Birdsong Recognition with DSP and Neural Networks

Alex L. McIlraith and H.C. Card
Department of Electrical and Computer Engineering
University of Manitoba
Winnipeg, Manitoba
R3T 5V6
Email: mcilrth@ee.umanitoba.ca, hcard@ee.umanitoba.ca

Keywords -- **speech recognition, birdsong, temporal processing, artificial neural networks, LPC analysis.**

Abstract -- **Analysis of speech often begins with study of the vocal tract that created it. Bird vocalizations and human speech are generated by similar processes. This suggests that LPC coefficients extracted from birdsong samples could retain enough information to permit identification of species. In this paper we train a back-propagation neural network to recognize bird songs. We generated test and training data sets using 133 songs from six common bird species. Initially, identification performance was good for some species, and poor for others. We attributed this to a lack of temporal context information in the data. By changing the type of spectral information presented to the network, we were able to improve performance. We conclude that a neural network combined with digital preprocessing can be used to identify a bird by its song.**

I. INTRODUCTION

Natural selection has generated a diversity of songs, coloration and behaviour in birds. While the ultimate purpose of these characteristics is to attract mates, they also provide reliable clues about a species' identity. In fact, song is so distinct in birds that one can often identify a species by its song alone. Geographically isolated populations within a species can sometimes be distinguished by humans in this way [1]. It is often possible to identify an individual by its song. This is analogous to the concepts of language, regional accents and speaker dependence variation in human speech.

Although there is much ecological and acoustical research interest in bird song there appears very little published work on automated bird song recognition. Several facts suggest however that this task is tractable. Birds generate sound in a similar manner to the way in which humans generate speech. In voiced speech, the vocal tract, mouth and nose act as a filter that shapes the periodic excitation from the vocal cords [2]; this is the source-filter model of speech production. In birds, variation in the shape of the respiratory tract, air flow rates, tongue position and muscle control of the syrinx (the structure responsible for excitation of the vocal tract in birds) all affect the sounds produced [3],[4]. Hearing and auditory processing are similar in birds and humans [5]. These similarities coupled with the observed ability of humans to distinguish birds by their song suggest that techniques used in speech analysis could be applied to bird identification by song.

If one assumes that vocalizations of both birds and humans can be modeled by source-filter interactions, one may then characterize these vocalizations by extracting filter coefficients from the time domain waveforms. Linear predictive coding (LPC) fits the impulse response of an all-pole filter to such a waveform. LPC coefficients are calculated for a frame of a sampled vocalization by first calculating reflection or autoregression parameters. Performing an FFT on time domain LPC coefficients then provides a smoothed representation of the spectral envelope [6]. LPC analysis has been used successfully in speaker identification [6], in speech recognition [7] and in speech compression [8].

As with isolated word recognition, there are several difficulties to overcome before we can attempt automated birdsong recognition. Variability of vocalizations due to dialect and individual differences obscures word and song identity. The temporal nature of sound also creates both dimensionality and context problems.

IEEE CAT. NO. 95CH3581-6/0-7803-2741-1/95/$3.00 © 1995 IEEE

To deal with variability, and the lack of theoretical information on the structure of bird songs, we chose to use a neural network for the recognition task. Their ability to perform non-linear mapping without requiring parametric assumptions has been successfully employed for pattern classification and speech recognition [7],[9].

Using LPC coefficients automatically reduces the dimensionality of input vectors. Taking the FFT of the LPC coefficients also provides a means of data reduction. Phase information can be discarded since human speech processing appears to be relatively unaffected by phase information [10]. Thus use of LPC analysis and removal of phase information results in useful lossy compression of sampled data.

Temporal context can be incorporated into a word recognition system that utilizes neural networks in several ways [7]. The most obvious method is to present a network with the entire spectrogram as an input vector (using FFT coefficients for all frames in a vocalization). Such a system has been used effectively to model the target recognition that bats appear to perform [11]. This technique, however, has a high computational cost since large numbers of weights must be updated in training. Others have used time-delay neural networks, in which both present and a few past frames provide input parameters to the network. Waibel used this technique together with the concept of large modular networks to recognize consonants [12]. Recurrent networks that have context feedback units have also found applications [13].

One problem with word recognition is that it is computationally intensive. With neural networks, a way to reduce computation is to limit the number of nodes in the network. In addition, proper coding of input data is critical [14]. For this reason, we chose to encode temporal context through preprocessing rather than expecting the network to discover temporal context from a high-dimensional input. Our goal was to demonstrate that a simple neural network could identify a bird species based upon its song; we refer to this as species recognition.

II. METHODS

A. *Sampling*

Data for the study was extracted from audio tapes and compact disks that were intended for use by people wishing to identify species by song [15-20]. Birds have a variety of vocalizations. It is generally the male that sings in order to attract mates, and to mark its territory. Calls and call notes are produced by both sexes. They are generally short in duration and are used to signal such things as alarm or recognition. We chose to analyze only male song, rather than calls or call notes, because the latter are often similar between species [4]. Six species of bird native to Manitoba were chosen: Song Sparrow (SON), Fox Sparrow (FOX), Marsh Wren (MWR), Sedge Wren (SWR), Yellow Warbler (YLW) and Red-winged Blackbird (RWB) [21]. They were chosen to provide a representation of both long and short songs. 133 songs from these species were digitized using a PC sound card (11.025kHz sampling rate, 8 bit resolution). Digitization was performed with AGC, and levels were adjusted to give maximum amplitude without clipping. This was done to remove variation in the FFT magnitudes due to amplitude variation.

B. *Preprocessing*

Data was left-justified by hand and manipulated by the software package Hypersignal Plus, which preprocessed the data in several steps:

1. Framing -- a non-overlapping Hamming window, 256 samples (approx. 23.2 msec) wide was applied. Frame widths for speech recognition are normally in the order of 10 msec [7].

2. LPC -- 16 time domain coefficients for a 15th order LPC filter were generated for each frame.

3. FFT -- a fast fourier transform of the 16 LPC coefficients was produced with 9 unique spectral magnitudes.

This procedure was repeated with a 1024 sample window for all songs. The resulting data were exported from Hypersignal, with each song being represented by a number of records, each containing nine spectral coefficients. Further processing of the data was required before it could be used for training the neural network. Initial investigation revealed to us that the overall length of a bird's song was an important cue in species identification. Adding such a variable often helps a network learn from hints [22]. For this reason, we included a variable to represent the length of a song. Its

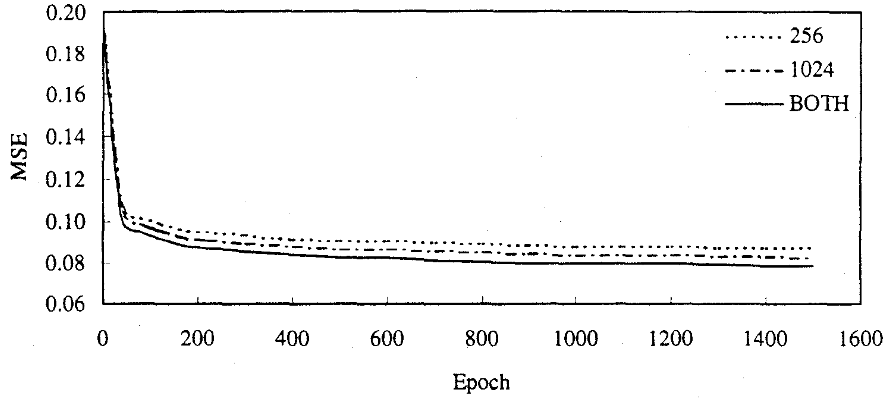

Fig. 1. Learning curves for 256 sample window, 1025 sample window and combined data sets.

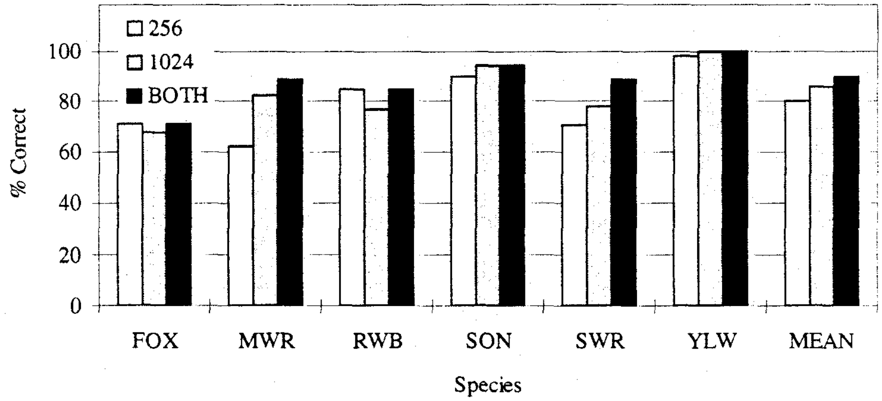

Fig. 2. Species recognition performance for networks trained with the three types of data set.

value was the same for each record within a given song. Spectral and time variables were normalized to a mean of zero and a standard deviation of one. Variables were squashed using a logistic function with a gain of unity.

Three types of data set were created. Two had 10 variables, with data either from the 256 sample window (9) or from the 1024 sample window (9) and the song length (1). The third included all 19 variables; records from the 1024 sample window were repeated in a manner similar to that used for song length.

In this study we chose 'vanilla' back-propagation, without momentum or higher order derivative information [23] as our learning model. Training was accomplished using the PDP [24] back propagation algorithm. Some experimentation suggested that a network of 10 or 19 inputs, 12 hidden nodes and 6 outputs would work well. The learning rate was set to 0.2. Target values of 0, 1 were changed to 0.2 and 0.8 to accelerate learning..

Songs were divided between test and training data sets in random order (records within songs were not randomized at this stage). The proportion of data used for testing was 25% in all runs. To enhance cross-validation, 10 training and 10 test sets were generated for each of the three data set types, and the network was trained with new initial weights each time. The 10 test and 10 training sets were generated in the same order for all three data sets in order to facilitate comparisons. Preliminary training runs of 10000 epochs suggested that a run length of 1500 epochs was sufficient for the mean sum-of-squares error (MSE) to reach a stable value in all cases.

Final weights were read by a C program that computed the six output values for each of the test set records and an error sum-of-squares (for target vs. expected outputs). For each song, any activation > 0.6 was tallied for the 6 species outputs. The species class with the largest tally was considered to be the winning class. If two classes were tied for the maximum tally, classification was considered to be incorrect.

III. RESULTS

Although the differences appear small in Fig. 1, the mean MSE during training is largest for the 256 sample window data, smaller for the 1024 window and is smallest for the combined data set. These differences are significant over the duration of the training cycle. Two-tailed t-tests [25] comparing the mean MSE (between 256 and 1024, 1024 and 'Both', n = 10) at epoch 1500 indicate p-values << 0.000001. This means that there was less than one chance in a million that the mean final MSE values were different due to chance.

Fig. 2 shows that the network was able to recognize the six species by their song. Overall performance ranged from 80% to 85% correct identification. Except for the FOX and RWB, mean performance increased as the input was changed progressively from 256 to 1024 and to combined data sets.

Table I shows that there existed considerable overlap in song length. Song lengths of YLW, MWR and SON had bimodal frequency distributions. When we examined the raw results from the runs, we noticed that some songs were consistently misidentified. In two cases, a song was either atypically long or short, and was misidentified regardless of the data set used. Other songs were misidentified with 256 and 1024 window data sets, but were correctly identified with the combined data set.

TABLE I
LENGTH OF SONG

Statistic	RWB	YLW	MWR	SWR	FOX	SON
Mean	445	525	789	857	1244	1466
Standard deviation	83	154	209	110	141	227
n	17	27	32	13	13	31

Times are in milliseconds.

IV. DISCUSSION

This study can be described as speaker-independent song recognition. Bird songs used in this study had been recorded from individuals in a variety of locations, with differing quality and amounts of background noise. The bimodal song length distributions of some species may indicate that these species have two classes of songs, or that there are two dialects in the data set. For the RWB songs, a human listener can detect distinct song classes. The overall performance achieved with the 256 sample window was acceptable; however, increasing the window size and combining two resolutions of spectral data substantially improved performance of the neural network. Both of these modifications add temporal context in addition to that provided by the length variable. The type of information available to the network due to preprocessing is not unlike that extracted by the vertebrate auditory system.

It is well known that the cochlea performs a time to frequency transformation. Recent work with bats [26], and barn owls [27] points to sophisticated preprocessing of spectral data in vertebrates. Both appear to have neurons that serve as delay lines and coincidence detectors. Bats measure the time difference between sonar emissions and echoes, and thus can assess the distance between themselves and an object. The barn owl accurately locates a prey item if it makes a sound. It does so by comparing the time of arrival and amplitude of sounds reaching its two ears. Carr [28] indicates that such delay elements may be widespread in the nervous system of vertebrates. Dear *et al.* [29] have shown that bats have neural mechanisms to assemble a complete sonar image 40 msec after emitting a pulse. They do so because the neurons activated by the rapidly returning echoes have a long latency; they remain activated until the late arriving echoes arrive.

It is not unreasonable to expect that populations of neurons with latency variation could permit vertebrates to compile an integrated picture of a short sound. This suggests that vertebrates may also be able to combine short-term and long-term spectral information, and easily learn to distinguish sounds. Our neural network / preprocessing scheme is one model of such a system.

V. CONCLUSIONS

The use of preprocessing to reduce data dimensionality, and to encode temporal context, permits a simple neural network to recognize bird songs. The discovery of similar preprocessing in the vertebrate brain would not be surprising based on performance of this simple recognition system. Continuing research is exploring the application of unsupervised learning and the community of experts network model to this problem.

ACKNOWLEGEMENTS

We wish to thank Dean McNeill and members of the Ganglion research group for their helpful input. This research was supported by NSERC and by a University of Manitoba Fellowship to A.L. McIlraith.

REFERENCES

[1] G. Thielcke, "Geographic variation in bird vocalizations," in *Bird Vocalizations*, R. H. Hinde, Ed. Cambridge, UK: Cambridge University Press, 1969, pp. 311-339.

[2] P. Lieberman and S. E. Blumstein, *Speech Physiology, Speech Perception and Acoustic Phonetics*. Cambridge, UK: Cambridge University Press, 1988.

[3] J. H. Brackenbury, "Functions of the syrinx and the control of sound production," in *Form and Function in Birds* , A. S. King and J. McLelland, Eds. London: Academic Press, 1989, pp. 193-220.

[4] W. H. Thorpe, *Bird - Song*. Cambridge, UK: Cambridge University Press, 1961.

[5] R. J. Dooling, "Auditory perception in birds," in *Acoustic Communication in Birds*, D. E. Kroodsma, E. H. Miller and K. Ouellet, Eds. New York: Academic Press, 1982, vol. 1, pp. 95-129.

[6] B. S. Atal, "Linear predictive coding of speech," in *Computer Speech Processing*, F. Fallside and W. A. Woods, Eds. London: Prentice Hall, 1985, pp. 81-124.

[7] R. P. Lippmann, "Review of neural networks for speech recognition," *Neural Computation*, vol. 1, pp. 1-38, 1989.

[8] L. R. Rabiner, "Applications of voice processing to telecommunications," *Proc. IEEE*, vol. 82, pp.199-228, 1994.

[9] R. P. Lippmann, "Pattern classification using neural networks," *IEEE Comm. Magazine*, vol. 27, no. 11, pp. 47-64, 1989.

[10] P. B. Denes and P. N. Elliot, *The Speech Chain*. New York: W. H. Freeman and Co., 1993, 2nd ed.

[11] I. E. Dror, M. Zagaeski and C. F. Moss. "Three-dimensional target recognition via sonar: a neural network model," *Neural Networks*, vol. 8, pp. 149-160, 1995.

[12] A. Waibel "Modular construction of time-delay neural networks for speech recognition," *Neural Computation*, vol. 1, pp. 39-46, 1989.

[13] M. D. Hanes, S. C. Ahalt, and A. K. Krishnamurthy, "Acoustic-to-phonetic mapping using recurrent neural networks," *IEEE Trans. Neural Networks*, vol. 5, pp. 659-662, 1994.

[14] J. Hertz., A. Krogh and R. G. Palmer, *Introduction to the Theory of Neural Computation*. Redwood City, California: Addison-Wesley Publishing Co., 1991.

[15] M. Brigham, *Bird Sounds of Canada*. Mount Albert, Ontario: Holborne Dist. Co. Ltd, no date.

[16] D. J. Borror, *Songs of Eastern Birds*. New York: Dover, 1970.

[17] D. J. Borror, *Common Bird Songs*. New York: Dover, 1967.

[18] L. Elliot and T. Mack, *Wild Sounds of the Northwoods*. Ithaca, N.Y.: NatureSound Studio, 1990.

[19] R. K. Walton and R. W. Lawson, *Birding by Ear (Eastern / Central) - A Guide to Bird-song Identification*. Boston: Houghton - Mifflin, 1989.

[20] P. P. Kellogg, R. T. Peterson and W. W. H. Gunn, *A Field Guide to Western Bird Songs*. Boston: Houghton - Mifflin, 1975.

[21] C. S. Robbins, B. Bruun, and H. S. Zim, *Birds of North America - A Guide to Field Identification*. New York: Golden Press, 1983.

[22] S. Haykin, *Neural Networks - a Comprehensive Foundation*. New York: Macmillan College Publishing Co., 1994.

[23] D. E. Rumelhart, F. E. Hinton and R. J. Williams, "Learning representations by back-propagation of errors," *Nature*, vol. 323, pp. 533-536, 1986.

[24] J. L. McLelland and D. E. Rumelhart, Explorations in Parallel Distributed Processing. Cambridge, Mass.: MIT Press, 1989.

[25] G. K. Bhattacharyya and R. A. Johnson, *Statistical Concepts and Methods*. New York: Wiley., 1977.

[26] N. Suga, "Biosonar and neural computation in bats," *Sci. Amer.*, vol. 262, no. 6 , pp. 60-68, 1990.

[27] M. Konishi, "Listening with two ears," *Sci. Amer.*, vol. 268, no. 4, pp. 66-73, 1993.

[28] C. E. Carr, "Processing of temporal information in the brain," *Annu. Rev. Neurosci.*, vol. 16, pp. 223-263, 1993.

[29] S. P. Dear, J. A. Simmons and J. Fritz, "A possible basis for representation of acoustic scenes in auditory cortex of the big brown bat," *Nature*, vol. 364, pp. 620-623, 1993.