:: **Refereed Article A2:**

An investigation and comparison of speech recognition software for determining if bird song recordings contain legible human voices

Tim D. Hunt
The Waikato Institute of Technology (Wintec), New Zealand
tim.hunt@wintec.ac.nz

Grant Ryan
The Cacophony Project, New Zealand
grant@hunchcruncher.com

Cameron Ryan-Pears
The Cacophony Project, New Zealand
cameron.ryan.pears@gmail.com

Abstract

The purpose of this work was to test the effectiveness of using readily available speech recognition API services to determine if recordings of bird song had inadvertently recorded human voices. A mobile phone was used to record a human speaking at increasing distances from the phone in an outside setting with bird song occurring in the background. One of the services was trained with sample recordings and each service was compared for their ability to return recognized words. The services from Google and IBM performed similarly and the Microsoft service, that allowed training, performed slightly better. However, all three services failed to perform at a level that would enable recordings with recognizable human speech to be deleted in order to maintain full privacy protection.

Keywords

speech recognition api, the cacophony project, pest control and eradication, audio recording, privacy

1. Introduction

Introduced pests have caused, either directly or indirectly through habitat destruction, the reduction and extinction of many of New Zealand's species of native birds (Brown, Elliott, Innes, & Kemp, 2015). The Cacophony Project, (Moore's Law for New Zealand Birds, 2016) aims to use technology to greatly increase the effectiveness of pest control and eradication in New Zealand and the work described in this paper contributes to that effort. In particular, this paper describes the use of low specification (old) mobile phones to record bird song in a way that avoids the problem of storing recordings that have inadvertently captured human conversations.

1.1 The Cacophony Project

The Cacophony Project (Moore's Law for New Zealand Birds, 2016) is a New Zealand

based open source project that has the aim of dramatically increasing the number of native birds in New Zealand's environment. The botanist Joseph Banks who arrived in New Zealand in 1769 with James Cook recorded the following in his journal:

"This morn I was awakd by the singing of the birds ashore from whence we are distant not a quarter of a mile, the numbers of them were certainly very great who seemd to strain their throats with emulation perhaps; their voices were certainly the most melodious wild musick I have ever heard, almost imitating small bells but with the most tuneable silver sound imaginable to which maybe the distance was no small addition. On enquiring of our people I was told that they have had observd them ever since we have been here, and that they begin to sing at about 1 or 2 in the morn and continue till sunrise, after which they are silent all day like our nightingales." (Banks, 2006).

The Cacophony Project aims to use technology to improve the effectiveness of removing pests that kill birds and eat the vegetation that they are dependent on, and in doing so, help to restore the dawn chorus of bird song to many parts of New Zealand. Technology will be used to monitor changing numbers of both pests and birds in order to assess the effectiveness of various interventions.

1.2 Pest control and eradication

If a pest eradication program is to be ultimately successful, authors such as Miller have stated "It is important that the remaining animals be quickly eradicated because they, surrounded by increased available resources, could repopulate the island in a relatively short time" (Miller, 1993, p. 5). Further, Brown et al (2015) noted "The biggest current issue with stoat trapping is that not all stoats enter the traps, especially when alternative food is plentiful" (p9) and go on to say "The development of a lure that will entice all stoats to enter traps is the 'Holy Grail' of stoat trapping" (p. 9). Parkes (1993) observed the exponential increase in cost per goat killed as the pest density neared zero, highlighting the difficulty in eradicating pests as opposed to merely continually managing the problem.

Brown et al. (2015) gave a review of the "techniques, successes and challenges" (p1) of controlling pests in New Zealand and concluded with a list of eleven "Where to from here?" (p25) issues and recommendations. Of particular interest to this work is the observation that "There will always be a need for ground-based control..." (p25) leading to the recommendation to "Continue development of new long-life lures, traps and toxins that will continue to incrementally improve the effectiveness of ground-based pest control" (p26).

In July 2016, the government announced a goal for "New Zealand to be Predator Free by 2050" (Department of Conservation, 2016) which has the ambitious aim of removing possums, rats and stoats from all of New Zealand. It is worth noting that ridding the environment other pests, such as feral cats and ferrets that are a threat to New Zealand birds, especially when other pests is not mentioned, but possibly implied.

1.3 Educational aspects

One of the authors of this work teaches Information Technology at a New Zealand Institute of Technology and was able to incorporate the research as an example case study in the class room. Details of this are given in section 5.4.

2. Audio recordings to establish a base line

One aim of The Cacophony Project is to create an inexpensive device allowing anyone to determine "The Cacophony Index" in their own local area and to ascertain whether the number and type of birds is increasing or decreasing, and at what rate. Only these objective measures for different predator control methods will allow effectiveness to be compared.

This work has looked at using low specification (old or unwanted) mobile phones to automatically make periodic audio recordings and upload them to a remote server for later analysis. Ultimately it is envisioned that phones will be setup in remote locations utilizing solar power and mobile data connections. However, initial recordings will be made in locations near to power sources and possibly Wi-Fi access, as would probably be the case in many urban or rural locations near a dwelling where monitoring will provide insightful information.

As noted by Brown et al. (2015) the mast seeding in beech and podocarp forests also greatly affects the population of pests. Calculating a base line will need to take these, and possibly other factors, into account when determining the effectiveness of an intervention.

2.1 Privacy issues and speech detection

Initially recordings were only made in remote locations where the possibility of inadvertently recording private conversations was deemed to be low. Recordings were kept confidentially by only one or two researchers. However, the aim of the project is to have thousands of devices, some in urban areas, and to make all the recordings "open source' to enable anyone to download and analyze them. There is therefore a need to find an automated solution that deletes any recordings containing human speech.

2.2 Speech recognition software

In this work, we investigated the feasibility of using "off-the-shelf" speech recognition services to determine the presence of spoken words in a recording. Although, for this work, it is not necessary to know what words are present in a recording, the existence of robust and readily available speech recognition services offered the possibility of integrating such a service into the recording process allowing any recordings containing recognizable words to be deleted. This work is not a complete review of available solutions, but rather has considered the robust services from some of the major companies currently making their algorithms available through easily accessible cloud based interfaces.

2.3 Audio analysis

Sound is basically a change in air pressure over time and a microphone is a transducer that measures these pressure changes and converts them into an electrical signal. These values are recorded at regular intervals (the sampling rate/frequency) and encoded using a variety of techniques resulting in a wide variety of common digital formats (Zölzer, 2008). The rate of change (frequency) of air pressure can be calculated (a technique known as Fourier analysis (Howell, 2017)) and is a common technique for analyzing audio recordings where a plot of frequency intensity versus time is used. This technique (in other work) is also being evaluated as a method for determining if a recording contains human voices to facilitate their removal.

2.4 Aim of this work

The aim of this work was to automatically identify sound recordings that contained human speech that could be understood by someone listening to the recording. This would then enable the deletion of these recordings before they are made available to a wider audience. Before this work started, the authors did not know how far from a phone a human could still be understood when the recording was listened to by a human and certainly did not know if speech recognition software would perform better or worse and so for this reason only, recordings were made at various distances from the phone. It should be noted that the exact distance from the phone that a recording was made was not important, but rather what was important was whether a human listening to the recording could understand what was being said when the automatic system could not.

3. Experimental method

Recordings were made using a HTC One V phone using the default "Voice Recorder" app that came with the phone. The Adaptive Multi-Rate (AMR) (3gpp, 2016) encoding option was used (the other option was AAC_LC, Advanced Audio Coding), no other options were available. The sample rate was determined (by using the QuickTime player (Apple Inc, 2016) to display the file properties) to be 8000Hz.

A series of recordings were made in real conditions with the phone placed on a wooden platform in a domestic garden with birds clearly audible to the human ear in the background. The first recording was made standing next to the phone, and then each subsequent recording was an additional step (approximately 90 cm) further from the phone, up to 15 steps for the furthest recording. Thus the volume/quality of the spoken phrases as captured by the phone decreased with distance from the phone.

Some of the recordings also contained other background noises such as vehicles and planes.

The Cacophony project intends to capture bird singing in a variety of settings from urban situations similar to that used in this work, through to wilderness locations many kilometers from buildings and surrounded by much denser tree cover.

3.1 Technology investigated

Three speech to text services were investigated: Cloud Speech API from Google (Google, 2016), Watson Speech to Text from IBM (IBM, 2016) and Custom Recognition Intelligent Service (CRIS) from Microsoft (Microsoft, 2015) which was in "private preview' at the time of use.

Google Cloud Speech API

The Google Speech API server does not allow for any configuration and recommends to not perform any processing of the files before uploading them. It does suggest using a sampling rate of 16000 Hz or higher and recommends using FLAC (Xiph.Org Foundation, 2016) or an uncompressed lossless codec. Although AMR uses compression, the raw AMR files from the phone were accepted by the API service.

IBM Watson Speech to Text

IBM Watson required files in the FLAC format and the VLC Media player (VideoLan Organization, 2016) was used to convert the AMR files to FLAC format.

Microsoft Custom Recognition Intelligent Service (CRIS)

CRIS requires the recordings to be in the Waveform Audio File (WAV) format (Wikipedia, 2016) sampled at 8000Hz, Pulse-Code Modulation (PCM) using sixteen bit integers. The VLC Media player was used to convert the files to WAV format, with the sample rate set to 8000 Hz, a bit rate 128 kb/s and one channel. There was no option to set a PCM setting (which didn't seem to prevent the recordings being accepted by CRIS).

CRIS allows for the creation of a custom acoustic model for a particular domain: "For example, if you would like to better recognize speech in a noisy factory environment, the audio files should consist of people speaking in a noisy factory" (Microsoft, 2015). The creation of the model is achieved by uploading sound files along with a transcription of the audio. CRIS also allowed for the creation of custom language models, but this feature was not investigated in this study and the default US English model was used.

4. Results

4.1 Google Cloud Speech API and IBM Watson Speech to Text

Initially sixteen recordings were made with the following sentence repeated twice "This is test X standing Y steps from the phone" with X and Y increasing by 1 for each additional step from the phone. There will have been some variability in the volume of the speech and the exact distance from the phone of each recording, however listening to the recordings confirmed the general trend of reduced quality of recording as the speaker moved further away.

Although even when standing next to the phone neither of the services gave perfect speech recognition results, both gave responses that indicated that words had been detected. For example, the result returned from the Google Cloud Speech API contained the phrase "this is test free" with a confidence level of 0.93 and alternatives of "standing Two Steps From the Sun" with confidence level of 0.86 and "this is free-standing Two Steps From the Sun" with confidence level of 0.82. Similar results were obtained up to six steps (apart from no recognized words at five steps with the Google Cloud Speech API) with the confidence reducing to 0.60. Neither services returned detected words at more than six steps, yet a human can still discern words in the recordings up to fifteen steps from the phone.

4.2 Microsoft CRIS

The first test created a model using the first two recordings (zero and one step), and

then the model was tested with recordings taken from further steps away (Acoustic Model 1). As could be expected, the results did show an improvement over the other services with the result at two steps being "this is test free standing two steps from the phone this is test free standing two steps from the phone". However, it also failed to return any recognized words in recordings made more than six steps from the phone.

Further training was then performed using recordings at zero to six steps (Acoustic Model 2) but this only slightly improved the performance with some words being recognized at up to seven steps. More training was performed using the recordings of up to ten steps from the phone (Acoustic Model 3), and although the same recordings were then used to try to recognize words, the service failed to recognize words in the recordings at nine and ten steps, although some words "this is tom eleven" were returned for the eleven steps recording but none at twelve steps.

Further training

The training used so far consisted of very short phases consisting of a limited range of words. A new set of training data (seventeen separate recordings) was created at fifteen steps from the phone and a new acoustic model was created (Acoustic Model 4). Table 1 gives the transcriptions of some of the training data used to create this model.

Table 1. Transcriptions of examples of training recordings used to create CRIS Acoustic Model 4

Although rabbits are the main diet of ferrets, this mustelid will attack, kill and eat native species. Ferrets are small predators and are members of the mustelid family, along with stoats, weasels, badgers, mink and otters.
Stoats, weasels and ferrets, the only mustelids present in New Zealand, pose a serious threat to our threatened wildlife. In addition to New Zealand's wild populations, ferrets are kept as domestic pets and farmed for export markets in Asia.
Ferrets were introduced to New Zealand from Europe in the 1880s, along with stoats and weasels, to control rabbits that were breeding out of control. By 1900, ferrets were well established in the wild and definitely played a role in the decline of native birds like the kiwi, weka and blue duck, and the extinction of kakapo on the mainland.
When these closed down due to a downturn in the market, many ferrets escaped or were set free, contributing to the expansion of ferrets northwards into some of New Zealand's remaining prime kiwi habitat.

Once the model had been trained, it was tested with the phrases at between five and 15 steps and the results are given in Table 2.

Table 2. Transcriptions of test phrases and results returned from the CRIS service when using Acoustic Model 4.

Distance (Steps from phone)	Transcription of test phrase.	Result returned from service
Five	This is test six standing five steps from the phone This is test six standing five steps from the phone	test six five cents from son has sex and then five cents from the start
Six	This looks like a good place to stop for lunch can you hear the birds	this looks like a good place for lunch
Seven	This is test eight standing seven steps from the phone This is test eight standing seven steps from the phone	Unspecified Error
Eight	This is test nine standing eight steps from the phone This is test nine standing eight steps from the phone	Unspecified Error
Fifteen	Not all ferret breeders were registered under the regulations. Ferrets are now unwanted organisms under the Biosecurity Act. This has stopped the sale and reduced the public acceptance of ferrets as pets. You can help Signs that ferrets are present Strong eye-shine at night.	Unspecified Error

These results were still not good enough for a working solution and so a new model was created, Acoustic Model 5, that used recordings of the same phrases previously used, but this time at 8 steps and the results are given in Table 3.

Table 3. Results of testing Acoustic Model 5 that was created with eleven separate recordings similar to those given in Table 1. It can be seen that some words have been returned at eight steps, but no words are recognized at further distances.

Distance (Steps from phone)	Transcription of test phrase.	Result returned from service
Five	This is test six standing five steps from the phone This is test six standing five steps from the phone	test six five cents from son has sex and then five cents from the start
Six	This looks like a good place to stop for lunch can you hear the birds	this looks like a good place for lunch
Seven	This is test eight standing seven steps from the phone This is test eight standing seven steps from the phone	Unspecified Error
Eight	This is test nine standing eight steps from the phone This is test nine standing eight steps from the phone	Unspecified Error
Fifteen	Not all ferret breeders were registered under the regulations. Ferrets are now unwanted organisms under the Biosecurity Act. This has stopped the sale and reduced the public acceptance of ferrets as pets. You can help Signs that ferrets are present Strong eye-shine at night.	Unspecified Error

Acoustic model 5, created using training recordings at eight steps has given slightly better results than the model created at fifteen steps, but the results show that the speech recognition service still does not return words at greater than eight steps.

5. Conclusion and discussion

The speech recognition services from Google and IBM gave similar results with words recognized when the speaker was up to six steps from the phone. CRIS, the service from Microsoft gave similar results when the acoustic model was created with recordings of the speaker at zero and one step from the phone and this result was improved on slightly with words recognized at eight steps when the acoustic model was created with recordings of the speaker at eight steps. Audio recordings with the speaker between nine and fifteen steps from the phone, contained words that could be recognized by a human but not by any of the services.

5.1 Algorithm improvement

The algorithms tested in this work are "black-boxes" from the end-user's perspective. It is likely that improvements to the algorithm's will continue and might allow them to be successfully used at some point in the future. For example, recent work (Yin, et al., 2015) on speech recognition has considered the possibility of over-fitting models when training deep neural networks in noisy conditions. They report improved performance by intentionally adding noise during the training process.

5.2 Student involvement

One of the authors teaches an Artificial Intelligence (A.I.) course on the third year of a Bachelor of Information Technology degree. The course introduces the students to several concepts of A.I. including: rule based/fuzzy expert systems, artificial neural networks, genetic algorithms and knowledge engineering. A major assessment is the creation of a weekly food menu using genetic algorithm methods to minimize cost while adhering to some constraints such as not repeating meals or avoiding certain food types.

The author undertakes research to inform his teaching and could see that his current research might interest the students. He presented the scenario of the research as a case study and the current impasse of the best speech recognition software currently available being unable to detect speech in recordings in situations where a human could. Although he was not expecting any great incites from the students, he was hoping to generate discussion, encourage thought and elicit feedback.

One of the students observed that the results from the Google service all had confidence levels above 0.59 and suggested that the API be configured to return results of lower confidence. This observation was a good example of looking at a problem with a new set of eyes and the student was excited to able to contribute to the lecturer's research.

The author followed up with Google as they do not offer this configuration, but have accepted a feature request and have passed the request on to their engineering team for consideration. The student and class were informed of the positive response to the idea from the student.

It is possible that the comparison of speech recognition algorithms could be set as a practical and hands on case study for future students. This would be especially appealing to students if Google go ahead with the implementation of the student's suggestion.

5.3 Creation of a custom algorithm

The need to recognize words is not required for the purpose of determining whether human voices are present in recordings of bird song. Instead an algorithm that just returns yes or no (human voice detected) would be sufficient and, thus, creating this algorithm may be a more appropriate method for solving the problem.

5.4 Implications for sleep apnea research

Most speech recognition research and applications consider the ability to determine actual words spoken. This work was different in that it only needed to determine if people were talking from a privacy protection view point i.e. it did not need to know what words were detected. A similar situation occurs in the research into sleep apnea where people are often recorded in the privacy of their bedrooms: for example, a study (Schwab, et al., 2013) regarding "The Optimal Monitoring Strategies and Outcome Measures in Adults' discuss data security and privacy and safety issues of monitoring. In sleep apnea work it is important to record snoring but it would be very

useful to prevent private conversations being captured. It has been noted that as both snoring and speaking occupy similar frequencies it can be hard to separate them using conventional signal processing techniques (Hicks, 2012; Pevernagie, Aarts, & De Meyer, 2010). Behar et al (2013) have developed a smart phone application for at home screening of sleep apnea patients. Applying the concept of repurposing the speech recognition systems investigated in this paper is likely to improve the privacy of those people being monitored for sleep apnea.

5.5 Implications for electronic waste reduction

There is a growing concern regarding the increase of electronic waste ending up in land fill (Awasthi, Zeng, & Li, 2016; Razi, 2016; Tan, et al., 2017). Mishima, Rosano, Mishima, & Nishimura (2016) discuss the strategies for reducing electronic waste and conclude with the suggestion of investigating the value of increasing the reuse rate of mobile phones. The Cacophony Project fits well with this idea.

6. Summary

The concept of using an "off-the-shelf" speech recognition service for determining if recordings contain human voices has been tested. After testing three services from well known organizations, it has been found that the system worked for recordings made close to the phone, but as the speaker moved further than around eight steps from the phone, none of the services returned any recognized words even after training custom acoustic models.

The Cacophony Project will need to look at other techniques for ensuring the privacy of those who may be inadvertently recorded by phones located for the purpose of recording bird song.

Acknowledgements

References

3gpp. (2016). *Mandatory speech CODEC speech processing functions; AMR speech Codec; General description*. Retrieved from https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=1386

Apple Inc. (2016). *QuickTime Player Support*. Retrieved from Support: https://support.apple.com/quicktime

Awasthi, A. K., Zeng, X., & Li, J. (2016). Comparative Examining and Analysis of E-waste Recycling in. *Procedia Environmental Sciences*, 676-680.

Banks, J. (2006). *The Endeavour Journal of Sir Joseph Banks.* Cirencester: Echo Library. Retrieved from Project Gutenberg Australia.

Behar, J., Roebuck, A., Shahid, M., Daly, J., Hallack, A., Palmius, N., . . . Clifford, G. D. (2013). SleepAp: An Automated Obstructive Sleep Apnoea Screening Application for Smartphones. *Computing in Cardiology*, 257-260.

Brown, K., Elliott, G., Innes, J., & Kemp, J. (2015). *Ship rat, stoat and possum control on mainland New Zealand: an overview of techniques, successes and challenges.* Nelson: Department of Conservation. Retrieved from http://www.doc.govt.nz/Documents/conservation/threats-and-impacts/animal-pests/ship-rat-stoat-possum-control.pdf

Department of Conservation. (2016). *Predator Free New Zealand 2050*. Retrieved from Department of Conservation Te Papa Atawhai: http://www.doc.govt.nz/our-work/predator-free-new-zealand-2050/

Google. (2016). *Cloud speech API*. Retrieved from Google Cloud Platform: https://cloud.google.com/speech/

Hicks, D. R. (2012, March 24). *How to distinguish voice from snoring?* Retrieved from Signal Processing, Stack Exchange: http://dsp.stackexchange.com/questions/1828/how-to-distinguish-voice-from-snoring

Howell, K. B. (2017). *Principles of Fourier Analysis.* Boca Raton: CRC Press.

IBM. (2016). *Speech to Text*. Retrieved from Watson Developer Cloud: https://www.ibm.com/watson/developercloud/speech-to-text.html

Microsoft. (2015). *Custom Recognition Intelligent Service (CRIS)*. Retrieved from

CRIS: https://cris.ai/

Miller, C. J. (1993). An evaluation of two possum trap types for catch-efficiency. *Journal of the Royal Society of New Zealand*, 5-11. doi:10.1080/03036758.1993.10721213

Mishima, K., Rosano, M., Mishima, N., & Nishimura, H. (2016). End-of-Life Strategies for Used Mobile Phones Using Material Flow Modeling. *Recycling*, 122-135.

Moore's Law for New Zealand Birds. (2016). Retrieved from The Cacophony Project: https://cacophony.org.nz/

Parkes, J. P. (1993). The ecological dynamics of pest-resource-people. *New Zealand Journal of Zoology*, 223-230. doi:10.1080/03014223.1993.10420333

Pevernagie, D., Aarts, R. M., & De Meyer, M. (2010). The acoustics of snoring. *Sleep Medicine Reviews*, 131-144.

Razi, K. M. (2016). Resourceful recycling process of waste desktop computers: A review study. *Resources, Conservation and Recycling*, 30-47.

Schwab, R. J., Badr, S. M., Epstein, L. J., Gay, P. C., Gozal, D., Kohler, M., . . . Weaver, T. E. (2013). *An Official American Thoracic Society Statement: Continuous Positive Airway Pressure Adherence Tracking Systems The Optimal Monitoring Strategies and Outcome Measures in Adults.* American Thoracic Society. doi:10.1164/rccm.201307-1282ST

Tan, Q., Dong, Q., Liu, L., Song, Q., Liang, Y., & Li, J. (2017). Potential recycling availability and capacity assessment on typical metals in waste mobile phones: A current research study in China. *Journal of Cleaner Production*, 509-517.

VideoLan Organization. (2016). *VLC media player*. Retrieved from http://www.videolan.org/index.html

Wikipedia. (2016). *WAV*. Retrieved from Wikipedia The Free Encyclopedia: https://en.wikipedia.org/wiki/WAV

Xiph.Org Foundation. (2016). *what is FLAC?* Retrieved from flac free lossless audio codec: https://xiph.org/flac/

Yin, S., Liu, C., Zhang, Z., Lin, Y., Wang, D., Tejedor, J., . . . Li, Y. (2015). Noisy training for deep neural networks in speech recognition. *EURASIP Journal on Audio, Speech, andMusic Processing*, 1-14. doi:10.1186/s13636-014-0047-0

Zölzer, U. (2008). *Digital Audio Signal Processing.* Chippenham: John Wiley & Sons Ltd.