A New General Purpose, PC based, Sound Recognition System

Neil J Boucher (1), Michihiro Jinnai (2), Ian Gynther (3)

(1) Principal Engineer, Compustar, Brisbane, Australia
(2) Takamatsu National College of Technology, Japan
(3) Conservation Services, Queensland Parks and Wildlife Service, Brisbane, Australia

ABSRACT

We describe a method of sound recognition, using a novel mathematical approach, which allows precise recognition of a very wide range of different sounds. The mathematical approach is based on the use of the LPC transform to characterize the waveform and the Geometric Distance to compare the resultant pattern with a library of reference patterns of different sounds. The PC hardware consists of a dual processor P4 Pentium computer running at 3.0 GHz or faster. The system can use multiple sound cards and process ten or more sources recording at 44 kbps simultaneously. The initial application for this system is to monitor on a 24/7 basis, the calls of a rare parrot, reporting any detections in real-time by SMS and email. We show recognition capability that is orders of magnitude better than expert human listeners.

INTRODUCTION

A Coxen's Fig Parrot (illustration by Sally Elmer)

This project began with the development of software that is capable of recognizing the call of an endangered parrot called the Coxen's Parrot, found only in parts of Queensland and New South Wales. The bird is difficult to detect in its rainforest habitats using standard visual surveys and a novel approach to locating this species is required This project is being undertaken in cooperation with the EPA (Environmental Protection Agency) of the State of Queensland, Australia. Ian Gynther of the EPA is coordinating the project.

The objective is to use computer sound recognition software and the appropriate hardware to monitor sites occupied by the parrot on a 24-hour basis. The computer detects calls in real-time and alerts officials once a target call has been detected.

To be able to detect the calls, we needed some indication of how many different types of calls the bird could be expected to make and something of their nature. We were surprised to learn that this kind of information was not readily available (for any bird) and that the first job for the detector would be to examine recordings and identify and count the different types of calls.

Human observers had previously classified the calls of the Coxen's Parrot and other parrots into five types. However our software soon revealed at least twenty call types and maybe a lot more (it is difficult sometimes to decide if a variant of a call is indeed a variant or if it is a "new" call altogether). For our purposes if a call is significantly different from others and can be measured as such, then it is a "new" word. With that definition we soon had one hundred plus distinct calls to contend with.

As an example here is a call that you might hear if you were listening to a Coxen's Fig Parrot. Its wave form is seen in Figure 1. This call is a depiction of the recorded .WAV file.

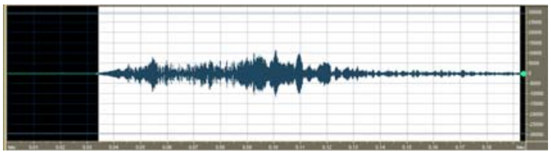
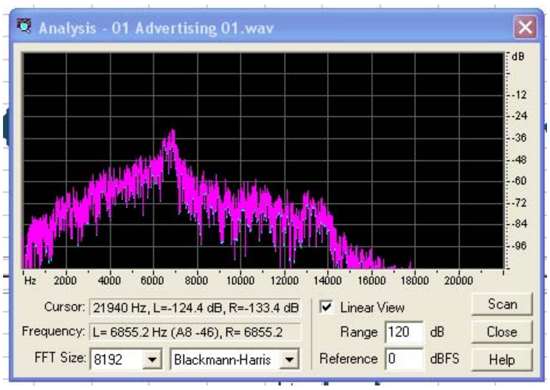


Figure 1. Time-domain form of the parrot call.

In the frequency domain it is even more interesting as seen in Figure 2. Most importantly the center frequency is around 7 kHz.



Cursor: 21940 Hz, L=-124.4 dB, R=-133.4 dB

☑ Linear View

Frequency: L= 6855.2 Hz (A8 -46), R= 6855.2

Range 120 dB

FFT Size: 8192 Blackmann-Harris

Reference 0 dBFS

Figure 2. Frequency domain analysis of the parrot call.

Now to understand what this means for recognition let's look at a similar recording of one of us (Boucher) saying "bird call" in the time domain as is shown in Figure 3. The first thing to notice is that the complexity of this phrase is less than that of the call in Figure 1. Hence it is possible (probable) that the bird call contains more information than a few words.

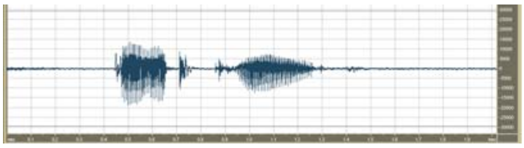
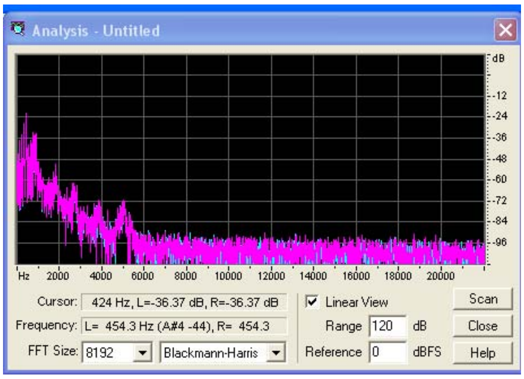


Figure 3. Waveform of the human phrase "bird call" seen in the time domain.



Cursor: 424 Hz, L=-36.37 dB, R=-36.37 dB

☑ Linear View

Frequency: L= 454.3 Hz (A#4 -44), R= 454.3

Range 120 dB

FFT Size: 8192 Blackmann-Harris

Reference 0 dBFS

Figure 4 The phrase "bird call" in the frequency domain.

Now notice where the voiced sound is centered (at around 300 Hz). Most of the parrot call is outside the hearing range of humans (admittedly some young people can hear to 20 kHz, but they can't get much useful information from any signal above about 5 kHz). The parrot on the other hand is mostly "talking" at a frequency of 7.5 kHz.

This information was both surprising and a little alarming. We are dealing with a sound that is structurally more complex than human voice, and we have to reliably detect it. A search of other attempts to identify bird calls soon revealed that there was very limited success in this field. In particular AI detectors, which worked well for simple calls like frogs and crickets, were very poor at discriminating bird calls.

After a few optimistic false starts, using traditional recognition techniques, the call complexity led us to the conclusion that a wholly new approach was called for. One technique that was firmly ruled out was speech recognition software. This software. Which has matured remarkably over the last few years, has become excellent for detection of voiced sounds, but in the process has become so specialized that it cannot handle non-voiced sounds.

TRADITIONAL RECOGNITION

Traditional recognition techniques rely on taking the recordings as .WAV files and performing an FFT (Fast Fourier Transform) on the signal. The FFT removes a lot of the redundancy in the call and so makes the recognition task easier. This same process occurs in human sound recognition as the hair cells within the cochlea respond to different frequencies and present outputs which are resolved in the frequency domain, in a way that mimics the FFT.

In traditional computer recognition, once we have the signal resolved into its frequency components, the FFT image of the sound is compared with a library of "known" sound FFT images. The computer recognition relies on comparing an FFT image (as seen in Figures 2 and 4 above) with those referenced in the library. So what we are really comparing is the relative shape of the transform The usual way to do this is to use a measure of similarity called the Euclidean Distance.

The Euclidean Distance is simply the RMS distance between the two patterns defined as in equation 1.

$$D = \sqrt{1/n\sum_{1}^{n}\left(A_i - B_i\right)^2} \quad ..1$$

And this distance $A_i - B_i$ is the distance as indicated in Figure 5, as the physical distance between the two waveforms.

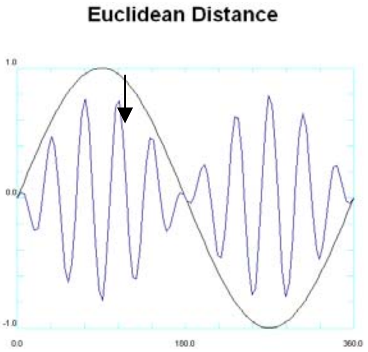


Figure 5. The Euclidean Distance between two graphs at a point is the actual distance between them.

This method of comparing shapes is widely used in image processing and voice recognition. Its downside is that the Euclidean Distance is rather sensitive to noise, and while it is very good at recognizing identical patterns it is not so good at recognizing similar ones. This posed a problem as early work had already indicated that, when examined carefully, a call from a single bird that repeats in a pattern is such that each successive call "burst" can be very different from the previous ones. In fact no two distinct call "bursts" have been found so far that are perfect matches to each other, either from the same bird or from within groups of similar calls from all the birds studied.

To make matters worse the only recordings we can get for the Coxen's Parrot of those of a related northern species, which is believed to be very similar (based on human listener reports) but we have no metric on the degree of similarity.

Exactly how the human brain processes sound is still the subject of much debate. But the human brain is reasonably good at recognizing similarity and rather poor at confirming perfect matches. Therefore it is unlikely that the human processing is even similar to the Euclidean method.

NEW METHOD

A new method of comparing patterns has been pioneered by one of us (Jinnai) and it uses the LPC (linear predictive coefficients) and Geometric Distance rather than the FFT and Euclidean Distance.

The LPC is widely used in digital speech encoders as Linear Predictive Coding. It is an efficient way of minimizing the number of bits that need to be encoded. The LPC in speech encoders also takes advantage of the speech vocal tract characteristics as a filter (which is something that is not used in our model). The LPC has a greater computational overhead than the FFT but it is more efficient at minimizing redundancies. This leads to a compromise situation where the LPC is calculated to a lower order than the FFT.

In Figure 6 we see an example of a .WAV file transformed to both an FFT and LPC. It can be seen that the LPC is a "cleaner" and simpler transform, which makes the pattern matching easier.

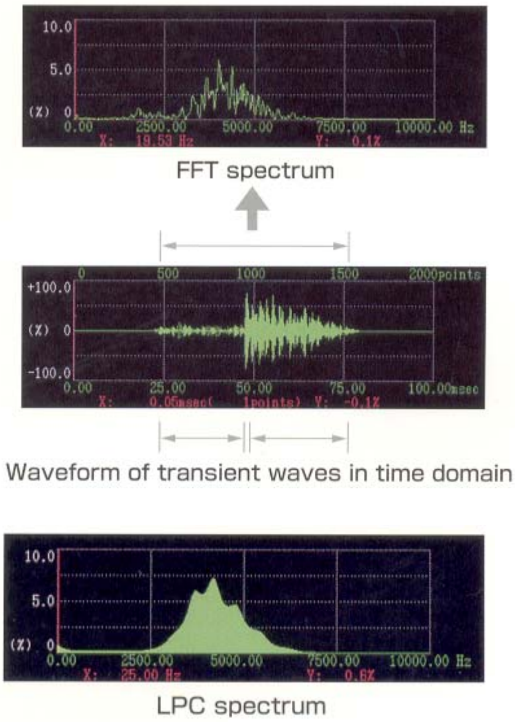


Figure 6. The FFT compared to the LPC

Once we have the transform, next we compare the transformed spectrum with the reference library of transformed spectra.

Our next depature from the converntional matching is that instead of Euclidean Distance we use the Geometric Distance. The Geometric Distance is again more computationally challenging than the Euclidean Distance, but our experience reveals that the Geometric Distance better classifies similar images than does the Euclidean Distance. The latter tends to classify patterns that to a human observer appear similar, as dissimilar. The Geometric Distance so calculated will always be in the range of -1 to +1. The geometric distance also performs better when it has to contend with noise and distortion.

The matching is further refined by using a weighting vector, which effectively assigns more weight to the most energetic part of the waveform, thus discounting the less prominent parts of the signal. This technique can also be used with the conventional matching techniques with good results.

The Geometric detection process is patented by one of us (Jinnai), and the code is made available as source code in VB6, .NET and as a .DLL

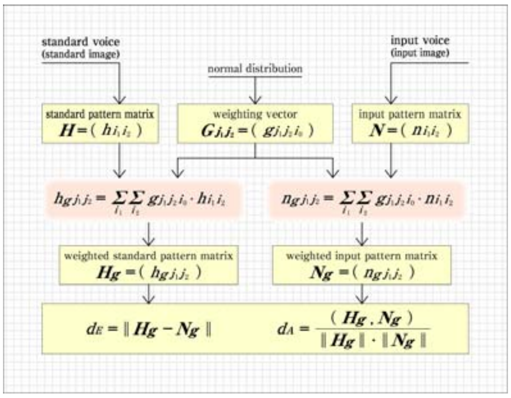

standard voice
(standard image)
normal distribution
input voice
(input image)
standard pattern matrix
$H = (h_{i_1 i_2})$
weighting vector
$G_{j_1 j_2} = (g_{j_1 j_2 i_0})$
input pattern matrix
$N = (n_{i_1 i_2})$
$hg_{j_1 j_2} = \sum_{i_1} \sum_{i_2} g_{j_1 j_2 i_0} \cdot h_{i_1 i_2}$
$ng_{j_1 j_2} = \sum_{i_1} \sum_{i_2} g_{j_1 j_2 i_0} \cdot n_{i_1 i_2}$
weighted standard pattern matrix
$Hg = (hg_{j_1 j_2})$
weighted input pattern matrix
$Ng = (ng_{j_1 j_2})$
$d_E = \| Hg - Ng \|$
$d_A = \frac{(Hg , Ng)}{\| Hg \| \cdot \| Ng \|}$

Figure 7 Process of the geometric transform.

THE HARDWARE

The heavily mathematical processes involved in this detection means that only top-end PCs are useful and we consider, as a minimum, a 3.0 GHz P4 is required. In the .NET versions of the hardware true multi-threading is available and multi-CPU processors can be used to full advantage as is 64 bit calculation.

It is possible to use the PC sound card as the input source. Not all sound cards are created equal, and the most important parameter is the signal to noise ratio which should be about 100 dB. Some sound cards are as noisy as 80 dB and should be avoided. Multiple sound cards can be used and most PCs have three to five PCI slots (the slots that take the sound cards). The sounds cards are mostly stereo and the two stereo channels can be used as independent channels, although in some applications the cross-talk between channels can be an issue. Cross-talk is typically 50-60 dB. Where this level of cross-talk is a problem it would be best to use only one channel per card

If more channels are needed than can be provided by multiple sound cards, outboard solutions are possible. The audio industry today uses PCs and sound cards for sound mixing. Solutions that allow 40 or more high quality sound channels are available at reasonable prices (about $40.00 per channel).

For rainforest applications where it is intended to detect parrots visiting their food trees (fruiting figs) we have developed waterproof radio microphones and acoustic parabolic dish receivers. The wireless microphones provide up to two months operation on rechargeable batteries or indefinite operation from small solar panels.

The microphones turned out to be more problematic that had been expected. As a economy measure it is desirable to maximize the acoustic range. The simplest way to do this is to add a preamplifier. This increases the range but it also increases the noise floor (mainly acoustic noise from the microphone). As the noise floor rises the headroom (the dynamic range between the noise floor and saturation of the sound card A/D decreases. We found a 40 dB preamplifier that left about 50 dB of headroom to be the best compromise. The resultant acoustic range is similar to that of the unaided ear.

Results

As already indicated the parrot calls are most demanding diction challenges and they were ideal for testing how well the system performs with complex sounds. One benchmark was to equal an expert human listener. Surprisingly there are not much data on how accurate human observers are when identifying bird calls. However experience with random English words suggests that about 95% accuracy is what can be expected from an average human listener (random words of course do not have the context clues that sentences provide).

We were provided with .WAV files of our target bird species and of species that were likely to be in the same location, and ones that were sometimes confused with the target species by human observers. Initially we looked at the files of calls that humans sometimes mismatch and we found that the software was not similarly confused by these calls. A hint of why this is so comes from comparing the bird call in Figure 1 with the voiced sound in Figure 3. The bird call lasts about 0.1 seconds and the voiced sound is about 1.0 seconds. It apparently is a fact that birds process calls at ten times the rate that humans do (notice also that the centre frequency and bandwidth of the bird call is about ten times higher than that of a human). Humans therefore fail to hear most of the detail in the bird call. In fact by slowing down a bird call a simple “chirp” is clearly heard to be much more complex, and is not “chirp-like” at all.

Next we needed to test how well the software classifies calls and how immune it is to mismatching calls. Using the software we found and identified groups of similar calls. Then we embedded these calls in a mix of all the calls that we had recorded (about 630 in all). Initially the result was that a call was wrongly identified only once or twice in each run (an error rate of about 0.3%, or a correct identification rate of 99.7%). This encouraging result led to some more development in the software that resulted in a false positive of zero and 100% correct identification of like calls.

This does not mean that the system has been totally perfected as the real world is likely to throw some more challenges, but within the confines of the original set of test calls provided the system has matured beyond expectations.

By tightening the value of the Geometric Distance that defines a group, the false positives can be reduced to zero, but at the cost of missing some positives. This in turn can be overcome by further subdividing the reference calls into more groups, but there is a limit to how far this process can be taken.

There will always be some compromise here as indeed there would be with a human observer, who hears a call that is like the target, but cannot be certain that match is perfect. If the human observer lowers the standard for matching there is a consequent increase chance of false positives.

CONCLUSION

As a result of our efforts to detect the most difficult of sounds (parrot calls) we now have a very useful general purpose, PC based sound detector. The system is robust, very capable and most importantly cheap enough for widespread use. The technique is in no way limited to wildlife and could be used for any sound detection purpose. It is especially good at detecting similarities, and in this respect is superior to conventional techniques. It is our intention to make this software available to researchers who may have an application for this technology.