A Cortically-Inspired Model for Bioacoustics Recognition

Linda Main and John Thornton

Cognitive Computing Unit, Institute for Integrated and Intelligent Systems
Griffith University, Gold Coast, Australia
l.main@griffith.edu.au, j.thornton@griffith.edu.au

Abstract. Wavelet transforms have shown superior performance in bioacoustic recognition tasks compared to the more commonly used Mel-Frequency Cepstral Coefficients, and offer the ability to more closely model the frequency response behaviour of the basilar membrane within the cochlea. In this paper we evaluate a gammatone wavelet as a preprocessor for the Hierarchical Temporal Memory (HTM) model of the neocortex, as part of the broader development of a biologically motivated approach to sound recognition. Specifically, we implement a gammatone/equivalent rectangular bandwidth wavelet transform and apply it, in conjunction with the HTM's spatial pooler, to recognise frog calls, bird songs and insect sounds. We demonstrate the improved performance of wavelets for feature detection and the potential viability of using HTM for bioacoustic recognition. Our classification accuracy of 99.5% in detecting insect sounds and 96.3% in detecting frog calls are significant improvements on results previously published for the same datasets.

Keywords: signal processing, wavelet transforms, bioacoustics, machine learning, Spatial Pooling, Hierarchical Temporal Memory, k-NN classifier

1 Introduction

In order to apply machine learning to auditory detection and classification, a suitable source signal preprocessing method is required. Such methods have the goal of revealing salient features in the data that will best facilitate the learning process. The choice of preprocessor is typically based on the nature of the signal and the desired properties of the extracted features, without considering the theoretical principles on which the learning algorithm is based.

Traditionally, Mel-Frequency Cepstral Coefficients (MFCCs) have been used to preprocess signals for audio recognition. MFCCs are obtained via a short time Fourier transform (STFT) to produce the power spectrum, which is thought to model human vocal tract characteristics [1]. The spectrum is then warped on the perceptual Mel-frequency scale in order to model the frequency response behaviour of the basilar membrane. The resulting features are therefore a combination of modelling both speech production and auditory response mechanisms.

An alternative approach is the extraction of audio features by means of a wavelet function [2]. The wavelet transform (WT) uses basis functions with limited duration which are isolated in time and frequency, where each wavelet has a characteristic location and scale. Wavelets offer a number of benefits: improved time-frequency resolution compared to MFCCs; the envelope of the wavelet is mathematically tractable [3]; there are a variety of established basis functions available; and they can be used to model aspects of mammalian auditory perception.

In this paper we evaluate WTs as a preprocessor for the Hierarchical Temporal Memory (HTM) model of the neocortex, and as part of the broader development of a biologically motivated approach to sound recognition. HTM is a high level implementation of mammalian neocortical structure that aims to minimise unexpected interactions with the environment by learning to predict future input. As such it fits within the framework of the free energy principle [4], but differs from Friston's Hierarchical Predictive Coding (HPC) model by implementing neocortical regions as networks of artificial mini-columns that learn sequences of input received from lower regions with the assistance of feedback from higher regions.

To the best of our knowledge, HTM has not previously been applied to bioacoustic recognition, nor has the idea of using a biologically-inspired model for processing bioacoustics. To address this we introduce a gammatone/equivalent rectangular bandwidth (ERB) wavelet transform as a model of biological audition and evaluate its performance on three bioacoustic datasets, using a k-Nearest Neighbour classifier. Our results show the gammatone/ERB WT outperforms the previously reported best classification accuracies on two of our datasets. However, we found that classification accuracies fall significantly when the WTs are further processed through an HTM spatial pooler, indicating that raw WTs are not the best form of input for an HTM system. On this basis we conclude that further work, modelling processes already occurring in the brainstem, will be needed before an HTM can perform competitively in this domain.

2 Related Work

The primary challenges in bioacoustic recognition are the handling of a diverse range of animal sounds [5] and the difficulty of identifying them against a background of ambient noise. While MFCCs have been popular, it has been shown that other feature detection methods may offer better performance.

In 2006, Mitrovic et al. described a set of time-based low-level features which they compared to MFCCs. They found that a combination of their new features provided significant improvement over MFCCs in bioacoustic classification [6]. Gonzalez compared MFCCs against a selection of spectral features on a range of sound classes (e.g. music, frog calls, rain, etc.), and demonstrated that variations to the Principle Component Analysis (PCA) approach were competitive [7]. In the LifeCLEF 2014 Identification Challenge for birdsong classification, Stowell and Plumbley's winning audio-only submission used spherical k-means to

learn features from PCA-whitened Mel spectral frames which significantly outperformed MFCC-based approaches [8]. Wavelets have also outperformed spectrogram template matching techniques for classifying Humpback whale song [9], and STFT and MFCC for classifying bat echolocation calls [10].

3 Wavelet Transforms

MFCCs, while efficient to compute and considered to perform well, are susceptible to noise and have poor time-frequency resolution. This means that while the transform is able to extract frequencies with a high degree of accuracy, the time at which they occur within the signal is lost. As sound has a fundamentally temporal nature, the loss of timing information is likely to impact on a system's ability to perform classification. The STFT attempts to address this shortcoming by applying the Fourier transform in sliding windows that move with time. The downside is that the length of the window limits the frequency resolution according to the well-known Heisenberg Uncertainty relationship [3].

In order to achieve greater time resolution while maintaining good frequency detail, the WT may be used. The translation of the wavelet basis function across the signal allows identification of the temporal location of the obtained frequencies, and by varying the scale of the wavelet, high frequency resolution is maintained. This multi-scale approach makes wavelets ideal for extracting features from the non-stationary signals typical of bioacoustics.

An attractive aspect of WTs is the ability to modify the envelope (or window) of the wavelet and thereby optimise the quality of extracted features with respect to the target application. Various implementations using wavelet bases have been developed, with the discrete wavelet transform (DWT) being one of the most widely used. This is due to its non-redundant and invertible nature, which are key requirements for techniques aimed at signal compression and decompression (e.g. JPEG 2000).

Morlet Wavelet: The Morlet (or Gabor) wavelet was one of the first basis functions developed. The sinusoid of the Morlet wavelet is modified by a smooth Gaussian window, producing a waveform that is symmetrical about the peak amplitude. In a Morlet WT the translation and scaling factors are typically calculated as a linear progression.

Gammatone Wavelet: The mechanical frequency analysis of the cochlea is often modelled using a gammatone filter, which is considered to give a reasonable first-order approximation of basilar membrane impulse responses [14]. A gammatone filter is the product of a gamma distribution function and a sinusoidal tone centred at frequency f_c, calculated as:

$$g(t, B, f_c) = K\, t^{(n-1)} e^{-2\pi B t} e^{j2\pi f_c t} \qquad t > 0 \tag{1}$$

where K is the amplitude factor; n is the filter order; f_c is the central frequency in Hertz; and B represents the duration of the impulse response [15].

Basilar membrane impulse responses are nearly linear for frequencies ranging between 20–1,000 Hz and approximately logarithmic between 1–20 kHz. Glasberg and Moore's Equivalent Rectangular Bandwidth (ERB) calculation may be used to model the basilar membrane's progressive bandwidth scaling [16]. By modifying the B term of the gammatone wavelet function according to the ERB scale, we obtain a filter bank considered to be a close match to the biology:

$$B = ERB(f) = 2.47 \times (4.37 \times f + 1) \tag{2}$$

Using the gammatone/ERB wavelet transform, which models only biological audition, as a preprocessor for HTM, which models the neocortex, we can construct a biologically plausible processing pipeline which is coherently focussed on auditory processing.

4 Hierarchical Temporal Memory

The theoretical principles of HTM have been developed as a set of Cortical Learning Algorithms [12], which implement sparse coding, distributed representation, Hebbian learning, and inhibition techniques. HTM is distinguished from other related models (such as HPC) by the integration of *sequence prediction* as the *primary function* of the system. These properties are implemented in the interaction of artificial cortical mini-columns. Research on HTMs has steadily increased over the past five years, with a focus on image processing [13].

HTMs are constructed by hierarchically arranging regions of cortical columns, where each column is a set of neurons with associated dendrites and synapses. Within each region, two functional processes cooperate to learn temporal sequences from their input, and then pass their learned patterns to the region above. A Spatial Pooler (SP) operates on the input first, with the objective of learning sparse, distributed representations. The spatial codes are then used by a Temporal Pooler (TP) to learn sequences within the data stream.

The spatial and temporal patterns learnt by HTM are represented by the activation levels of columns, rather than the responses of neurons. The role of neurons in HTM is to collectively determine the activity of the column. By adding columns as a feature of the model, a closer match to cortical structure [17] is achieved which permits more sophisticated processing.

Unlike other models, where synapses are associated with weights that modulate input signals, HTM dendrites are associated with *potential* synapses which become connected and *active* when their *permanence value* passes a certain threshold. Only dendrites with connected synapses relay their input to the column. All other synapses remain inactive, but potentially active if the column has not sufficiently participated in the learning process. The participation level of columns is controlled by inhibition, where strongly activated columns compete with and inhibit less active neighbours. The activation level of a column is determined by the sum of the inputs from its connected synapses.

Learning in SP is based on how well synapses of a column match the input to which they are connected. It is implemented by increasing permanence values of

potential synapses connected to active input, and decreasing the same parameter for active synapses connected to inactive input. This method of altering synapse permanence values models the well established principle of Hebbian learning.

In order to focus on the relationship between data preprocessing and the initial operations of HTM, i.e. SP processing, we did not make use of the TP in this study. We refer the interested reader to [12].

5 Experimental Study

Datasets: Three bioacoustic datasets were used in this study. 'Frogs' are a set of 1,629 recordings of 73 different species of native Australian frog calls [18]. They are 250 milliseconds in duration, sampled at 22.05 kHz and 16 bits. The 'Insects' dataset consists of 381 insect species sounds, 5 seconds in duration, sampled at 44.1 kHz and 16 bits. The insects are categorised into four families: katydid, cricket, cicada, and others (i.e. bee, beetle, fruitfly, midges, mosquito, wasp).[1] The 'Birds' dataset was taken from the ICML 2013 Bird Challenge.[2] We used only the training files as the ground truths for the test set are not available. The training set comprises song recordings of 35 species of birds, 150 seconds in duration, sampled at 44.1 kHz and 16 bits.

Following the method of Gonzalez [7], we processed Insects using 1,024 samples per frame and no attempt was made to detect and remove ambient noise from the recordings. In order to accurately capture the lower frequencies typical of Frog calls, we increased the frame size to 5,120 samples. Because noise was not removed from the data, the Birds set produced a very large number of noise-only frames which distorted results. To counterbalance the extreme ratio of feature-to-noise frames, we increased the Bird frames to 32,768 samples.

Preprocessing: Wavelet features were obtained using Matlab R2012a and the UviWave.300 wavelet toolbox[3] running on a MacBook Pro with OS X Mavericks version 10.9.7. We used the UviWave.300 Morlet WT to produce individual scalograms for each sample frame. For gammatone/ERB wavelet features, we extended the UviWave.300 toolbox by developing an ERB scaled gammatone wavelet function to replace the Morlet function when producing scalograms.

To dimensionally reduce the scalograms we took the mean of each frequency band. For each dataset we produced both Morlet and gammatone/ERB features of two different sizes, either 36 or 100 coefficients. These feature set sizes were chosen as being the closest possible match to the feature set sizes used in [7], and which could be processed by the SP (which, for optimal performance, currently relies on input being a square matrix).

[1] Compiled by Gonzalez [7] from various internet resources.
[2] Kaggle. `https://www.kaggle.com/c/the-icml-2013-bird-challenge/`
[3] Universida de Vigo, Spain. `http://www.tsc.uvigo.es/~wavelets/uvi_wave.html`

Spatial Pooling: Over the past few years the SP has been incrementally developed and used in a range of vision processing studies. The current version allows input of multiple channels per instance as separate matrices, but in this study we disabled this feature, and used only a single input matrix. The dimension of SP columns was altered to match the dimension of input features, e.g. for 36 wavelet coefficients, 36 columns arranged in a six-by-six matrix were used.

The SP was run either to convergence, or for a maximum of 500 iterations at which time the 'best state' was used to obtain SP column codes. The best state was determined as the iteration during which the least number of synapses had their permanence values altered. SP column codes were output as the level of column activation, i.e. the sum of the input for all active synapses of the column.

Classification: Using a k-Nearest Neighbour (k-NN) classifier ($k = 1$), we performed ten-fold cross validation to evaluate all feature sets, which included features obtained after preprocessing by WTs, and those output by the SP. In [7], ten-fold cross validation using a k-NN classifier was also employed, so we are able to compare our Frog and Insect results against those achieved using spectral features.

6 Results and Discussion

The results of using WTs and the SP in this study are reported as the percentage of correcly classified instances. Table 1 summarises the results obtained using the Morlet and gammatone/ERB wavelets. Results previously obtained in [7] are provided for comparison. Due to our not being able to validate the test set of the ICML 2013 Bird Challenge, we cannot directly compare our results with those achieved in the competition. Nevertheless, we provide results on the Birds dataset as an extension to the range of bioacoustics investigated in this study.

Table 1. Percentage of correctly classified instances using the Morlet and gammatone/ERB WTs. Results from [7] are listed for comparison in the right hand columns.

	36 Features		100 Features			
	Morlet	+SP	Morlet	+SP		
Frogs	**90.7%**	53.3%	90.0%	59.6%		
Insects	**78.4%**	69.5%	78.1%	71.5%		
Birds	**71.3%**	43.5%	70.1%	53.9%		
	36 Features		100 Features		32	96
	G/ERB	+SP	G/ERB	+SP	Features	Features
Frogs	96.2%	60.5%	**96.3%**	69.9%	90.5%	87.0%
Insects	99.3%	91.1%	**99.5%**	94.6%	99.2%	98.6%
Birds	**92.8%**	47.3%	93.5%	53.6%	—	—

Both the gammatone/ERB and Morlet wavelets performed well on the Frogs dataset. The 100 coefficients of the gammatone/ERB feature set produced a correct classification rate of 96.3%, closely followed by 36 features which produced 96.2%. These results are a significant improvement on the 90.5% from [7].

The gammatone/ERB features performed best for both the Insects and Birds sets, achieving 99.5% with 100 features for Insects, and 94.2% with 36 features for Birds. The gammatone/ERB result for Insects is an improvement on the previously reported classification rate of 99.2%, and worth noting as any increase at these high levels of classification is difficult to achieve.

The use of the biologically-inspired gammatone/ERB wavelet consistently outperforms the linearly scaled Morlet wavelet on these datasets. We attribute this to the finer acuity achieved by the scaled wavelets, particularly in the higher frequency ranges typical of Insect and Bird sounds. The improvement on Frogs is less pronounced, as the ERB formula is less applicable at the lower frequency range for our limited sample frame size.

The inclusion of the SP reduced all classification accuracies (although not to the same degree for the Insects set[4]) suggesting that the SP encoding is degrading the salience of the gammatone/ERB wavelet features, at least the features relevant for a k-NN classifier. There are several possible explanations for this poor performance: (i) that the gammatone/ERB wavelet is not a sufficiently accurate model of the cochlea signal; and/or (ii) that HTM is not a sufficiently accurate model of the neocortex; and/or (iii) that additional processing of the cochlea signal in the brainstem[5] changes the characteristics of the auditory signal so that it becomes suitable for neocortical processing.

7 Conclusion

We have presented details of preliminary work aimed at developing the HTM model for auditory recognition and classification of bioacoustics. A biologically-inspired processing pipeline using WTs and the HTM SP was applied to three bioacoustic datasets and evaluated based on classification accuracy. These results showed that using gammatone/ERB WTs alone produced superior performance over previously published results for both frog call and insect sound classification. However, the inclusion of the SP caused classification rates to decline across all datasets. This suggests that the combination of gammatone/ERB WTs with an HTM spatial pooler does not accurately model the biological interaction between

[4] Unlike Frogs and Birds, the Insects samples are continuous and do not contain ambient noise. This suggests that the different context of the noise frames is impacting on the performance of the SP for Frogs and Birds.

[5] The cochlear nucleus of the brainstem provides considerable input to auditory processing due to a wide variety of neurons having distinct temporal and spectral response properties, e.g. cells of the posteroventral cochlear nucleus respond strongly to temporal features of complex tones. Higher within the brainstem, the superior olivary complex engages in binaural processing, while other regions of the brainstem handle reflexive and emotional responses to sound.

the cochlea and the neocortex. Further work is therefore required, particularly in studying and modelling the effects of brainstem activity on the auditory signals reaching the first region of the neocortex. We conjecture that such additional effects may be necessary for the neocortex (and an HTM system) to effectively classify bioacoustic data streams.

References

1. Murty, K.S.R. and Yegnanarayana, B.: Combining evidence from residual phase and mfcc features for speaker recognition. Signal Processing Letters, IEEE 13(1), 52–55 (2006)
2. Strang, G.: Wavelet transforms versus Fourier transforms. Bulletin of the American Mathematical Society 28(2), 288–305 (1993)
3. Mallat, S.: A wavelet tour of signal processing: The sparse way. Academic Press (2008)
4. Friston, K.: The free-energy principle: a unified brain theory? Nature Reviews Neuroscience 11(2), 127–138 (2010)
5. Towsey, M.W., Planitz, B., Nantes, A., Wimmer, J. and Roe, P.: A toolbox for animal call recognition. Bioacoustics: The International Journal of Animal Sound and its Recording 21(2), 107–125 (2012)
6. Mitrovic, D., Zeppelzauer, M. and Breiteneder, C.: Discrimination and retrieval of animal sounds. Multi-Media Modelling Conference Proceedings, 12th International, 339–343 (2006)
7. Gonzalez, R.: Better than mfcc audio classification features. The Era of Interactive Media, 291–301 (2013)
8. Stowell, D., and Plumbley, M.D.: Audio-only bird classification using unsupervised feature learning. Working Notes of CLEF 2014 Conference, (2014)
9. Seekings, P. and Potter, J.R.: Classification of marine acoustic signals using wavelets & neural networks. 8th Western Pacific Acoustics Conference (Wespac8), Proceeding of, (2003)
10. Mirzaei, G., Majid, M.W., Ross, J., Jamali, M.M., Gorsevski, P.V., Frizado, J.P. and Bingman, V.P.: The bio-acoustic feature extraction and classification of bat echolocation calls. Electro/Information Technology (EIT), 2012 IEEE International Conference on, 1–4 (2012)
11. Hawkins, J., and Blakeslee, S.: On Intelligence. Henry Holt, New York (2004)
12. Hawkins, J., Ahmad, S. and Dubinsky, D.: Hierarchical temporal memory including HTM cortical learning algorithms. Tech. Rep., Numenta, Inc., Palto Alto (2011)
13. Cowley, B., Kneller, A. and Thornton, J.: Cortically-inspired overcomplete feature learning for colour images. PRICAI 2014: Trends in AI, 720–732 (2014)
14. Schnupp, J., Nelken, I. and King, A.: Auditory Neuroscience: Making Sense of Sound. MIT Press (2011)
15. Valero, X. and Alías, F.: Gammatone wavelet features for sound classification in surveillance applications. Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European, 1658–1662 (2012)
16. Glasberg, B.R. and Moore, B.C.J.: Derivation of auditory filter shapes from notched-noise data. Hearing Research 47(1), 103–138 (1990)
17. Mountcastle, V.B.: Introduction to the special issue on computation in cortical columns. Cerebral Cortex 13(1), 2–4 (2003)
18. Stewart, D.: Nature Sound. Australian Frog Calls: Subtropical East. [Audio Recordings] Available from `http://www.naturesound.com.au/cd_frogsSE.htm`