll \mathbf{T}_{full}$, a non-trivial tree, and $T_{(l,j)}$ as the branch of T rooted at $(l,j) \in \mathcal{I}(T)$, we can define $\rho_T(l,j) \equiv I(m_{T_{(l,j)}}(X); Y|X_j^l)$, where the following pseudo-additive property holds [6]:

$$\rho_T(l,j) = \Delta\rho(l,j) + \rho_T(l+1,2j) + \rho_T(l+1,2j+1), \quad (5)$$

where $\Delta\rho(l,j) \equiv I\left(X_{2j}^{l+1}, X_{2j+1}^{l+1}; Y|X_j^l\right)$ denotes the MI gain of splitting the band (l,j) of a non-terminal node of T. In Eq.(5), $(l+1,2j)$ and $(l+1,2j+1)$ represent the left and right children of (l,j). Noting that $\rho_T(l,j) = I(T_{(l,j)}) - I(X_j^l;Y)$ $\forall (l,j) \in \mathcal{I}(T)$, we can generalize Eq.(4) by:

$$\mathbf{T}_v^{k*} \equiv \arg \max_{\substack{T\ll\mathbf{T}_{full} \\ v\in T, |T_v|=k}} I(T_v) = \arg \max_{\substack{T\ll\mathbf{T}_{full} \\ v\in T, |T_v|=k}} \rho_T(v), \quad (6)$$

$\forall v \in \mathcal{I}(\mathbf{T}_{full})$, $\forall k \in \left\{1,..,\left|\mathbf{T}_{full_v}\right|\right\}$. In Eq.(6) using Eq.(5), we have a way for characterizing $I(T_v)$ as an additive combination of a root dependent terms, $\rho_T(\cdot)$, evaluated in its left and right branches. In fact, the following result, extended from the *minimum cost tree pruning problem* [7] for additive tree functionals, can be stated as follows:

THEOREM 1 *(Proof in [6]) Let $v \in \mathcal{I}(\mathbf{T}_{full})$ and let us denote its left and right children by $l(v)$ and $r(v)$ respectively*[5]*. Assuming that we know the solution Eq.(6) for the child nodes $l(v)$ and $r(v)$*[6]*, then the solution of Eq.(6) for v is given by:*

$$\mathbf{T}_v^{k*} = \left[v, \mathbf{T}_{l(v)}^{\hat{k}_1*}, \mathbf{T}_{r(v)}^{\hat{k}_2*}\right], \quad (7)$$

where [7]

$$(\hat{k}^1, \hat{k}^2) = \arg \max_{\substack{(k_1,k_2)\in\left\{1,..,\left|\mathbf{T}_{full_{l(v)}}\right|\right\}\times\left\{1,..,\left|\mathbf{T}_{full_{r(v)}}\right|\right\} \\ k_1+k_2=k}} \left[\rho_{\mathbf{T}_{l(v)}^{k_1*}}(l(v)) + \rho_{\mathbf{T}_{r(v)}^{k_2*}}(r(v))\right], \quad (8)$$

[5]From Fig. 1.A, $l(v) = (l+1, 2j)$ and $r(v) = (l+1, 2j+1)$.

[6]We know $\left\{\mathbf{T}_{l(v)}^{k_1*}, \mathbf{T}_{r(v)}^{k_2*} : k_1 = 1,.., \left|\mathbf{T}_{full_{l(v)}}\right| \; k_2 = 1,.., \left|\mathbf{T}_{full_{r(v)}}\right|\right\}$.

[7]$[v, T_1, T_2]$ represents a rooted binary tree T with root v, $T_{l(v)} = T_1$ and $T_{r(v)} = T_2$.

$\forall k \in \left\{1,..,\left|\mathbf{T}_{full_v}\right|\right\}$. *In particular, for v_{root} the solution for the original optimal pruning problem, Eq.(4), is given by*

$$\mathbf{T}^{k*} = \left[v_{root}, \mathbf{T}_{l(v_{root})}^{\hat{k}_1*}, \mathbf{T}_{r(v_{root})}^{\hat{k}_2*}\right]. \quad (9)$$

This result shows a way to solve our optimal tree pruning problem using a dynamic programing (DP) approach. This DP solution is a direct consequence of solving the optimization problem for the parent node as a function of the solutions of the same problem for its direct descendants [8].

4. EXPERIMENTS

For evaluating the performance of the optimal filter bank decomposition, we consider a simplified speech recognition scenario, where filter banks have widely been 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[9].

The corpus comprises about 1 hour 30 minutes of spontaneous conversational speech from a native American English speaker. sampled at 16 Khz. The spoken content was human segmented into utterances and transcribed at the word level. Word level transcriptions were used for generating phone level time segmentations on the acoustic signals by using automatic force Viterbi alignment techniques. The standard frame by frame analysis was performed on those acoustic signals, where every 10ms (frame rate) a segment of the acoustic signal of 64ms around a time center position was extracted. Finally 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 for purposes of this evaluation we considered one segment of the data comprising $N = 14979$ supervised sample points, $\mathcal{D}_N$.

We used this supervised data for addressing the problem of optimal filter bank decomposition using Daubechies' maximally flat filter (db4) for the WP basis family [2], and the energy on the resulting bands as the measurements framework, again motivated by standard feature representations used in automatic speech recognition. We first present some analysis of the minimum cost tree pruning in terms of topology of

[8]The pseudo-code for implementing this solution is presented in [7, 6].

[9]e.g. mel frequency filter bank.

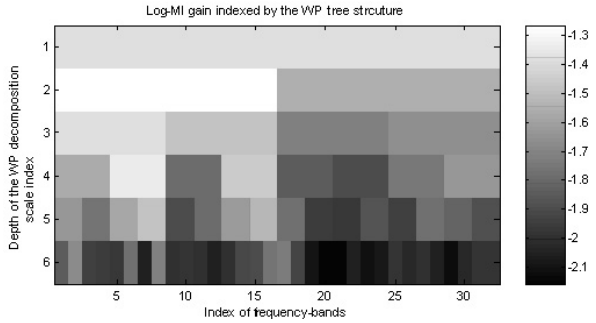

Fig. 2. Graphical representation of the MI gain by splitting the basic two channel filter bank across scale (level of decomposition, vertical axes) and bands (horizontal axes) in the WP decomposition.

those solutions (the optimal filter bank decomposition problem) and then we evaluate performances associated with those solutions.

4.1. Analysis of the MI Gain and Optimal Tree Pruning

Fig. 2 represents the MI gain across scale and frequency as a consequence of iterating the basic two channel filter bank that induces the WP basis family, $\rho_T(l, j)$ in Eq.(5) (estimated by a non-parametric density estimation approach [6]). The global trend 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 there is higher discrimination for lower frequencies than higher frequencies in the auditory range of 55Hz-15Khz. 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 signal (not reported here for space considerations). In this respect, exact Wavelet type of filter bank solutions (the type of filter bank structure induced from human perceptual studies, MEL scale) were obtained for solutions associated with small dimensions but those start deviating from this scenario in the process of exploring higher dimensional solutions.

4.2. Frame Level Phone Recognition

The solutions of the rate-distortion problem on the raw acoustic data were used as feature representations for purposes of frame level phone recognition. In particular, we evaluated solutions associated with the following dimensions: $4, 7, 10, 13$, $19, 25, 31, 37, 43, 49, 55$ and 61. A ten-fold cross validation was used, using a non-parametric classification approach, K-nearest neighbors (KNNs). As a reference, we consider the standard 39 Mel-Cepstrum (Mfccs) using same frame rate and window length, where the correct phone classification rate obtained was 68.71%. The performances of the rate-distortion solutions, mean (standard deviation), obtained across dimensions, **d:**, where:
d:4 $29.07\%(1.13)$; **d:7** $45.70\%(1.07)$; **d:10** $53.89\%(0.92)$; **d:13** $59.35\%(1.17)$; **d:19** $63.57\%(1.30)$; **d:25** $65.85\%(1.23)$; **d:31** $67.41\%(1.24)$; **d:37** $70.02\%(1.15)$; **d:43** $71.75\%(1.02)$; **d:49** $72.80\%(0.85)$; **d:55** $73.70\%(0.91)$; **d:61** $74.24\%(1.01)$.

Remarkably these results show that the rate-distortion solutions provide a level of improvement with respect to Mfccs, in particular considering the same number of dimensions. This interesting scenario shows that the proposed pruning algorithms present consistent solutions and competitive performance values with well understood empirically motivated feature extraction techniques. While the proposed information theory driven feature extraction offers promising phone recognition results, a more systematic evaluation of speech recognition experiments still remain to be done.

Furthermore, exploring solutions of the rate-distortion problem provides the flexibility for finding the optimal operational fidelity-complexity tradeoff for given classification scenario, such as in terms of number of training points available for the problem and the type of learning framework considered.

5. CONCLUSIONS AND FUTURE WORK

The solution for the minimum probability of error signal representation (MPE-SR) for Wavelet packets (WPs) feature representations is formally presented in this paper using a rate-distortion optimality criterion. MPE-SR solutions show competitive classification performances with respect to standard feature extraction techniques for the frame level phone recognition task. Furthermore, solutions for the rate-distortion problem reflects in considerable extent the expected frequency decomposition in this scenario. Future efforts will be devoted to extending the MPE-SR approach for the two dimensional WP families, widely used in various image classification problems, and in exploring conceptual connection with related tree-indexed optimization problem — decision trees and tree-structured vector quantizations (TSVQ).

6. REFERENCES

[1] N. Vasconcelos, "Minimum probability of error image retrieval," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2322–2336, August 2004.

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

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

[4] J. Silva and S. Narayanan, "Minimum probability of error signal representation based on a rate-distortion optimality criterion," *submitted for review*, 2006.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Interscience, New York, 1991.

[6] J. Silva and S. Narayanan, "Optimal wavelet packets decomposition based on the minimum probability of error signal representation principle," *submitted for review*, 2006.

[7] C. Scott, "Tree pruning with subadditive penalties," *IEEE Transactions on Signal Processing*, vol. 53, no. 12, pp. 4518–4525, 2005.