Text-independent Speaker Identification in Birds

E.J.S. Fox[1], J.D. Roberts[1], M. Bennamoun[2]

[1]School of Animal Biology, University of Western Australia, Australia
[2]School of Computer Science and Software Engineering, University of Western Australia, Australia
foxe02@student.uwa.edu.au

Abstract

Speaker recognition is used to identify individual humans, but has rarely been applied to other species. To be applicable to the wide variety of bird species, text-independent speaker identification would be the most effective method. This is the first paper to report results of this technique in a species other than humans. Mel-frequency cepstral coefficients were extracted from recordings of three bird species and a multilayer perceptron was used as the classifier in each species. First, the song types used in training and testing were not controlled for, and these conditions gave an accuracy of 68-100%. Next the recordings of the wagtails and scrub-birds were split into their respective song types, a network was trained with one song type from each individual and tested with a different song type. With these purely text-independent conditions the accuracy was 71-96%.

Index terms: speaker recognition, artificial neural network, mel-frequency cepstral coefficients

1. Introduction

Many animal species are currently under threat and in decline. In order to know how to best conserve these species it is necessary to fully understand their biology, many aspects of which can only be determined through the study of known individuals over time. Most commonly these individuals are identified through the addition of external marks (for example radio transmitters, or leg bands on birds). However, this requires that animals are caught at least once and has the potential to influence survival and behaviour through stress, increased predation rates and other effects [1,2]. These methods are also of little use in species which are nocturnal, cryptic, difficult to catch or particularly prone to disturbance.

Individual identification based on aspects of natural variation, e.g. marks, colours, patterns or sounds, eliminates most of the problems associated with artificial marking. Many bird species produce songs which can be recorded at a distance, with minimal impact on the individual. This provides the opportunity to use speaker identification techniques to identify the individual being recorded.

To date much work has been carried out in the area of individual recognition of birds from their songs, but this has focused on using the gross morphology and time varying characteristics of the song obtained from the spectrogram, such as the song or syllable length, maximum and minimum frequency, or change in frequency over time [3,4]. The classifiers used are similarly simple, including visual comparison of spectrograms, discriminant function analysis, and cross-correlation. These methods are often highly time intensive and subjective. A further problem is that each of these methods can only compare the same song type (i.e. it is text-dependent). However, in some bird species individuals produce a variety of songs which may not be shared amongst the entire population, while in other species individuals will regularly change their song types. These species therefore require a method of text-independent speaker identification.

Speaker identification in humans has received interest for use as a biometric to assist with secure access control [5]. Most speaker identification systems use short-time spectral analysis, and assume that speech is stationary over these periods. This short-term spectrum is then transformed into a set of feature vectors that represent the individual characteristics present in the speech signal. Speech analysis is based on the source-filter model, represented by

$$y[n] = s[n] * h[n]$$

where y[n] is the speech signal, s[n] is the excitation, and h[n] is the vocal tract filter. In humans the excitation signal is produced by the vocal folds, and this signal is then filtered by the vocal tract and articulators. In order to extract the individually characteristic features of the vocal tract filter, it is necessary to deconvolve s[n] and h[n]. The two main deconvolution methods are cepstral analysis and linear predictive coding.

The most commonly used features for human speaker identification are the mel-frequency cepstral coefficients (MFCCs) [5,6]. The MFCCs include information on the human auditory ability, and have also shown resilience to noise. They capture the vocal tract resonances, while excluding the excitation patterns.

While work on speech and species recognition has had some research in animals, only recently has the area of speaker identification in animals received interest. Recent studies have shown speaker identification accuracies of 82.5% in African elephants [7], and 76%-99% for the Norwegian Ortolan bunting [8]. However, these were all text-dependent tests.

This paper gives the first results for text-independent speaker identification in birds.

2. Approach

Speaker recognition follows the general method for any pattern recognition task, consisting of data collection, pre-processing, feature extraction and classification (Figure 1). Each of these steps is explained in greater detail below.

2.1 Data collection

Eight willie wagtails (*Rhipidura leucophrys*) were recorded between November 2004 and January 2005 at a variety of

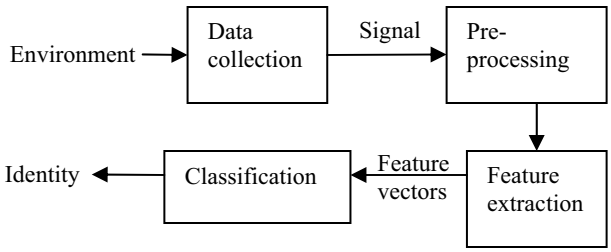

Figure 1 *General model for speaker recognition*.

locations around Perth, Western Australia. Birds were recorded at night (2000h to 0400h) during which time each bird would sit in a single location and sing.

The songs of eight noisy scrub-birds (*Atrichornis clamosus*) were recorded in December 2001 at Two People's Bay Nature Reserve (34˚59'22"S, 118˚11'4"E) on the south coast of Western Australia. Singing males were recorded between 0530h and 1830h.

The final data set was of eight singing honeyeaters (*Lichenostomus virescens*). Each bird was recorded before sunrise, between 0300h and 0500h, when they would sit and sing in a single location. Honeyeaters were recorded between November 2004 and January 2005 from street verges in the suburb of East Victoria Park, Western Australia.

Recordings of the scrub-birds were made using a Sony Walkman WMD6C with either a Sennheiser ME67 shotgun microphone or a Beyer Dynamic M88N(C) directional microphone. All other recordings were made using a Marantz PMD670 Solid State Recorder with a Sony ECM672 unidirectional microphone. The analogue recordings of the scrub-birds were digitized at 44.1kHz, while the other species were all recorded digitally at 48kHz.

2.2 Pre-processing

A recording from each individual had all periods of silence removed using the silence removal feature in Cool Edit Pro [9] plus some additional manual deletion, based on viewing the spectrogram and listening to the recording, to leave a signal of continuous bird song. The silent frames contain no speech information and discarding them improves computational efficiency. Each bird produced several different song types within a single recording. Some song types were specific to the individual, while others were shared between a few birds.

Since all recordings were made in the field they had background noise, particularly from wind, passing cars and other animals. To remove some of this noise a bandpass filter was applied to the signal to remove frequencies outside the range 1,000 Hz – 14,500 Hz for willie wagtails and noisy scrub-birds and 800 Hz – 14,500 Hz for the singing honeyeaters. The songs for all three species were within these ranges. Spectral subtraction using Goldwave's [10] Noise Subtraction function was also used, in which a sample of noise is analysed and this noise is then subtracted from the entire signal. Tests showed that this method of noise removal increased accuracy.

2.3 Feature extraction and classification

After noise removal, a 30 ms Hamming window was applied to the recording every 15 ms and the 12th order MFCCs were calculated for each frame. A window length of 30 ms is similar to that used in human speaker recognition, where windows are usually 10-30 ms in length. MFCCs are the most commonly used features for speaker recognition, having shown good results for both text-dependent and -independent recognition. They are based on the mel-frequency scale of human perception, and show a good ability for capturing the vocal tract resonances while excluding the excitation patterns. The first 12 MFCCs formed the feature vectors for the classifier.

Each recording was split into three sections. The first 10 seconds was used to train the classifier, the second 10 seconds was used for validation to improve generalisation and to prevent the classifier from overtraining, and the rest of the recording was used as the testing data. The data was tested in 2 second segments.

Text-independent recognition requires a classifier that is not temporally based. Of the classifiers commonly used for text-independent speaker identification, a back-propagation neural network, the multilayer perceptron (MLP), was chosen for this task. MLPs are able to classify input regions that either intersect each other or are disjoint as they are able to generalize from the information presented in the training data. MLPs have shown comparable results to another commonly used speaker recognition tool, vector quantization [11]. For further information on MLPs see [11]. The neural network toolbox in Matlab was used to design and implement the neural networks. The network had one hidden layer with 16 neurons, log sigmoid transfer functions and a Levenberg-Marquadt training function. Training continued until the error of the validation data started to increase.

3. Results

Speaker identification was carried out separately for the three species. In each species seven or eight of the eight individuals were correctly identified (i.e. had more than half the tests assigned to the correct bird), with an overall accuracy of 100% for willie wagtails, 68% for noisy scrub-birds, and 95% for singing honeyeaters. The confusion matrices are shown in Figure 2. For these tests the recordings were not split into their different song types, so the song types used for training and testing were a random assortment based on the order sung by the bird. Therefore, the song types present in the testing data may or may not have been present in the training data. In order to confirm that the technique is text-independent, further tests were carried out on the wagtail and scrub-bird recordings (seven wagtail and five scrub-bird recordings were able to be used).

The recording from each individual was separated into its different song types, with each song type assigned a letter. This was done via a visual inspection of the spectrograms. Each song type is highly stereotyped, even between individuals, making them simple to distinguish. Each willie wagtail had between two and four song types, with two being made frequently and any others only made

occasionally. Each noisy scrub-bird had between two and six song types sung in roughly equal proportions.

A network, one for each species, was trained with one song type from each bird and tested with a second song type. The same procedure as described above was used to extract the MFCCs and train the neural network. The network correctly identified all wagtails and four out of the five scrub-birds, with an overall accuracy of 96% and 71% respectively. The confusion matrices are shown in Figure 3.

4. Discussion and conclusions

This paper gives the first results for text-independent speaker identification in birds. The high results from the speaker identification tests (68-100%) are comparable to what is achieved in humans. They are also comparable to the results achieved for text-dependent identification in the Ortolan Bunting [8] which showed 85-95% accuracy for eight birds, depending on the song type, and in the African Elephant which showed an accuracy of 82.5% for six animals [7].

Text-independent recognition is typically more difficult than text-dependent recognition, so the high results achieved are particularly encouraging. There are many bird species in which individuals have a variety of song types, and in some species these song types can change over time. Therefore, a method of text-independent recognition is required for the application of this technique in the identification of individual birds in the field.

The lower result observed for the noisy scrub-birds is likely to be due to the higher amount of background noise present in these recordings. The willie wagtail and singing honeyeater recordings were made at night, or just before sunrise, when there is typically less wind and traffic and fewer birds and animals calling in the background. Therefore, they had much lower levels of background noise compared to the noisy scrub-birds which were recorded during the day.

Training and testing with different song types from each individual clearly showed that the MFCCs and the neural networks are capable of purely text-independent recognition. This was particularly highlighted in the results from the willie wagtails. In this test two song types (B and K) were used for both training and testing in different individuals (for example song type B was used for training in bird 5, and used for testing in bird 6). In both cases when these song types where tested they were successfully classified to the correct individual, rather than to the same song type.

The results given here do need to be treated with some caution since they are taken from a single recording for each bird. It is possible that recordings of the same bird taken at a different time may show lower accuracy due to the mismatched conditions between the recordings. In addition, only eight individuals were used and, as shown in [9], the accuracy can drop significantly as the number of individuals to be identified increases. However, the results are highly promising, particularly given that the methods used were those that have been developed for humans. Few alterations were made to either the features or the classifier to better suit the higher frequency and complex songs of the birds. The MFCCs are based on the human auditory ability which, while similar to that in birds, could be altered further to better suit the avian auditory ability. This will be the focus of future research.

The results given here show that text-independent speaker identification is possible is birds and, even using standard speaker recognition techniques, yields high accuracies. The next phase in this work will involve identifying an individual from recordings taken over time. This will be done by recording birds both in the laboratory (resulting in good quality recordings) and in the field (resulting in poorer quality recordings). From this the robustness of the technique can be determined, and hence its plausibility as a field tool.

5. Acknowledgements

Thanks to Allan Burbidge and Bill Rutherford for their help with banding willie wagtails and to Dean Portelli for supplying me with noisy scrub-bird recordings. Funding was supplied by the UWA School of Animal Biology, the Birds Australia Stuart Leslie Bird Research Award, and the Janice Klumpp Award.

6. References

[1] N. Burley, G. Kramtzberg, and P. Radman, "Influence of colour-banding on the conspecific preferences of zebra finches," *Animal Behaviour*, vol. 30, pp. 444-455, 1982.

[2] A. Berggren, and M. Low, "Leg problems and banding associated leg injuries in a closely monitored population of North Island robin (Petroica longipes)," *Wildlife Research*, vol. 31, pp. 535-541, 2004.

[3] T.M. Peake, P.K. McGregor, K.W. Smith, G. Tyler, G. Gilbert, and R.E. Green, "Individuality in corncrake Crex crex vocalizations," *Ibis*, vol. 140, pp. 120-127, 1998.

[4] D.N. Jones, and G.C. Smith, "Vocalisations of the marbled frogmouth: II. An assessment of vocal individuality as a potential census technique," *Emu*, vol. 97, pp. 296-304, 1997.

[5] J.P. Campbell, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, pp. 1437-1462, 1997.

[6] T.F. Quatieri, *Discrete-time speech signal processing: principles and practice*, Prentice Hall, New Jersey, 2001.

[7] P.J. Clemins, M.T. Johnson, K.M. Leong, and A. Savage, "Automatic classification and speaker identification of African elephant (Loxodonta Africana) vocalizations," *Journal of the Acoustical Society of America*, vol. 117, pp. 1-8, 2005.

[8] M.B. Trawicki, M.T. Johnson, and T.S. Osiejuk, "Automatic song-type classification and speaker identification of Norwegian Ortolan bunting," *IEEE International Conference on Machine Learning in Signal Processing*, 2005, in press.

[9] Syntrillium Software Corporation, Cool Edit Pro, v2.1, Phoenix, 2003.

[10] GoldWave Inc., GoldWave, v5.10, St. John's, 2005.

[11] R.P. Ramachandran, K.R. Farrell, R. Ramachandran, and R.J. Mammone, "Speaker recognition – general classifier approaches and data fusion methods," *Pattern Recognition*, vol. 35, pp.2801-2821, 2002.

A.

Classification \ Identity	1	2	3	4	5	6	7	8
1	12	0	0	0	0	0	0	0
2	0	39	0	0	0	0	0	0
3	0	0	53	0	0	0	0	0
4	0	0	0	24	0	0	0	0
5	0	0	0	0	20	0	0	0
6	0	0	0	0	0	24	0	0
7	0	0	0	0	0	0	16	0
8	0	0	0	0	0	0	0	26

B.

Classification \ Identity	159	325	4	40	41	42	43	9
159	16	6	2	0	7	0	4	0
325	0	11	0	0	2	2	0	0
4	1	5	9	0	0	4	0	0
40	0	0	4	6	0	9	2	0
41	0	1	0	0	8	2	0	0
42	0	7	0	0	8	22	3	0
43	1	1	0	1	0	2	53	9
9	0	0	0	1	0	0	0	54

C.

Classification \ Identity	2	6	10	12	14	15	16	21
2	14	1	0	0	0	1	2	1
6	0	48	2	0	0	0	0	0
10	0	0	100	5	0	0	0	2
12	0	0	0	31	0	0	0	0
14	0	1	0	0	60	0	0	0
15	0	0	0	0	0	27	0	0
16	1	0	0	0	0	0	92	0
21	1	1	0	2	0	0	2	63

Figure 2 *Speaker identification results for (A) willie wagtails, (B) noisy scrub-birds, and (C) singing honeyeaters.*

A.

Classification \ Identity	2 D	3 H	4 K	5 B	6 K	7 N	8 K
2 E	10	1	0	0	0	0	0
3 H2	0	22	0	0	0	0	0
4 L	0	0	5	0	1	0	0
5 C	0	0	0	11	0	0	0
6 B	0	0	0	0	12	0	0
7 K	0	0	1	0	0	5	0
8 P	0	0	0	0	0	0	10

B.

Classification \ Identity	159 A	4 G	42 M	43 M	9 I
159 B	3	10	2	0	0
4 H	0	22	2	3	0
42 N	0	0	8	0	0
43 N	1	1	5	12	0
9 Q	0	0	0	0	14

Figure 3 *Speaker identification when text-independent for (A) willie wagtails and (B) noisy scrub-birds.*