Discriminant Wavelet Basis Construction for Speech Recognition

C.J.Long and S.Datta
Department of Electrical and Electronic Engineering,
Loughborough University,
Loughborough,
Leics LE11 3TU
UK
Email:C.Long@iop.bpmf.ac.uk or S.Datta@lboro.ac.uk

10.21437/ICSLP.1998-535

ABSTRACT

In this paper, a new feature extraction methodology based on Wavelet Transforms is examined, which unlike some conventional parameterisation techniques, is flexible enough to cope with the broadly differing characteristics of typical speech signals. A training phase is involved during which the final classifier is invoked to associate a cost function (a proxy for misclassification) with a given resolution. The sub spaces are then searched and pruned to provide a Wavelet Basis best suited to the classification problem. Comparative results are given illustrating some improvement over the Short-Time Fourier Transform using two differing subclasses of speech.

1.1 INTRODUCTION

Multi-scale feature extraction is an attractive option when representing non-stationary real world signals such as speech. Coupled with integrated optimisation of the feature extraction and classification stages the aim is to provide an overall improvement in recognition performance. The problem is relevant because as modelling techniques have become vastly improved in recent years, further gains in recognition accuracy are likely to come from the preprocessing stage.

Wavelets and related techniques like subband coding have been applied with considerable success to speech processing applications such as compression [3],[4], and to a more limited extent on feature extraction for speech recognition / classification [5],[6].
Their main advantages are a somewhat richer multiresolution representation of the acoustic signal and the flexibility to use one of a number of basis functions. Subsequent refinements that aim to efficiently model signal statistics by choosing the depth of projection and amount of signal reduction adaptively [1] serve to improve accuracy of the model further.

Learning from the training set the best set of subspaces in which to model the data, results in a discriminant basis set which will highlight using the expansion coefficients of the wavelet transform (preferably just a few) the major differences between classes. If feature reduction is subsequently carried out, then the final classifier is designed in lower dimensional space. Assuming the data is well modelled in the first place, then there is a better chance of the classes being well separated by the classifier.

In this paper, we propose an implementation of this theoretical framework for tackling phoneme classification problems. The method is outlined in the next section.

2.1 Method

Let us first define the Discrete or Dyadic Wavelet Transform. The wavelet transform can be developed from a number of existing theories, here we will consider the extension of the DWT from its continuous counterpart; the CWT since this is intuitively similar to the Short Time Fourier Transform. The basic *analysing* or *mother* wavelet is given by:

$a^{-1/2}h(t-\tau/a)$ where τ and a are time shift and scale respectively. This shifted scaled set of functions forms an orthonormal family if sampled appropriately see [7] for further details of this. The *h(t)* furthermore, satisfy a number of constraints to enable them to be wavelets. For example, most well designed wavelets have *compact support* both in time *and* frequency enabling good feature localisation in the respective domains. Wavelet *regularity*, *vanishing* moments and *orthogonality* are design parameters which influence factors such as reconstruction fidelity, degree of compression achievable, or type of signal most suitable for decomposition in that wavelet basis. A wealth of literature exists on this subject see [7], [8], [9] for details.

If we take $a = a_0^m$ and $\tau = nb_0a_0^m$, where *n* and *m* are the discretisation integers on the dyadic grid, the resulting wavelets then become

$$h_{m,n}(t) = a_0^{-m/2}h(a_0^{-m}t - nb_0) \qquad (1)$$

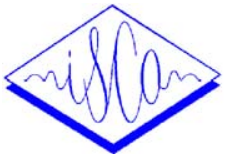
with the added constraint that $\int h(t)dt = 0$.

The discrete wavelet transform is just a projection of a given signal onto these analysing functions:

$$c_{m,n}(f) = \langle h_{m,n}, x \rangle \qquad (2)$$

The algorithm used in the following experiments is connected to the Local Discriminant Bases [1] developed for classification as a direct extension of the original Best-Basis algorithm [2]. The LDB uses dictionaries of Wavelet Packets and Local Cosine Transforms, which will be defined shortly, to form a library from which the best basis dictionary may be chosen using, as criterion, one of a number of cost functions. These cost functions, of which there are a number of differing types, are generally additive, but all essentially provide a measure of 'energy concentration' of the vector.

ISCA Archive
http://www.isca-speech.org/archive



5th International Conference on Spoken Language Processing (ICSLP 98)
Sydney, Australia
November 30 - December 4, 1998

Definition: An additive cost function ϑ^{add} from a sequence $\{x_i\}$ to $\Re$ is additive if $\vartheta(0)=0$ and $\vartheta(\{x_i\})=\sum_i (x_i)$.

In LDB, the cost function used is *relative entropy* which should be a good measure of the power of discrimination of each subspace. If we consider a simple two class case, where $\boldsymbol{p}=\{p_i\}_{i=1}^n$, $\boldsymbol{q}=\{q_i\}_{i=1}^n$ are two normalised energy distributions of signals belonging to class 1 and class 2 respectively. The Relative Entropy is then given as :

$$RE(\boldsymbol{p},\boldsymbol{q}) \equiv \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \qquad (3)$$

Typically one first computes an estimate of the class probabilities by calculating a time-frequency energy map for each signal class from the Wavelet packet / Cosine packet transforms.

Definition: The Wavelet/Cosine Packet Transform is a generalisation of the standard discrete wavelet transform given in (2).If the signal subspace is given as $\Omega_{j,k}$ ie the coarsest resolution, then each node is split recursively in a manner similar to the DWT to form a binary tree of subnodes. If j is the depth and k the subspace number (either 0 or 1), the first level will have two subspaces, $\Omega_{j+1,k}$ and $\Omega_{j+1,k+1}$.The next will have four and the J^{th} 2^j. In total there will be $2^{2^{J-1}}$ possible subnodes in the tree and the issue is to extract a non redundant signal representation by assigning criterion such as Relative entropy to each node and pruning/growing a tree to maximise this measure.

The DWT on the other hand is iterated only on the Low Pass part of its decomposition and is such that a non-redundant representation is guaranteed. Wavelet Packets, on the other hand, have the advantage of covering signal space entirely and provide an unfixed resolution tiling of the time-frequency plane although the Heisenberg inequality principle still of course holds. However they are overcomplete and require some kind of pruning if orthogonality is to be achieved. This search will be fast if the cost function is additive.
Local Trigonometric Transforms or Sine/Cosine Packet Transforms are exactly analogous to the WP transform except that they partition the time instead of the frequency axis smoothly.

Here is the LDB algorithm used in the experiment. Assume $\Phi_{j,k}$ is the discriminant measure, whether additive or not, let $D_{j,k}$ represent the Best Discriminant Basis and $R_{j,k}$ the fully expanded, redundant basis:

0) Choose to use either trigonometric dictionaries or Wavelet Packets for the transform.

1) Expand every signal in the training set into its wavelet packet table.

2) Determine the set of most discriminant subspaces using a top down pruning methodology by testing the efficacy of each subspace for discrimination.
i.e. set a temporary array $\Im_{j,k} = \Phi_{j,k}$

if $\Im_{j,k} \geq \Im_{j+1,2k} \cup \Im_{j+1,2k+1}$; $D_{j,k}=R_{j,k}$;
else
$D_{j,k}=D_{j+1,2k} \oplus D_{j+1,2k+1}$ and set
$\Im_{j,k} = \Im_{j+1,2k} \cup \Im_{j+1,2k+1}$

3) Rank the expansion coefficients according to their discriminant power and from these select the top $k \leq n$ features (where $n = 2^{no}$ is the dyadic length of the signal) for each signal in the training class to construct the final classifier.

The LDB gained from step two is an orthonormal basis, also if the cost function is additive, this step will be fast.
Step 3 isn't necessary since we can still design the classifier on all the features, however if the dimensionality of the problem is reduced, this step will reduce the number of interfering components in the decomposition, making the class-specific features more robust. Computational training times will simultaneously be reduced. In practice one can rank the expansion coefficients by a) Finding the discriminant validity of a particular basis function in the LDB expansion. b) Use Fishers class separability index to rank the coefficients.

Results

In the following experiments, the above algorithm was implemented using the standard LDB configuration: an additive cost function of Relative Entropy and the best-$k \leq n$ chosen using the same criterion.

This approach was compared with a configuration using non-additive costs; a proxy for LDA-derived misclassification rate was used and the expansion coefficients ranked using Fishers class separability criterion. In addition in this case, we applied a small non-linear thresholding to the subspace vectors prior to calculating the misclassification rate. The final classifier in both cases was LDA thus in case 2 the same optimality criterion was used both in the evaluation of suitable features for class separability as for the final classification estimate. The wavelet used in all cases was the Daubechies 6^{th} order wavelet.

The phoneme classification problems broached dealt two extreme cases: first, three well behaved (in the statistical sense), well separated vowels aa,ax,iy corresponding to the back, mid and front positions of the tongue during voicing were examined. Secondly, the three unvoiced stops, p,t,k were discriminated against one another. In both cases, the

phonemes were extracted from dialect region 1 of the Timit database from all speakers both male and female to ensure a good statistical representation of each sound. The speech datasets used were sampled at a rate of 16Khz, thus the 32ms window, which we assumed, was composed of ~512 samples.
The results gained using the methods outlined plus a benchmark version of the STFT, commonly used in speech parameterisation are given in Table 1.

Technique	Error Rate (Training)	Error Rate (Testing)	Problem
LDA on STFT64	9.39%	10.35%	iax
LDA on LDB60	8.53%	**9.40**%	iax
LDA on LDBuLDA60	9.2%	10.1%	iax
LDA on STFT64	33.51%	43.87%	ptk
LDA on LDB60	31.41%	**39.68%**	ptk
LDA on LDBuLDA60	30.68%	42.58%	ptk

Table 1: Misclassification rates of the feature extraction techniques when applied to two phoneme classification problems. LDA,STFT64,LDB60 indicate the type of final classifier used, 64 short-time fourier transform gained from whole 512 via decimation, the top 60 expansion coefficients extracted using standard LDB.LDBuLDA60 is the top 60 coordinates obtained using LDA-derived optimality criterion.

CONCLUSIONS

With regard to the number of features chosen, approximately 10% of the original signal dimensionality was used. The performance of the wavelet methods was noticeably better than the STFT. The initial computational cost of the Wavelet Packet related methods is always going to be greater since there is a significant cost in the pruning part of the algorithm not present in FT methods - especially if LDA is used at this stage. However this is only a training cost, once a basis tree is worked out, all subsequent signal known to belong to a broad phonetic subclass can be decomposed in a comparably fast manner. It should also be emphasised, in particular for the ptk experiment that this is a difficult classification problem, we ourselves would generally use context and higher level knowledge to characterise these. The type of system proposed has been shown to provide some improvement over a standard widely used parameterisation technique in two situations, it is likely to be of robustly similar performance in other recognition scenarios. As a preprocessing technique to standard modelling conventions e.g. HMM it certainly shows some promise. It is likely anyway that a better recogniser would highlight improvements between Wavelet over Fourier decompositions, it has been noted in [10] that LDB derived features appeared "oblique" in a sense and this is borne out in some of our other experiments where the true multiresolutional advantages of wavelet appeared much superior. Better performance could also be had by using some standard preprocessing of which none was done here since for the purposes of comparison this was irrelevant.
With regard to the decrease in performance between standard LDB and LDB using an LDA-derived non-additive cost, we felt was perhaps due to non-linear relations within the training set not being exploited. Instead of using LDA, in future we will try a neural network to provide a cost and incorporate this seamlessly into the whole design.

REFERENCES

[1] N.Saito, "Local Feature Extraction and its Applications using a Library of Bases," *A Dissertation*, Yale University, Dec. 1994.
[2] R.R.Coifman and M.V.Wickerhauser, "Entropy-Based Algorithms for Best Basis Selection," *IEEE Transactions on Information Theory*, vol. 38, no.2, pp. 713-718, March 1992.
[3]M.V.Wickerhauser, "Acoustic Signal Compression with Wavelet Packets," In Chui C.K.(ed) Wavelets: A Tutorial in Theory and applications (1992).
[4]J.A.Thripuraneni et al, "Mixed Malvar Wavelets for Non-stationary Signal Representation," *Proc.ICASSP vol.1pp.13-16, Atlanta* (1996).
[5]C.D'Alessandro and G.Richard, "Random wavelet representation of unvoiced speech," *Proc. IEEE-SP Int.Symp. on Time Frequency and Time Scale Analysis*, pp. 41-44 (Oct.1992).
[6]S.Kadambe and G.Faye Boudreaux-Bartels, "Application of the Wavelet Transform for Pitch Detection of Speech Signals," ," *IEEE Transactions on Information Theory*, vol. 38, no.2, pp. 713-718, March 1992.
[7]I.Daubechies, "Orthonormal bases of compactly supported wavelets," Comm in Pure and Applied Math., vol.41 No.7, pp.909-996, 1988.
[8]S.Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," IEEE Trans. Pattern Anal. Machine Intell., vol. 31, pp. 674-693, 1989.
[9]O.Rioul and M.Vetterli, "Wavelets and Signal Processing," IEEE Signal Processing Magazine, October 1991.
[10]N.Saito, "Classification of Geophysical Acoustic Waveforms using Time-Frequency Atoms," *Proceedings of Statistical Computing, Amer. Statist. Assoc.* (1996).