Sensor Network for the Monitoring of Ecosystem: Bird Species Recognition

Jinhai Cai [1], Dominic Ee [2], Binh Pham [1], Paul Roe [1], Jinglan Zhang [1]
Microsoft-QUT eResearch Centre, Faculty of Information Technology,
Queensland University of Technology, Brisbane, QLD 4001, Australia
[1] *j.cai,b.pham,p.roe,jinglan.zhang@qut.edu.au*
[2] *m.ee@student.qut.edu.au*

Abstract

In this paper, we investigated the performance of bird species recognition using neural networks with different preprocessing methods and different sets of features. Context neural network architecture was designed to embed the dynamic nature of bird songs into inputs. We devised a noise reduction algorithm and effectively applied it to enhance bird species recognition. The performance of the context neural network architecture was comparatively evaluated with linear/mel frequency cepstral coefficients and promising experimental results were achieved.

1. Introduction

Since the early of 2007, we have established a sensor network at the Samford Ecological Research Facility (SERF), in Brisbane, Australia. This facility, which is very close to Brisbane urban suburbs and Brisbane Forest Park, is an idea place to study the impact of urbanisation of neighbouring suburbs on the ecological system of Samford. Our sensors are designed to record sounds and images. We find that both sounds and images are very useful for different applications. Using images is advantageous over using sounds in research areas related to botany. However, there are many advantages in using sounds in research areas related to ornithology and acoustical-environmental science. Sounds have been widely used in the recognition of bird and insect species, and the estimation of acoustical environment health. In this paper we focus on the recognition of bird species using acoustic signals, as bird species recognition is one of vital tasks of our project: Sensor Network for the Monitoring of Ecosystem.

Currently, there are two main groups of scientists who are interested in bird species recognition: ornithologists and pattern recognition researchers. Usually, ornithologists study acoustic signals by listening and then identify birds using their professional knowledge, where most ecological sampling is conducted at small spatial scales or consists of infrequent or one-time sampling [1]. As the manual process conducted by most ornithologists is very tedious and time consuming, it is impractical for large scale surveys or studies. Recently, many pattern recognition researchers are interested in bird recognition due to potential applications such as minimising bird striking risks in aviation industry and preserving biodiversity in government development planning. Pattern recognition researchers have developed several algorithms for automatic bird species recognition [2], [3], [4].

Sensor networks bring ornithologists and pattern recognition researchers together to make some applications possible. These applications include real-time assessing risks from potential bird collisions between birds and airplanes, unobtrusive observations (where the presence of humans changes some animal behaviours), and the studies of spatial and temporal variation in biological processes [1].

As far as bird species recognition algorithms are concerned, most algorithms for speech recognition have been attempted by many researchers. Among them, the most popular algorithms for bird species recognition are hidden Markov models, multilayer perceptron (MLP) neural networks, and support vector machines. A hidden Markov model is able to integrate high level language statistics and low level feature statistics into one speech recognition system. Due to this, hidden Markov models are vital parts of all commercial speech systems. Support vector machines have become a popular tool in machine learning tasks. One major advantage of using support vector machines is the superior generalization properties they offer when compared to other types of classifiers. Therefore, the performance of support vector machines is comparable to that of other classifiers if the training set is small. We are particularly interested in using neural networks for bird species recognition. Currently, most neural based systems using frame-based features as inputs. However, it is difficult for this approach to model the dynamic process of bird songs. To solve this problem, we use features from "past" and "future" frames as well as the current frame as inputs to the neural network ("context" MLP neural network) for bird species recognition.

Section 2 presents a novel noise reduction algorithm which is capable of estimating noise from frames with or without signals (noise-only). Section 3 deals with the feature extraction process using mel-frequency cepstral coefficients (MFCC) as features. The design of the context neural network architecture is covered in Section 4, while the comparative analysis of its performance versus that of other approaches is discussed in Section 5. Section 6 provides conclusion and the direction for our future work.

2. Pre-processing: Signal Enhancement

The quality of bird databases used by most pattern recognition researchers is generally high, therefore signal enhancement is not applied on original data in many systems [2], [3], [4], [5], [6]. However, the good quality of the published sound databases is the result of manual selection and editing process, which is to select small part of good quality data from raw data. Consequently, data with low Signal-to-Noise Ratio (SNR) caused by rains and winds are removed. Due to this kind of important information loss, it is impossible for us to investigate bird behaviours/songs under rainy and windy conditions from these databases.

Sensor networks can recode data according to a predefined schedule and send data back to our server. The SNR of recorded data varies from 30 dB to -10 dB mainly due to different weather conditions, the distances of birds from the sensors, and surrounding environments. The noise level can change from time to time due to different conditions such as windy weather or flyover airplanes. Therefore, it is essential that the signal enhancement algorithm is able to handle different noises and to track the noise level. In the past decades, researchers have developed many noise reduction algorithms including Wiener filter methods [7], signal subspace methods [8], and statistical methods [9], [10]. All these algorithms assume that the first few frames of the signal consist of noise only so that the initial noise estimation can be obtained from the first few frames. The noise estimation is usually updated from noise-only frames which are determined by voice (signal) activity detectors (VADs). This approach is reasonable for speech enhancement as speakers are cooperative. However we cannot assume that the first few frames of the recoded data by our sensors are noise only without bird calls. Sometimes there is almost no gap between bird calls; therefore there are not enough noise-only frames to update noise estimation. Moreover, VADs do not work well in low SNR signals.

A novel noise reduction algorithm developed by Cai is shown in Figure 1. This algorithm is able to estimate noise from any frames with or without signals. Therefore, it does not need a VAD, which is error prone. We now briefly discuss the algorithm and give some examples to demonstrate properties of the algorithm.

In the first example, a sensor in our network recoded calls of a powerful owl in the distance together with calls of crickets and other insects. Figure 2 shows waveforms of the original sound and the sound after noise reduction. Figure 3 shows spectra of the original sound and the sound after noise reduction. This example clearly shows that our algorithm improves the sound quality significantly even though the SNR of the original sound is very low (about -10 dB).

The second example is a sound record without noise-only frames. Figre 4(a) shows spectra of the original sound and the sound after noise reduction by our algorithm. Figure 4(b) shows the spectrum after noise reduction by an algorithm based on minimum mean square error (MMSE) [9]. This example shows that our algorithm is able to get reasonable noise estimation from any frames. The MMSE-based algorithm [9] is one of the best algorithms for noise reduction, but it is not designed to estimate noise from frames with signals.

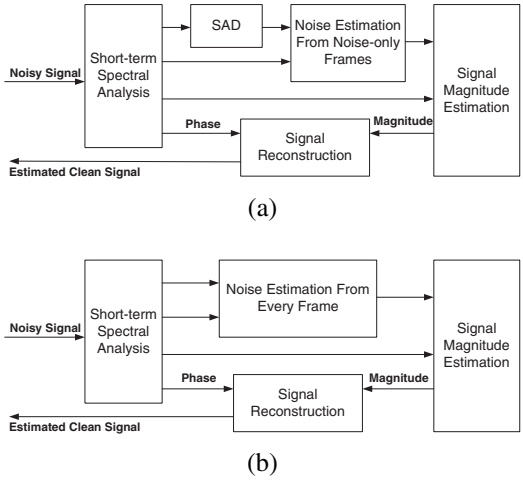


Fig. 1: Signal enhancement: (a) traditional MMSE; (b) new method without VAD

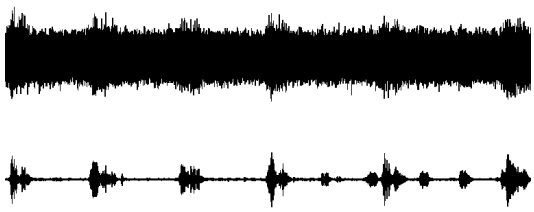


Fig. 2: Waveforms: calls of a powerful owl and insects.

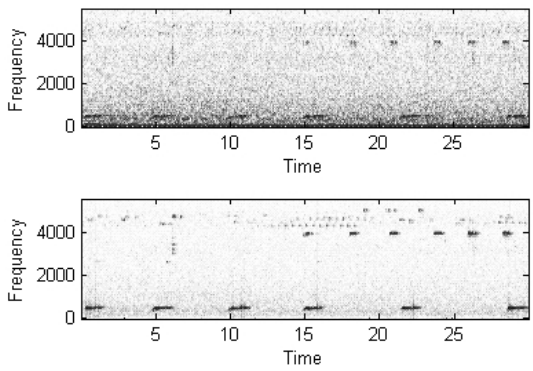
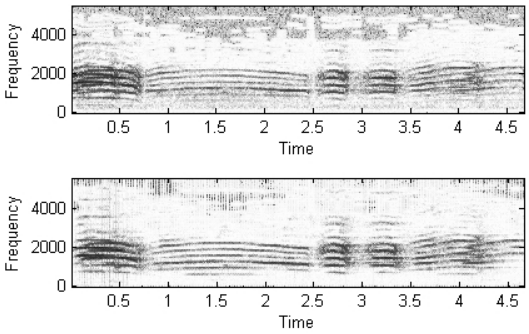


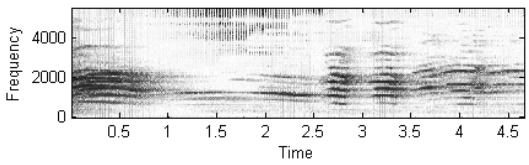
Fig. 3: Spectra: calls of a powerful owl and insects.

3. Feature Extraction

Human speech and birdsong have numerous parallels in terms of communication and its development. In both songbirds and humans, these sounds are produced by the flow of air during expiration through a vocal tract [11]. The bird's vocal tract, similar to the human vocal tract, acts as an acoustic filter that controls the sound. Hearing and auditory processing are also similar in birds and humans.



(a)



(b)

Fig. 4: Spectra: calls of two crows.

Due to these similarities, it is natural for researchers to adopt feature extraction methods used in speech analysis. Chen and Maher used spectral track method to extract features for bird classification and achieved good results ranging from 99% to 95% in terms of accuracy. This feature extraction method is similar to the formant tracking method used in speech processing in the 1980's. The main advantage is that spectral peaks are robust to noises. However, it cannot be used for aperiodic signal processing. Therefore, this method can be only used to extract features from a limited list of bird's songs.

Recently, almost all speech recognition systems use mel-frequency cepstral coefficients (MFCC) as features. The main advantages are as follows:

- The mel scale frequency is consistent with human auditory perception. Mel frequency features have been proven to be able to serve speech recognition systems better than linear frequency features.
- We are able to extract MFCCs from both aperiodic and periodic signals.
- Cepstral coefficients can achieve significant data reduction with reasonable information loss as demonstrated in Figure 5.

Due to above reasons, we used MFCCs as features for bird species recognition. The feature extraction method is depicted in Figure 6. For the purpose of performance comparison, we use linear scale cepstral coefficients as features as well. The method for extracting linear scale cepstral coefficients is the same as that for extracting MFCCs but without the mel scale conversion. The conversion from linear scale spectrum to mel scale spectrum is achieved by

$$F_{mel} = \begin{cases} F_{Hz} & \text{if } F_{Hz} \leq 1000 \text{ Hz}, \\ \dfrac{1000}{log2} log[1 + \dfrac{F_{Hz}}{1000}] & \text{if } F_{Hz} > 1000 \text{ Hz}, \end{cases} \quad (1)$$

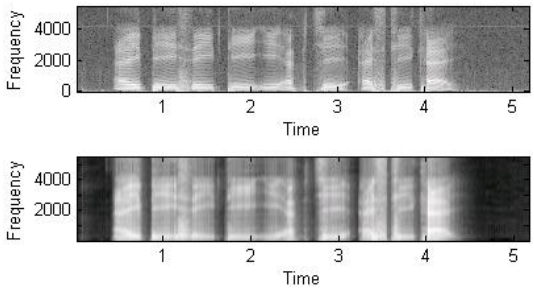


Fig. 5: Spectral data reduction. Top: Original spectrum using FFT with frame length of 512 points. Bottom: Spectrum recovered from 13 cepstral coefficients.

in which F_{mel} is the perceived frequency in mels and F_{Hz} is the real frequency in Hz.

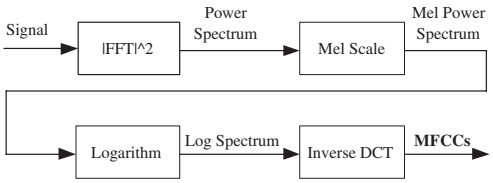


Fig. 6: Feature extraction method.

4. Artificial Neural Network with Dynamic Features

Artificial neural networks provide a general and practical machine learning approach that is robust to errors in the training data. It was inspired by the way that nervous systems in some mammals process information. It has been successfully applied to problems such as speech recognition and visual scene interpretation. Artificial neural networks have also been used to classify bird sounds. As multilayer perceptron neural networks (MLP NNs) are suitable for pattern recognition, they have been widely used for bird classification. In most neural-based bird recognition systems, frame-based features are directly used as inputs to MLP NNs [5]. However, it is difficult to extract temporal features from individual signal frames; while these temporal features are very sensitive to human auditory perception and visible to the unaided human eyes from spectra of bird songs. In order to alleviate this problem, two main techniques are commonly used:

- Differential features, such as delta MFCC features and delta-delta MFCC features, are used to model the changes between neighbouring frames.
- Time delay neural networks are used [12], where time delay neural networks take information from "past" as well from the current frame.

Dynamic Features: We use dynamic features for our neural network. The dynamic features are extracted by using simple differential operations:

$$\Delta\mathbf{MFCCs}(t) = \mathbf{MFCCs}(t+1) - \mathbf{MFCCs}(t-1), \quad (2)$$

where $\mathbf{MFCCs}(t)$ is a vector of mel-frequency cepstral coefficients at frame t. Now, we obtain the frame-based feature vector at time t as follows:

$$\mathbf{V}(t) = \{\mathbf{MFCCs}(t), \Delta\mathbf{MFCCs}(t)\}. \quad (3)$$

The input vector to the neural network at time t is

$$\mathbf{Input}(t) = \{\mathbf{V}(t-p), \cdots, \mathbf{V}(t), \cdots, \mathbf{V}(t+p)\}. \quad (4)$$

Neural Network: There are two kinds of time delay units in neural networks based on two principles. One is that the decision should be made based on the decision history as well as current observations. Many real world applications in speech recognition and bird identification support this principle as the duration of a syllable in speech and bird calls is longer than one frame. This principle results in time delay neural units illustrated in Figure 7. Another is that the decision should be made based on the acoustic context instead of the current observations only. Neural networks with input delays can perform context sensitive decisions, which resembles human perception. This principle results in time delay neural units for input features as shown in Figure 8. For speech phoneme recognition, time delay neural networks (TDNN), whose units are similar to that in Figure 8, can achieve excellent performance, which is comparable to the best performance of any other methods [12]. Therefore, we will use a neural network with the similar architecture to that of time delay neural networks.

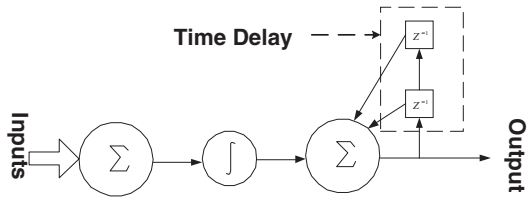


Fig. 7: Output time delay units for feedback neural networks.

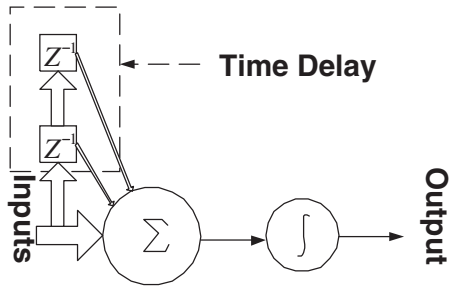


Fig. 8: Input time delay units for neural networks.

However, there is little information about the "past" for the onset of a phoneme in speech or an element of a syllable in bird song, thus the systems have to make decision mainly based on the information from the current frame only. While, human will delay the decision until we receive enough context information. Let take the recognition of '/too/ apples' and '/too/ big' as an example. Clearly, three words have the same pronunciation '/too/' in English, so we can't decide which word from the pronunciation '/too/' as there is no clue before '/too/'. But it is easy to know which word from the word after '/too/'.

In a similar way, we can overcome the problem of TDNN on the recognition of the onset of a syllable. We use features from "future" frames as well as the current frame and the "past" frames. This results in a context neural network architecture as shown in Figure 9.

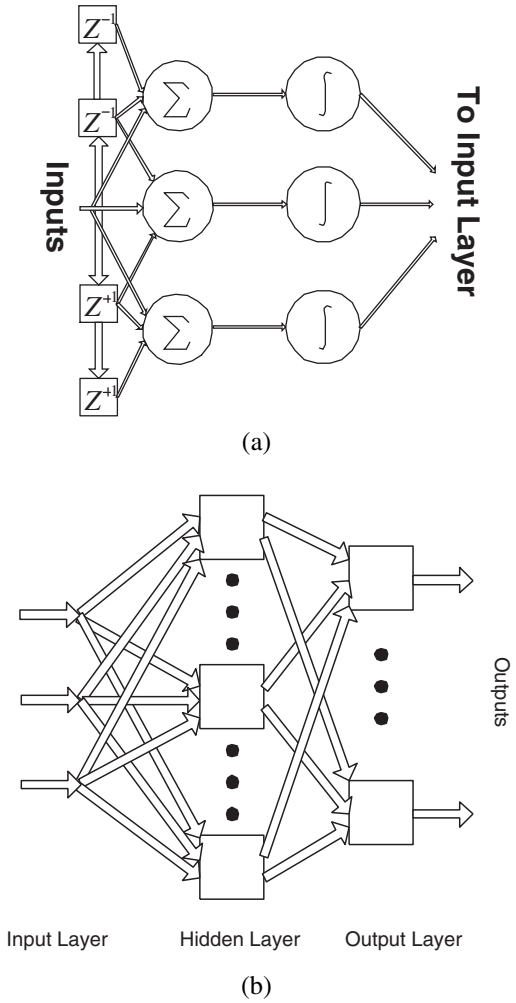


Fig. 9: Context neural network architecture: (a) Non-causal time delay units; (b) MLP units.

In our system, we use 13 dimensional $\mathbf{MFCCs}$ and 13 dimensional $\Delta\mathbf{MFCCs}$ as a feature vector for each frame. We set p in eqn(4) to 2, therefore there are 5 vectors as an input to the time delay units. The output of the time delay units contains tree 26-dimensional vectors, which are the input to the MLP part of the neural network. As the time delay units are non-causal, the input vector is delayed by two frames.

5. Recognition Results

In our experiments, the bird calls were from three sources: Birds in Backyards[14], Australian Bird Calls: Subtropical East [15] and Voices of Subtropical Rainforests [16], and the

data collected from our sensors at Samford. For each species, bird calls were divided into two sets: training set and test set in a ratio of 6 to 4.

In the first experiment, we used a small set of database containing calls from 4 bird species to test different neural network training methods for bird recognition. The four bird species are Golden Whistler (Pachycephala pectoralis), Eastern Rosella (Platycercus eximius), Eastern Spinebill (Acanthorhynchus tenuirostris), and Masked Lapwing (Vanellus miles). We employed 20 hidden units in the neural network. We used several algorithms to train the network and found that the Levenberg-Marquardt and the RPROP [13] algorithms are able to achieve the best performance, 98.7% of recognition rate, in terms of accuracy. The errors occur when the types of bird calls were not well represented in the training set and were closer to the types of calls from other birds. For example, some Golden Whistler's calls were misclassified as Eastern Rosella's calls. The misclassification is highlighted by a rectangle in Figure 10.

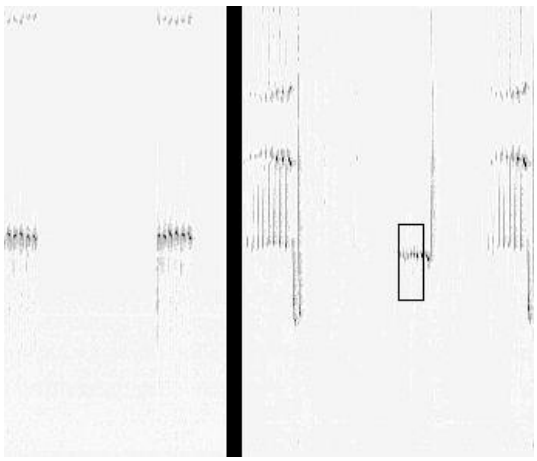


Fig. 10: Misclassification: Eastern Rosella and Golden Whistler.

As the RPROP [13] algorithm has a smaller memory footprint and requires much less training time to achieve the same performance as the Levenberg-Marquardt algorithm in the first experiment, we will use only the RPROP algorithm in the following experiments. In these experiments, calls of 14 bird species are used.

In the second experiment, we will evaluate the influence of noise reduction algorithms on the performance of bird species recognition. Two algorithms are evaluated in this experiment: our own method demonstrated in Section 2 and the MMSE [9]. The MMSE algorithm is one of best algorithms in speech enhancement. However it requires few silence/noise only frames at the beginning of each record. This assumption is reasonable in speech applications but not in our application. Our algorithm is also able to track slowly-changing noise. The performance in Table 1 confirms our analysis.

In the third experiment, we will evaluate the influence of hidden units of the neural network and features on the performance of bird species recognition. Table 2 shows the experimental results for the recognition of 14 bird species,

TABLE 1: THE PERFORMANCE OF 14 BIRD SPECIES RECOGNITION: COMPARISON BETWEEN TWO NOISE REDUCTION ALGORITHMS

	Accuracy % (Cepstral Coefficients)	
Hidden units	Our algorithm	MMSE
10	79.6	76.8
20	82.4	79.6
40	84.9	80.8
80	85.6	81.0
160	84.6	81.9

which are comparable to the results obtained by other reseachers [2], [5]. The results in this table have clearly shown that it is advantageous to use MFCCs as features over linear frequency cepstral coefficients for the recognition of bird species. The table also indicates that the number of hidden units in a neural network is important to the performance. In our experiments, the number of hidden units should be around 80 to give optimal results.

TABLE 2: THE PERFORMANCE OF 14 BIRD SPECIES RECOGNITION: CEPSTRAL COEFFICIENTS VS MFCCS

	Accuracy %	
Hidden units	Cepstral Coefficients	MFCCs
10	79.6	81.6
20	82.4	82.6
40	84.9	86.2
80	85.6	86.3
160	84.6	86.8

6. CONCLUSION AND FUTURE WORK

In this study, we have proposed the context neural network architecture to embed the dynamic nature of bird songs into inputs to a neural network. We have devised a noise reduction algorithm and effectively applied it to the pre-processing of bird songs. We have evaluated our proposed neural network architecture with the linear/mel frequency cepstral coefficients. We have achieved the recognition rate of 98.7% on a data set consisting of 4 bird species and 86.8% on an extended data set consisting of 14 bird species, which are comparable to the best published results in terms of accuracy.

We have analysed the performance of our system and found that the interference from calls of other birds or animals is still the major obstacle for bird species recognition. It is an open problem in this research field and our ongoing work will focus on this problem.

ACKNOWLEDGMENTS

This work is supported by Microsoft, Queensland State Government and QUT.

REFERENCES

[1] J. Porter, P. Arzberger, H-W Braun, etc, "Wireless Sensor Networks for Ecology", *BioScience*, Vol. 55, No. 7, 2005, pp. 561-572.

[2] C. Kwan, G. Mei, X. Zhao, etc., "Bird Classification Algorithms: Theory and Experimental Results", In: *Proc. Of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2004, Vol. V, pp. 289-292.

[3] P. Somervuo, A. Härmä, S. Fagerlund, "Parametric Representations of Bird Sounds for Automatic Species Recognition", *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 14, No. 6, 2006, pp. 2252-2263.
[4] Z. Chen, R. C. Maher, "Semi-automatic classification of bird vocalizations using spectral peak tracks Recognition", *Journal of Acoust. Am.*, Vol. 120, No. 5, 2006, pp. 2974-2984.
[5] A.L. McIlraith, H.C. Card, "A Comparison of Backpropagation and Statistical Classifiers for Bird Identification", In: *Proc. Of IEEE International Conference on Neural Network*, 1997, Vol. 1, pp. 100-104.
[6] A. Franzen, I. Gu, "Classification of Bird Species by Using Key Song Searching: A Comparative Study", In: *Proc. Of IEEE International Conference on Systems, Man and Cybernetics*, 2003, Vol. 1, pp. 880-887.
[7] J. S. Lim, A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech", *Proc. IEEE*, Vol. 67, 1979, pp. 1586-1604.
[8] H. Lev-Ari, Y. Ephraim, "Extension of the Signal Subspace Speech Enhancement Approach to Colored Noise", *IEEE Signal Processing Letter*, Vol. 10, 2003, pp. 103-106.
[9] Y. Ephraim, D. Malah, "Speech Enhancement using a Minimum Mean Square Error Short Time Spectral Amplitude Estimator", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 32, No. 12, 1984, pp. 1109-1121.
[10] Y. Ephraim, "Statistical model based speech enhancement systems", *Proc. IEEE*, Vol. 80, 1992, pp. 1526-1555.
[11] A.J. Doupe, P.K. Kuhl, "Birdsong and Human Speech: Common Themes and Mechanisms", *Annual Review of Neuroscience*, Vol. 22, 1999, pp. 567-631.
[12] M. Sugiyama, H. Sawai, A.H. Waibel, "Review of TDNN (Time Delay Neural Network) Architectures for Speech Recognition", In: *Proc. IEEE International Symposium on Circuits and Systems*, 1991, Vol. 1, pp. 582-585.
[13] M. Riedmiller, H. Braun, "A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm", In: *Proc. IEEE International Conference on Neural Networks*, 1993, pp. 586-591.
[14] Australian Museum, *http://www.birdsinbackyards.net/feature/top-40-bird-songs.cfm*.
[15] D. Stewart, *Australian Bird Calls: Subtropical East*, CD, Nature Sound, 2002.
[16] D. Stewart, *Voices of Subtropical Rainforests*, CD, Nature Sound, 2002.