Wavelet Based Feature Extraction for Phoneme Recognition

C.J.Long and S.Datta
Department of Electronic and Electrical Engineering
Loughborough University of Technology
Loughborough LE11 3TU, UK.

ABSTRACT

In an effort to provide a more efficient representation of the acoustical speech signal in the pre-classification stage of a speech recognition system, we consider the application of the Best-Basis Algorithm of Coifman and Wickerhauser. This combines the advantages of using a smooth, compactly-supported wavelet basis with an adaptive time-scale analysis dependent on the problem at hand.

We start by briefly reviewing areas within speech recognition where the Wavelet Transform has been applied with some success. Examples include pitch detection, formant tracking, phoneme classification. Finally, our wavelet based feature extraction system is described and its performance on a simple phonetic classification problem given.

1. INTRODUCTION

Speech recognition systems generally carry out some kind of classification/recognition based upon speech features which are usually obtained via time-frequency representations such as Short Time Fourier Transforms (STFTs) or Linear Predictive Coding (LPC) techniques. In some respects, these methods may not be suitable for representing speech; they assume signal stationarity within a given time frame and may therefore lack the ability to analyse localised events accurately. Furthermore, the LPC approach assumes a particular linear (all-pole) model of speech production which strictly speaking is not the case.

Other approaches based on Cohens general class of time-frequency distributions such as the Cone-Kernel and Choi-Williams methods have also found use in speech recognition applications but have the drawback of introducing unwanted cross-terms into the representation.

The Wavelet Transform overcomes some of these limitations; it can provide a constant-Q analysis of a given signal by projection onto a set of basis functions that are scale variant with frequency. Each wavelet is a shifted scaled version of an original or mother wavelet. These families are usually orthogonal to one another, important since this yields computational efficiency and ease of numerical implementation. Other factors influencing the choice of Wavelet Transforms over conventional methods include their ability to capture localised features. Also, developments aimed at generalisation such as the Best-Basis Paradigm of Coifman and Wickerhauser [1] make for more flexible and useful representations.

We consider the possibility of providing a unified wavelet-based feature extraction tool, one designed to contend optimally with the acoustical characterisics particular to speech, in the most computationally efficient manner.

The indications are that the Wavelet Transform and its variants are useful in speech recognition due to their good feature localisation but furthermore because more accurate (non-linear) speech production models can be assumed [2]. The adaptive nature of some existing techniques results in a reduction of error due to inter/intra speaker variation.

We shall begin by defining the wavelet transform.

2. WAVELETS AND SPEECH

2.1 The Discrete Wavelet Transform

The basic wavelet function $\psi(t)$ can be written

$$\psi_{\tau,a} = \frac{1}{\sqrt{a}}\psi\left(\frac{t-\tau}{a}\right) \qquad (1)$$

The *Continuous* Wavelet Transform is then defined as

$$X(\tau,a) = \frac{1}{\sqrt{a}}\int x(t)\psi^{*}\left(\frac{t-\tau}{a}\right)dt \qquad (2)$$

where $\psi(t)$ is known as the *analysing* wavelet or *prototype* function. Typically, these continuous wavelet functions are overcomplete and therefore do not form a true orthonormal basis. Redundancy may be eliminated by appropriately sampling the wavelet on a dyadic lattice, i.e. in a manner that reflects the tiling of the time-frequency plane as in figure 1. An orthonormal basis of compactly supported wavelets can then be obtained to span $L^2(\Re)$ (the space of all finite energy signals) by shifting and dilating the wavelet function $\psi(t)$ i.e.

$$\psi_{n,m}(k) = 2^{-n/2}\psi(2^{-n/2}k - mb_o) \qquad (3)$$

where n=1,2, . . . represents the scale and m=0,1, . . . the time shift. Note that the scaling factor a is here chosen as 2 in order that the frequency axis is decomposed in octaves. Now if one chooses a suitable wavelet, a true orthonormal basis will be obtained. This results in a multiresolutional analysis of a given signal over $L^2(\Re)$, yielding a time -scale decomposition similar to that exhibited in Figure 1. For further details on MRA, the reader is referred to the work of Mallat [3].

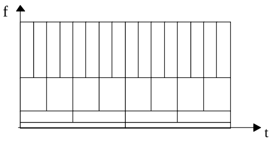


Figure 1: Tiling of time-frequency plane via the wavelet Transform.

2.2 Pitch and Formant Extraction using Wavelet Analysis

Kadambe & Boudreaux-Bartels [4] have used the multiresolutional properties of wavelets to propose an event - based pitch-detection system. Their method works by detecting the Glottal Closure Instant (GCI) and determines the pitch for each sample within a particular speech segment. This approach is particularly suitable for noisy speech.

Evangelista [5] has developed a ‘Pitch-Synchronous’ wavelet representation using a modified version of the QMF (Quadrature Mirror Filter) multiplexed filter bank outlined in [6]. Using the MRA properties of wavelets, the pitch-synchronous wavelet transform (PSWT) can be used for pitch tracking once the pitch has been extracted using conventional methods. Unique characterisation of speech events such as fricatives and occlusive unvoiced consonants may thus be achieved via the variation in the pitch of the signal.

Maes [7] reports success in the extraction of pitch and formants from speech. The speech signal is first decomposed into its subbands using the wavelet transform and the temporal behaviour of the speech in each subband is monitored using a ‘squeezing’ algorithm. Those components exhibiting similar temporal behaviour are then recombined and the resulting principle components represent the pitch and formant characteristics of the speech signal.

In [11], Wesfried introduces a speech representation based on the Adapted Local Trigonometric Transform. The window size into which the data is partitioned is dependent upon the spectrum it contains, and the transitions between windows is seen to be suitable for segmentation into voiced-unvoiced portions. A formant representaion is also introduced by carrying out the following compression: locating and retaining the centres of mass for the highest-value peaks of the transform. From this, the local spectrum is said to represent the formant of the speech signal.

2.3 Phoneme and Speaker Classification using Adaptive Wavelets

The adaptive wavelet transform and the concept of the *super wavelet* were developed as an alternative to existing wavelet representation schemes [8]. Given a wavelet function of the form shown in (2), the idea is to iteratively find the translation and dilation parameters, τ and a respectively such that some application-dependent energy function is minimised. With respect to the classification problem, a set of wavelet coefficients would normally be estimated to represent certain features of a given signal. Classification can then be performed by using the feature set as the input to a neural net classifier. The adaptive wavelet based classifier is given as

$$v(n) = \sigma(u_n) = \sigma\left[\sum_{k=1}^{K} w_k \sum_{t=1}^{T} x_n(t)\psi\left(\frac{t-\tau_k}{a_k}\right)\right] \qquad (4)$$

where *v(n)* is the output for the n^{th} training vector $x_n(t)$ and $\sigma(z) = 1/[1+\exp(-z)]$. For two classes, $w_{k,}a_{k,}\tau_k$ can be optimised by minimising the energy function in the least squares sense (see eq 5). In [2] then, two classification examples are considered with application to speech; classification of unvoiced phonemes and speaker identification.

$$E = \frac{1}{2}\sum_{n=1}^{N}\left(d_n - v_n\right)^2 \qquad (5)$$

The system first models the phonemes using a mother wavelet similar to Figure 2 (used because of its noise-like characteristics)only of order 3 and then presents the wavelet features to a 2 layer feed-forward neural network. Speaker i.d. is similarly achieved only using a Morlet wavelet to model the phonemes since these are voiced and hence semi-periodic and smooth. The classifier then attempts to identify a speaker by clustering the associated utterances into one class. Results reported are very high accuracy, although exhaustive testing on a larger database will be needed to evaluate the method more accurately.

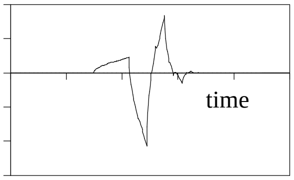


Figure 2: Daubechies Wavelet of order 4. This type of wavelet is used in Kadambe’s unvoiced sounds classifier because of its suitability for modelling high frequency noise-like signals.

3. THE BEST-BASIS ALGORITHM

A generalisation of the Wavelet Transform originally designed for signal compression is the Best-basis algorithm first described in [1]. The idea is to do transform coding on a signal by choosing a wavelet basis which is best suited for the given problem, resulting in an adaptive time-scale analysis. In particular, two possibilities are proposed, the smooth local trigonometric transforms which essentially performs local Fourier analysis on the signal, and its frequency domain conjugate, the wavelet packet which similarly partitions the frequency axis smoothly. Since these transforms operate on recursively partitioned intervals on the respective axis, the bases whether wavelet packet or local trigonometric are said to form a *library of orthonormal bases*. If these bases are ordered by refinement, they form a tree which can be efficiently searched to result in only those coefficients which contain the most information.

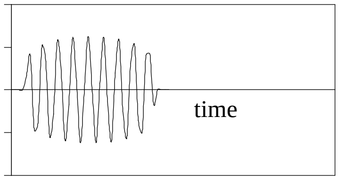


Figure 3: An example of a modulated smooth trigonometric packet. A localised sine dictionary, for example, would consist of a number of scaled, oscillatory versions of these.

In summary, the aim is to extract the maximum information or *features* from our signal by projection onto a co-ordinate system or basis function in which that signal is best (most efficiently) represented. What is meant by efficiency really depends on the final object. If compression is required, then the most efficient basis will be the one wherein most of the information is contained in just a few coefficients. On the other hand if we are interested in classification, a basis which most uniquely represents a given class of signal in the presence of other known classes will be most desirable.

Figure 4 shows the structure of the wavelet based acoustic-phonetic feature extractor used in the pre-classification stage. Our library of basis contained just two dictionaries, wavelet packets and smooth localised cosine packets, although others are certainly possible. Thus the first stage of the system is to choose the most suitable of these for the problem at hand. This is done in practice by simply picking the one which gives minimum entropy among them [10].

3.1 Experimental

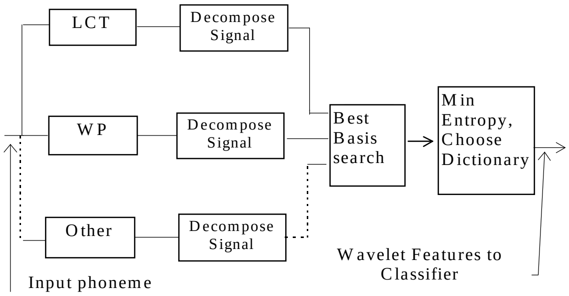


Figure 4: Wavelet-based feature extractor.

After Kadambe et al [2] who implement phonetic modelling and classification using adaptive wavelets, we use for training the voiced phoneme /d/ as in *had*, and the two unvoiced phonemes /s/ as in *ask* and /z/ as in *was*. These phonemes were extracted from a single male speaker in the TIMIT database. Each signal was low-pass filtered to resemble 8Khz band-limited telephone speech. The training features for the two layer feed-forward neural network were then obtained via the best-basis paradigm. A dictionary was chosen from our library of possible bases for each phoneme, dependent on which provided the minimum of a specified cost function, in this case entropy. As it turned out, the LCT (Local Cosine Transform) dictionary was selected for the voiced phoneme /d/ since these set of functions are smooth and most suitable for representing oscillatory signals. The /s/ and /z/ phonemes which correspond to different types of noise were best represented in terms of the wavelet packet with basis functions similar to Figure 2, i.e. a Daubechies wavelet of order 4. A fast search, (i.e. $O(n[\log n]^p)$ where p=0,1,2 depending on the basis type) was then performed in a binary tree similar to that of Figure 5.

The wavelet features of the training vectors obtained using this method are shown in Figure 6(a), (b), and (c) along with the original signals decimated to a length of 1024 samples. A restriction, in fact, of this method is that it requires a dyadic length which is a power of 2. To reduce the dimensionality of the training vectors, each signal was segmented into 4 equal parts. Similarly to Kadambe et al [2], we added Gaussian noise with $\sigma = 0.1$ independently to the first segment of each phoneme to give an extra ten training vectors for each class. Thus we obtained a total of 42 training vectors all normalised to unit norm. The neural network classifier had 5 nodes in its hidden layer after empirically determining that this number gave sufficient classification. When the classifier was tested on the training data it gave 100% accuracy with a 90% confidence threshold.

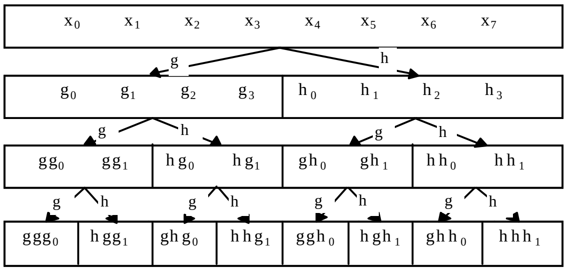

x_0 x_1 x_2 x_3 x_4 x_5 x_6 x_7
g
h
g_0 g_1 g_2 g_3 h_0 h_1 h_2 h_3
g h g h
gg_0 gg_1 hg_0 hg_1 gh_0 gh_1 hh_0 hh_1
g h g h g h g h
ggg_0 hgg_1 ghg_0 hhg_1 ggh_0 hgh_1 ghh_0 hhh_1

Figure 5: Best-Basis wavelet approximations organise themselves into a binary tree.

The next stage was to test the trained network on unseen data. We used the same kind of phonemes from the same speaker but uttered under different context, /d/ as in *dark,* /s/ as in *struggle,* and /z/ as in *Herb's* (see Figure 7). Overall classification was again 100% but with a lower confidence level, about 60%.

4. EVALUATION

The acoustic-phonetic feature extraction method described here takes advantage of the adaptive time-frequency localisation characteristics of the Best-Basis method to efficiently represent perceptually relevant acoustical events. That the features extracted are suitable for classification tasks has been illustrated by means of a simple training and test set consisting of those signal features contained in the wavelet coefficients. The results at this stage are promising and warrant the testing of this method on a larger database of speech data. It is interesting to note the structural similarities between the transformed data sets in Figures 6(f) and 7(a) of the contextually different phonemes /z/ used in the training and test phonemes respectively. The /s/ and /d/ phonemes exhibit a similar characterisitic.

5. REFERENCES

[1]Coifman, R.R. and Wickerhauser M.L. "Entropy based algorithms for best-basis selection," *IEEE Transactions on Information Theory*, vol.32, pp.712-718, March 1992.

[2] Kadambe, S; Srinivasan, P. "Applications of Adaptive Wavelets for Speech," *Optical Engineering* 33(7), pp.2204-2211 (July 1994).

[3] Mallat, S.A "Theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 31, pp.674-693, 1989.

[4] Kadambe, S. and Boudreaux-Bartels, G.F. “Application of the Wavelet Transform for Pitch Detection of Speech Signals, ” *IEEETransactions on Information Theory*, vol.32, pp.712-718, March 1992.

[5] Evangelista, G. “Pitch-synchronous wavelet representations of speech and music signals,” *IEEE Transactions on Signal Processing*, Vol. 41, No.12, December 1993.

[6] Evangelista, G. “Comb and multiplexed wavelet transforms and their application to speech processing,” *IEEE Transactions on Signal Processing*, Vol. 42, no.2, February 1994.

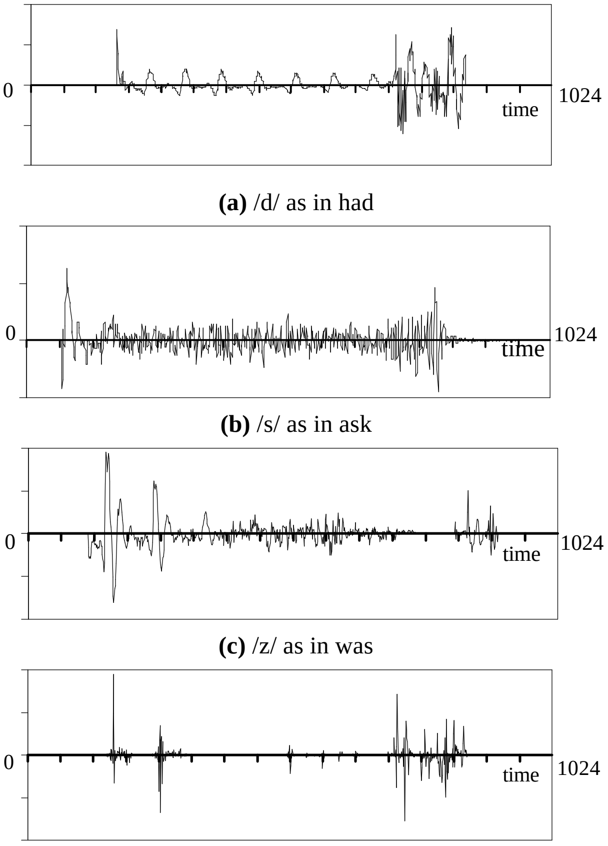
[7] Maes, S. “Nonlinear techniques for parameter extraction from quasi-continuous wavelet transform with application to speech,” *Proceedings of SPIE - The International Society for Optical Engineering* 1994. Vol.2093 pp. 8-19.

[8] Szu, H; Telfer, B; Kadambe, S. “Neural network adaptive wavelets for signal representation and classification,” *Optical Engineering*, vol.31 No.9 pp.1907 1916, September 1992.

[9] Saito, N. “Local feature extraction and its application using a library of bases.” *Phd thesis*, Yale University (1994).

[10] Buckheit, J.B. and Donoho, D.L. “Wavelab and reproducible research,” in *Wavelets and Statistics*, Springer-Verlag, New York (1995).

[11] Wesfried, E; Wickerhauser, M.V. “Adapted local trigonometric transforms and speech processing,” *IEEE SP* 41, 3597-3600 (1993)

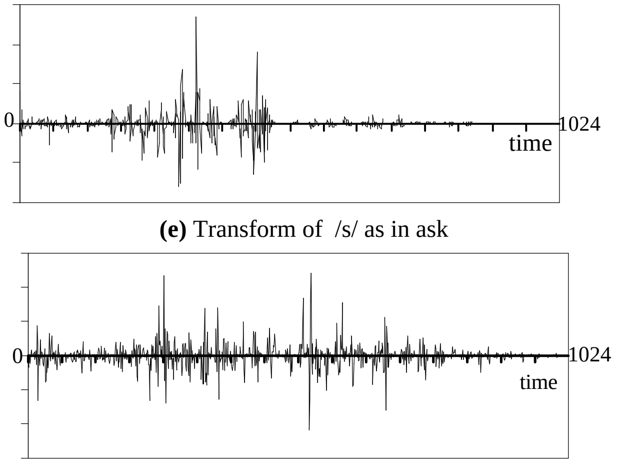


(a) /d/ as in had

(b) /s/ as in ask

(c) /z/ as in was

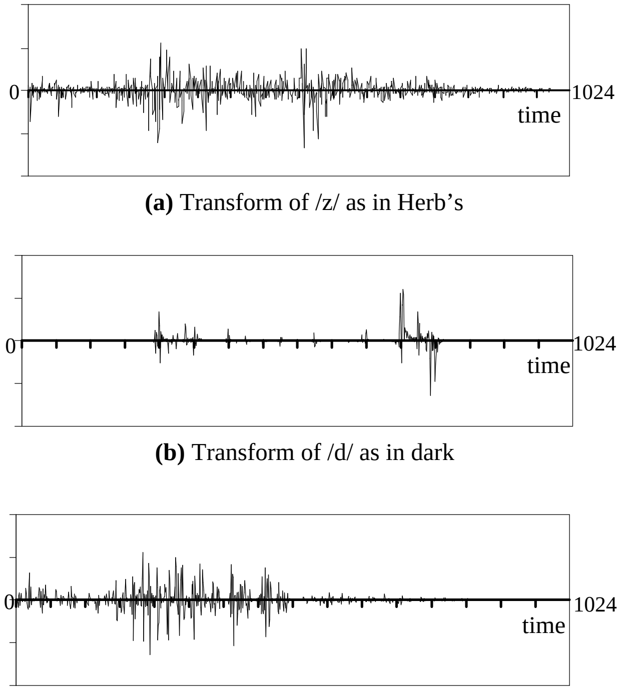
(d) Transform of /d/ as in had.



(e) Transform of /s/ as in ask

(f) Transform of /z/ as in was

Figure 6 (a)-(c): Original training signal. **(d)-(f)** Transform of respective signals.



(a) Transform of /z/ as in Herb’s

(b) Transform of /d/ as in dark

(c) Transform of /s/ as in struggle

Figure 7: Wavelet Transforms of test data**.** Note the correlation between transforms of contextually different phonemes.