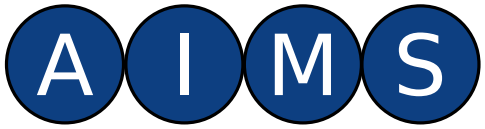


Autonomous Intelligent
Machines & Systems

UNIVERSITY OF OXFORD

DEPARTMENT OF ENGINEERING SCIENCE

Transfer of Status

Signal Detection: contributions from Image Recognition, Speech Recognition, and Acoustic Event Detection.

Author:

Ivan Kiskin

Supervisor:

Professor Stephen Roberts

Trinity 2017

1 Introduction

This literature review critically covers contributions to signal detection from deep learning in the context of speech detection, image recognition, and acoustic event detection. We define signal detection as the ability to discern information-bearing patterns from random patterns that contain no information. The research is examined in chronological order, following the progression of the state of the art in specific fields. We narrow our focus to audio signals and images and highlight potential areas in need of further improvements.

Section 2 focuses on the speech recognition field, due to its competitive nature and associated rich publication record. The progress in state-of-the-art speech recognition was fuelled by both industry and academia due to its importance in human-machine interaction systems, that arguably are essential to general artificial intelligence [39,23].

Section 3 focuses on image recognition, the most competitive field in recent years due to the multitude of established image recognition and object detection challenges (with over 50 participating institutions by 2014 in the ImageNet Large Scale Visual Recognition Challenge alone [43]). As a result, innovations have become influential to related fields in the scientific literature.

Our focus then shifts to bird species recognition and acoustic event detection in Sections 4 and 5 as fields where innovation was driven by progress in both speech and image recognition. In each field we remark how advancements are relevant to the subject of our paper: a practical application in mosquito detection.

The literature review is followed by the research proposal which includes a concise summary of completed work, future work, and a risk assessment. Future work is broken down into short, medium and long-term goals in Section 6 and presented on a Gantt chart. We conclude by summarising general scientific trends discussed in this literature review and our paper in Section 7.

2 Speech Recognition

Due to its significance in academia and industry, great attention is devoted to the field of speech recognition. Speech recognition has been a prominently researched field due to its critical role in human-machine

interactions. The ability to accurately transcribe spoken text would influence a diverse range of applications, and hence has been the subject of prominent study [23].

We choose to concentrate on work postdating approximately 1990, as it forms the basis of modern methods which are most applicable to current research. Major breakthroughs in early methods came from a paradigm shift to more statistics-based methods, such as the Hidden Markov Model (HMM) [23]. An HMM is a statistical model, where the system is modelled as a Markov process with discrete latent variables. An HMM can be interpreted as an extension of a mixture model [30] in which latent variables are related by a Markov process [1]. The discrete multinomial latent variables describe which component of the mixture is responsible for generating the corresponding observation.

By defining a state transition matrix (between states s_i) and the state output distribution, the HMM can be used as a generative classifier (for predicting y_i). An example scenario is given in Figure 1.

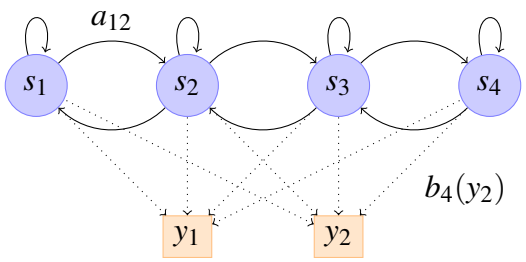


Fig. 1: An HMM with 4 states which can emit 2 discrete symbols y_1 or y_2. a_{ij} is the probability to transition from state s_i to state s_j. $b_j(y_k)$ is the probability to emit symbol y_k in state s_j. In this particular HMM, states can only reach themselves or the adjacent state.

The HMM formed the definitive state of the art in the early 1990s [38], with DSP hardware applications being sold in a market for office systems, manufacturing, telecommunications, and other areas [39, p.487].

To improve the quality of feature representations that are used as inputs to the HMM, there was a shift to hybrid Deep Neural Network Hidden Markov systems (DNN-HMMs) from Gaussian Mixture Model HMMs (GMM-HMMs). GMMs in GMM-HMM systems have a serious shortcoming in that they are statistically inefficient for modelling data that lie on or near a non-linear manifold in the data space [19]. This is problematic as speech is inherently highly non-linear due to the physical processes involved. DNN-HMM design follows traditional GMM-HMM systems. As the DNN is used to model the posterior

probability of a state given an observation vector, a prior is required. The prior is commonly obtained with the Viterbi algorithm of the GMM-HMM approach and is given in the form of an initial labelled state sequence [29]. In acoustic modelling for large vocabulary continuous speech recognition (LVCSR) tasks, DNN-HMMs were shown by Dahl et al. [4] to give significant gains over state-of-the-art GMM-HMM systems in a wide variety of small and large vocabulary tasks. Reductions in relative error of 16.0% and 23.2% were demonstrated using context-dependent DNN-HMMs over context-dependent GMM-HMMs. Additionally, it was found that DNNs gave much higher accuracy when large training datasets supported the greatly increased model capacity [7].

In GMM-HMM systems the feature representation is hand-crafted by the user, typically in the form of Mel-frequency cepstral coefficients (MFCCs). The Fourier transform can be temporally windowed with a smoothing window function to create a Short-time Fourier transform (STFT). MFCCs create lower-dimensional representations by taking the STFT, applying a non-linear transform (the logarithm), pooling, and a final affine transform.

MFCCs, GMMs and HMMs co-evolved in speech recognition in an era with limited computational power. MFCCs lose significant information from the sound wave, but preserve higher order low-dimensional information that is required for discrimination. This is to partially overcome the very strong conditional independence assumptions of HMMs [34], as well as improve tractability of the problem. Whereas MFCCs led to significant accuracy improvements in GMM-HMM systems despite their known loss of information introduced, further advancements came from reducing the specificity of the signal transforms that were used to form feature representations. For example, Mohamed et al. [34] showed significantly lowered automatic speech recognition errors using large-scale DNNs when moving from the MFCC features back to more primitive (Mel-scaled) filter-bank features. The results implicated that DNNs are able to learn a better representation from Mel-scaled features than the final step of the MFCC transform – the discrete cosine transform. The original use of the cosine transform is justified by its approximate de-correlation of feature components. This is of importance to using diagonal covariance matrices with GMMs. This restriction however does not apply to deep learning models, as their strength in modelling data correlation makes the transform redundant. There is a possibility of future work to focus on training

directly on raw waveforms, although current research has been unable to extract equivalent performance in audio applications, discussed further at the end of Section 5.

HMMs suffer inherent limitations in speech recognition applications. The assumptions that successive observations are independent, as well as the probability of being in a state only depending on the state at the previous time step, do not strictly hold true [38]. The final major change in recent literature was the replacement of the HMM in its entirety, as well as the introduction of convolutional neural networks (CNNs) and Long Short-Term Memory networks (LSTMs) for feature generation. In large-scale speech tasks, deep convolutional nets have been shown to significantly outperform DNN-HMM systems by Sainath et al. [44]. Their method trialled a combination of salient features (logarithmically spaced Fourier coefficients and other features stacked into one vector as used in [48]) to train the CNNs. Due to the relevancy to current research, as well as the method implemented in our paper [24], we choose to address particular implementation details. Compared to fully end-to-end trained DNNs, CNNs capture translational invariance with far fewer parameters by averaging the outputs of hidden units in different local time and frequency regions. The CNNs and fully-connected DNNs used 1024 hidden units per fully connected layer with sigmoid non-linearities. The last layer is a softmax layer with 512 output targets. The 512 targets were determined as a result of clustering context-dependent GMM-HMM states. The authors note decreasing performance when increasing layer depth beyond 2. This indicates that increasing network depth will not always yield performance benefits, as was consistent with our application [24, Section 3.2].

Further improvement was found with the application of recurrent neural networks by Graves et al. [16]. The authors noted at the time of publication in 2013 that while HMM-RNN systems had seen a recent revival, they did not perform as well as deep networks. As with Sainath et al., the authors favoured to train RNNs end-to-end fully, instead of combining RNNs with HMMs. This approach allowed RNNs to exploit their larger state-space capacity and richer dynamics, as well as avoid using potentially incorrect alignments as training targets. The combination of LSTMs with end-to-end training has proved especially effective for cursive handwriting recognition [15,17]. Graves et al. have shown that combination of deep, bidirectional LSTM RNNs with end-to-end training gave state-of-the-art results in phoneme recognition,

and note that the next step is to extend the system to large vocabulary speech recognition. The authors further propose to combine CNNs and deep LSTMs.

Notably, the authors use the same feature representation (Mel-scaled coefficients, with their first and second derivatives and a few extras). We propose the investigation of moving towards autonomous learning of the feature space (as discussed in the Proposal in Section 6) for further performance benefits.

Transferability Speech recognition, whether it is predicting particular phonemes, or their sequences, requires predicting a dynamic signal that is highly non-stationary and embedded in noise. The field has strong transferability to general audio event recognition, and indeed methods developed in speech recognition have been frequently applied as black-boxes with little modification to tasks such as bird recognition, which we discuss in Section 4. In addition, it has inspired many methods used in insect detection, further discussed in the literature review of our paper [24, Section 2.1].

3 Image recognition

Deep learning methods in image recognition have strong parallels with deep learning methods in acoustics. This stems from the fundamental frequency representation currently in use with sounds – the spectrogram. A spectrogram representation is a two-dimensional matrix with intensity measurements as entries. This is equivalent to a black and white image representation.

With a vast number of entries to a variety of image recognition tasks [43,42], a wide range of subfields in neural networks underwent incremental improvements. We structure this section by noteworthy changes in characteristics of the networks.

3.1 Depth

A primary driving factor of progress in deep learning has been network depth. This was first formally addressed by Simonyan and Zisserman [46] by steadily increasing the depth of the network by the addition of convolutional layers, while fixing the remaining parameters of the architecture. This was feasible due to the use of very small convolution filters in all layers. The effect was to increase model complexity by allowing more non-linearities, and hence make the decision function more discriminative. Without

spatial pooling in between convolutional layers, two 3×3 layers have an effective receptive field size of a single 5×5 layer. In addition, this results in a decrease of parameters (with an 81% reduction when replacing a 7×7 receptive field with three 3×3 layers). As a result, in addition to accuracy improvements, faster gradient back-propagation and hence training times were also achieved. This trend has continued in ILSVRC submissions, with challenge winners employing CNNs with depth 3, 19, and finally 152 [27,46,18], ranging from 2012 to 2016.

3.2 Residual learning

A recent advancement was presented by He et al. [18] as deep residual learning. Their submission resulted in first place placements on ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation in ILSVRC & COCO 2015 competitions. The authors argue that the breadth of categories in which excellent results are achieved is strong evidence that the method is applicable in other vision and non-vision problems. The causal mechanics are explained as follows. The authors define $\mathscr{H}(x)$ as an underlying mapping to be learnt by a few stacked layers of the network, with x denoting the inputs to these layers. The authors argue that "if one hypothesizes that multiple nonlinear layers can asymptotically approximate complicated functions, then it is equivalent to hypothesize that they can asymptotically approximate the residual functions, i.e., $\mathscr{H}(x) - x$". Instead of approximating $\mathscr{H}(x)$, the equivalent layers are approximating the residual function $F(x) := \mathscr{H}(x) - x$. The original function thus becomes $F(x) + x$.

As a result, the mathematically equivalent formulations can be implemented with architecture modifications via identity shortcut mappings [18, Figure 3]. The improved rate of learning deep architectures under this formulation allows the use of extremely deep networks. The 152-layer residual network was at the time of submission the deepest network ever presented at ImageNet, while still having lower complexity than VGG-19 proposed in [46]. The improvement is in line with the approach to depth discussed in the previous subsection.

3.3 Data augmentation

Prominent applications where neural networks significantly outperformed other means of classification all involved large datasets. Since its inception in 2009, the CIFAR10/100 dataset [26] served as an extremely useful testbed with over 1000 associated cited publications presently. The set contained 60,000

labelled samples, which were split into either 10 or 100 classes. At this date, since its introduction ImageNet contains over 14 million images, of which over a million come with bounding box annotations. For details of the ImageNet construction the interested reader is directed to [42, Section 3], which discusses the crowdsourcing strategy for its three categories: image classification, single-object localisation, and object detection. Simonyan and Zisserman [46] attribute much increase in performance due to computational hardware (GPU computing) and dataset availability. However with scarce data, and even with healthy datasets, it has become common practice to generate additional training data by applying transformations to the original data [20]. The practice can be traced back to [45], where the authors note that synthesising plausible transformations of data is simple, however learning transformation invariance by a model can be arbitrarily complicated. If the data is both scarce, and the distribution to be learned has transformation-invariance properties, generating additional data using transformations may improve performance. Indeed this practice has become widely adopted in the neural network literature, with common transformations including translations and horizontal reflections. In addition, modifications to intensities of RGB channels in training images [27] have been applied (similar to synthesising and adding various noise profiles for acoustics). The authors note that the transformations artificially increase the dataset size by over a factor of 2048 for the affine transformations alone. Despite the high inter-dependence of samples, the augmentation offers a form of normalisation, allowing the use of deeper networks while avoiding overfitting. As established, depth plays an important role in overall performance, which places great emphasis on producing as large of a training data set as possible. The augmentation has become common practice as evidenced in competition submission winners throughout the years [46,18,27]. This practice also served useful in audio applications, as discussed in Section 4. Due to the scarcity of samples in our application, we propose to improve on our paper by augmenting our datasets with transformed versions of the training samples. This could include a variety of noise profiles, slight alterations in audio intensity, as well as minor pitch shifts.

3.4 Deep Kernel Learning

To gain insight into understanding deep learning models, a practice gaining traction in recent years is to view CNNs and other network configurations as instances of Gaussian Processes, or other kernel-learning methods [10]. In transition to fully probabilistic models, some works attempt to build hybrid neural

network Gaussian process models to retain the performance benefit of deep learning methods, while also conforming to probabilistic predictions. Wilson et. al [51] showed that jointly learning deep kernel parameters has advantages over training a GP applied to the output layer of a trained deep neural network. This allows more principled handling of uncertainty. In addition, the jointly learned model outperformed the equivalent deep neural network, as well as more primitive CNN/DBN-GP combinations, in a range of regression tasks. A particular example is given on the benchmark MNIST digit recognition dataset. The work forms the basis of a medium-term proposed goal addressed in Section 6.

3.5 Transferability

Despite immense popularity in image recognition, until recently there were relatively few convincing applications in signal processing. Similarly to how handcrafted scale-invariant feature transforms, Difference of Gaussians, and Histograms of Oriented Gradients [31,5] were phased out with the advent of deep learning, handcrafted feature descriptors are beginning to undergo the same fate in acoustic signal processing. We illustrate this in Section 4 with a bird recognition case study equivalent to ILSVRC. We further support this claim by reviewing advancements in the general field of acoustic event detection in Section 5.

4 Bird recognition

A prominent example of the recent shift in methodology towards deep learning is the BirdCLEF bird recognition challenge. The challenge consists of the classification of bird songs and calls into up to 1500 bird species from tens of thousands of crowd-sourced recordings. The introduction of deep learning has brought drastic improvements in mean average precision (MAP) scores. The highest MAP score of 2014 was 0.45, achieved by combining two main categories of features for classification: parametric acoustic features consisting of spectral features, cepstral features, energy features and voicing-related features, and probabilities of species-specific spectrogram segments [13]. The feature sets were used in combination with unsupervised dimensionality reduction to train randomized ensemble decision trees. The runner-up used an MFCC feature foundation, with further PCA whitening, to train random forests (a full pipeline is given in [49, Figure 3]).

Participants in the following year of 2015 achieved slightly lower scores due to the increased difficulty of the task [14], with the best results achieved by the same participant in very similar fashion. However, 2016 marked the introduction of deep learning methods to the competition. Overall MAP scores improved to 0.69, outperforming the previously dominant method, which scored 0.58 on this occasion following small tweaks [22]. The impressive performance gain came from the utilisation of well-established convolutional neural network practice from image recognition. By transforming the signals into windowed Fourier transform segments, the training data is represented by 2D matrices analogous to single-channel image representations. Following median-based segment de-noising, the authors employed a 5-layer CNN network on input sections of approximately 3 seconds in length. This allowed capturing and encoding the temporal dependencies of repeating birdsong or call patterns. The authors also noted no performance benefit from using 1D convolution with height equal to spectral width (frequency). As with common practice in image recognition, data augmentation was performed in the form of noise injection. We note that the implementation, while effective, was unremarkable compared to state-of-the-art submissions of recent years seen in ILSVRC, so there may be further room for improvement. Coupled with the scarcity of overall submissions, there is strong evidence of a lack of maturity in this particular research area, which is expected to advance rapidly since the arrival of deep methods and over-saturation of submissions to image analysis challenges.

5 Acoustic Event Detection

To draw further parallels to the application in our paper, we study recent research in Acoustic Event Detection (AED). This is the field that deals with detecting and classifying non-speech acoustic signals, with intent to convert continuous signals into a sequence of event labels associated with start and end times [11]. In recent years the field has attracted more research effort, with the foundation of dedicated challenges such as such as CLEAR [35], and D-CASE [12], which involve detecting a known set of acoustic events indoors.

The increased attention to deep learning methods can be evidenced at the Music Information Retrieval Evaluation eXchange (MIREX), where example tasks include recognising musical chords, tempos, and genres [33]. The majority of challenges require detecting a specific composition of frequencies localised

in time. For instance, classifying musical instruments that play identical notes requires detection of the harmonic structure of the tones produced. This is analogous to many real-world problems, as the ratios of amplitudes of harmonics define particular sounds. Correct detection could allow discrimination of information-bearing harmonic stacks from background noise occurring at the same fundamental frequency. For example, the task may be to identify an insect flying in a room with constant noise produced by a fan at the same fundamental frequency. As part of our candidate challenge of accurately detecting mosquitoes in our paper [24], this case was evidenced in our collected training data. A meta-analysis of MIREX concluded that the state of the art stagnated in a variety of such classification and detection tasks (Humphrey et al. [21]) between 2007 and 2012. This was attributed to the use of traditional methods for feature extraction and classification. It was suggested deep learning can help overcome three deficiencies associated with traditional methods:

1. Hand-crafted feature design is not sustainable,
2. Shallow processing architectures struggle with latent complexity of real-world phenomena,
3. Short-time analysis cannot naturally capture higher-level information.

Evidence for (1) in this field was presented by Espi et al. [11], who argue in favour of using more generic feature representations at the input layer of CNNs. This allows the network to truly exploit the feature learning ability of deep learning. Such claims are in line with the move to less complex features discussed in speech recognition (Section 2). Indeed, since the assessment, there has been a paradigm shift. As an example, in 2016 challenge results [33], every (all-time) top performer in the four chord recognition challenges (Korzienowski, Widmer [25,32]) featured forms of deep networks operating on simple feature representations. Their contribution was the ability to discover feature representations from a basic frequency representation (the STFT or constant-Q transform) that were better suited to use with a conditional random field than hand-crafted representations. The authors note that classifiers relying on hand-crafted representations benefited from using a further transform on the STFT (a logarithmic tiling by means of a constant-Q transform [2]), whereas the neural network was insensitive to the base representation.

Further evidence for (1) is presented by Espi et al. [11] in a meta-study of acoustic events. The authors noted that pooling, especially in frequency, seemed to degrade the network's ability to detect events.

As a result, use of salient features (MFCCs) would result in worse overall performance than learning features with deep models directly from the spectrogram (STFT). Additionally, multiple time-frequency resolutions may be required to parameterise the STFT, which underpins a shortcoming of the windowed Fourier transform – a fixed time-frequency resolution. This shortcoming has been well-studied in the signal processing community, yet it is rare to find considerations of this at the intersection of acoustics and deep learning research. Therefore we believe this warrants basis for further research, as addressed in the Proposal in Section 6. For mosquito detection we addressed this shortcoming in our paper [24, Section 2.2].

Additional support for (1) is presented by Phan et al. [37], where the authors compare popular traditional classifier combinations to CNNs and DNNs. With the use of variable convolutional filter sizes, relatively shallow networks with 1-max pooling strategies significantly outperformed traditional approaches on their given dataset across a range of SNRs. Strongest improvements where evidenced in low SNR scenarios, where traditional methods would fully break down. The signal data was obtained from the Real World Computing Partnership Sound Scene Database in Real Acoustic Environments. The authors attempted to minimise bias by randomly sub-sampling the dataset into 2000, 500, and 1500 event instances for training, validation, and testing purposes, respectively.

Irrespective of finer details, a very clear benefit to CNNs over HMM/SVM combinations is evidenced with the introduction of noise. Additionally, there is a reliable benefit to CNNs over DNNs as consistent with speech recognition literature. Furthermore, a general trend for generic spectrogram features to be preferable to more salient features such as MFCCs and Gabor features was demonstrated with the use of CNNs. The CNN architecture was both simple and efficient. We note varying convolutional size filters were employed, that were learnt from data automatically. As often it is unclear over what time-scale events happen predictably exactly, this may be a highly useful feature to implement in future works.

Further CNN proponents (and support for claim (3)) include the ability to better capture long-term information (dependence of sample on previous samples). In the limit an example is given in the WaveNet paper [36] where each sample is conditioned on every preceding sample by using causal correlation filters. On the other hand, conventionally, features derived from short-time signals are limited to the information content contained within each segment. As a result, if some musical event does not occur

within the span of an observation—a motif that does not fit within a single frame—then it simply cannot be described by that feature vector alone. This is in part addressed by the Markov property as regularly used in speech detection, but imposes a limit on conditioning features on only a single previous observed feature.

Learning a frequency decomposition Taking Humphrey's claim (1) to the limit and to further facilitate autonomous behaviour, the next potential research challenge would be to operate on raw data, rather than raw spectrogram data, as alluded to frequently in the literature, even in significant papers [28]. Spectrogram representations are not "raw", as they require both a parameterised Fourier transform, as well as a smoothing window at a fixed resolution (or in some case multiple transforms at different resolutions to detect events occurring over differing time scales [11]) to construct.

While few papers have achieved noteworthy results in this field, Dieleman et al. [9] show promising results in a range of music information retrieval tasks. The authors note that while performance of STFT methods was not matched, the networks were able to autonomously discover frequency decompositions from raw audio, as well as phase and translation-invariant feature representations. Furthermore, successful applications have arisen in raw audio *generation* [36]. The authors employed a convolutional architecture, with the predictive distribution for each audio sample conditioned on all previous ones, to synthesise speech and music sequences with remarkable results. A possible stepping stone towards moving towards fully autonomous feature extraction may be learning wavelet bases that suit the data. We address this in our research proposal, and demonstrate practical performance improvements in our paper [24, Section 5].

6 Proposal

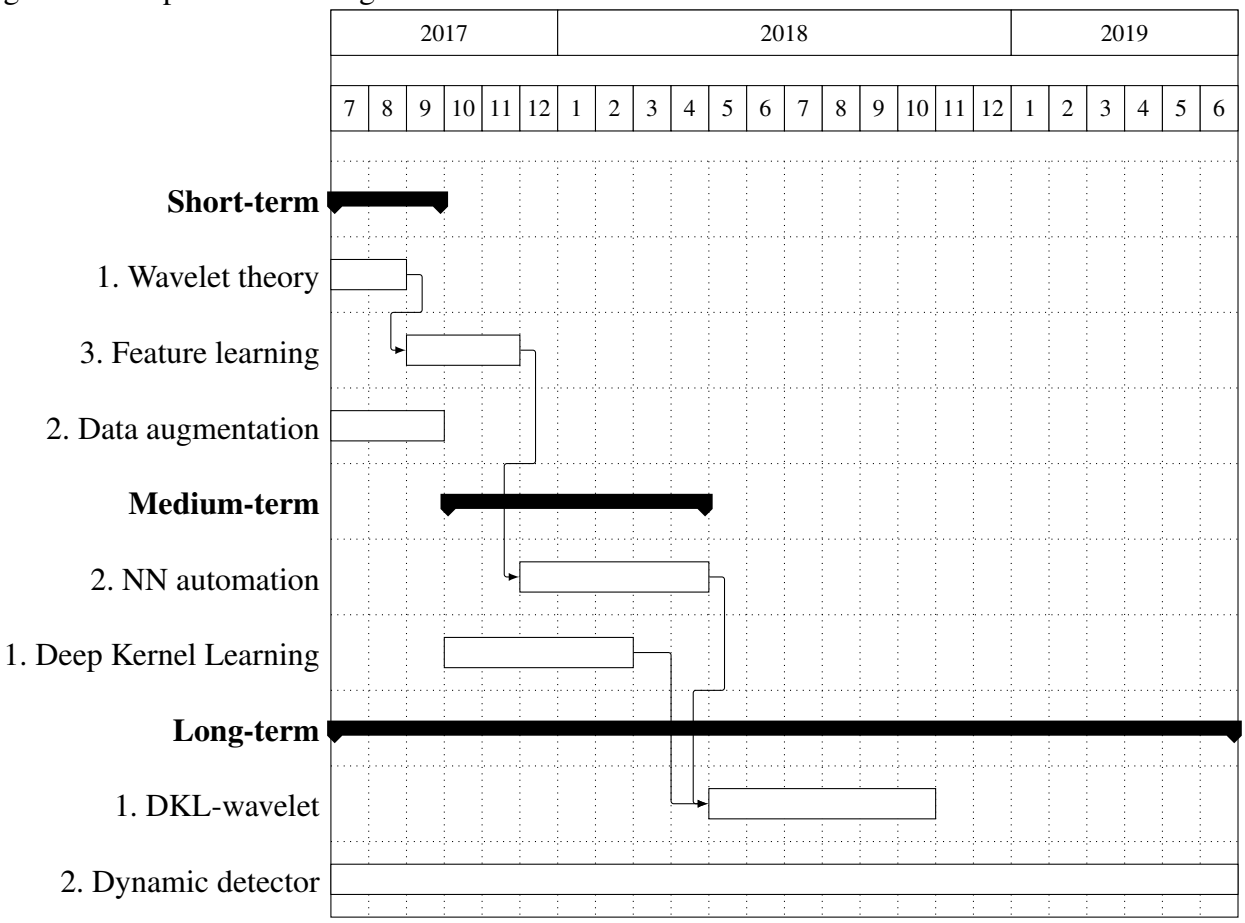
6.1 Completed Work

Completed work to date presents an application of deep learning for acoustic event detection in a challenging, data-scarce, real-world problem. Our candidate challenge was to accurately detect the presence of a mosquito from its acoustic signature. To this end we developed convolutional neural networks operating on wavelet transformations of audio recordings. This was accomplished with the aid of a contin-

uous wavelet transform with the bump mother wavelet – a popular choice in traditional time-frequency analysis methods in signal processing [6]. Furthermore, we interrogated network predictive power by visualising the statistics of network-excitatory samples. The visualisations offered insights into relative informativeness of components in the detection problem. It was shown that the network correctly learned acoustic properties of the mosquito, rather than peculiarities of the recording (such as microphone noise). Detection was achieved with performance metrics significantly surpassing those of existing algorithmic methods, as well as marginally exceeding those attained by individual human experts. The existing algorithmic methods trialled included Support Vector Machines, Random Forests, Naive Bayes classifiers, and Gaussian processes.

6.2 Further Work

Future proposed work is structured by approximate timelines. We define short-term goals as targets to address in two to three months, medium-term as three to ten months, and long-term as over ten months. The goals and dependencies are given in the Gantt chart below.



Short-term

1. Following our work on mosquito detection with the use of wavelet-conditioned convolutional neural networks, it is hypothesised that the wavelet transform is more effective than the STFT from an information-theoretic viewpoint. We wish to explore this hypothesis formally.
   (a) If proven true, this may have a profound effect on the deep learning literature, where the baseline state-of-art method involves taking the STFT before further learning. A possible way to test this would involve taking a pre-trained network from a paper, repeating all the steps with both STFT as a baseline, as well as the wavelet transform.
   (b) With the abundance of data, if deep representations learned from STFTs outperform those learned from wavelets, there is still valuable research to be made for data-scarce scenarios. As applications in the real world often involve expensive data collection and labelling, we consider this research avenue worthy of pursuit.

   Should benefits be demonstrated in either scenario, the focus of work will be to improve on our submitted paper, and further focus on the computational efficiency of the CWT algorithm implementation. A suitable alternative may be to use a discrete wavelet transform (DWT) that can be executed at the same (or even lower) computational complexity than the FFT [41].
2. A further short-term goal involves utilising data augmentation for training neural networks. In our mosquito detection application data is very expensive, so being able to generate data probabilistically using a scaffold model may help scale performance for multiple species and remain competitive with traditional classification methods, even in extremely data-limited scenarios. Furthermore, we will consider Generative Adversarial Networks [40], which may allow the generation of representative samples with our network, helping alleviate the poverty of our data.
3. Our current method uses a parameterised wavelet transform as feature space conditioning for neural network computations. As a short- to medium-term goal we wish to extend this by also learning wavelet (or other) feature representation spaces jointly with the the remaining network parameters. This involves learning a kernel for frequency transformations.

Medium-term

1. The first medium-term goal is to bring probabilistic reasoning to our detection schemes, while maintaining the impressive performance offered by deep learning approaches. A major downside with neural networks is the inability to obtain model uncertainty, despite extensions originally discussed by Denker and LeCunn [8]. The aforementioned does not however apply to neural networks trained with Deep Kernel Learning (DKL) [51]. As part of DKL, kernel parameters are learnt jointly with neural network hyperparameters the authors. The authors demonstrated that jointly learning all DKL parameters has advantages over training a GP applied to the output layer of a trained deep neural network. With availability of code, as well as demonstrated benefits on the MNIST dataset, a simple medium-term milestone would be to apply a DKL method with our equivalent wavelet-conditioned CNN to the mosquito data available at that stage of the project.
2. A further medium-term goal is to work on automation of neural network architecture optimization. This would build on work by Swersky et al. [50] which utilises a Bayesian Optimisation kernel [47] for conditional parameter spaces to infer model architecture and corresponding hyperparameters. In line with our short-term objectives to infer wavelet feature representations, we would also like to build upon the BayesOpt framework to include a mechanism for learning hyperparameters in hybrid wavelet-CNN networks.

Long-term

1. In the long-term we wish to extend DKLs to our model which we will develop to jointly infer wavelet feature representation spaces and neural network parameters.
2. The ultimate long-term goals are to build a dynamic detector that is able to also respond to changes in its environment. This is particularly relevant to signal detection over traditional image detection, due to the increased variability of the stimulus in time-series compared to static image representations. As part of the adaptability, the algorithm is required to recognise when the model is insufficient for explaining novel data. This may be incorporated using feedback mechanisms by drawing from ideas in evolutionary algorithms [3]. We may wish to feed back uncertain results to a crowdsourcing platform to aid with creating a lifelong learning algorithm.

6.3 Risk Assessment

We provide a risk assessment to identify critical points or uncertainties, and indicate how we will manage these. The assessment is categorised by high and medium risks, as well as their impact factors.

High-level risks

1. High impact: Benefit to using wavelets may have been purely due to saliency of feature, in particular the logarithmic tiling of the time-frequency plane, working well for small datasets
2. Medium impact: Advances in deep learning render our classification algorithms obsolete
3. Low impact: Changes in data acquisition render our algorithm impractical or irrelevant (affecting the need to port or research efficiency of algorithms such as the DWT)

Mitigating measures

1. Research why performance benefit was shown for small datasets, and investigate CNN performance relative to other salient features. The CNN was trialled with a set of salient features, which were shown not to improve performance, so it is unlikely that only the saliency is responsible for the boost in performance when combining CNNs with the wavelet transform.
2. This would not discredit our scientific contribution, but would affect practical implementations of algorithms on portable devices. This has an effect on the outcome of the HumBug project, but not on the course of the DPhil. A further mitigation measure is to combine future advances in the field with our own research advances.
3. As with (2), this affects practical implementations. We could shift our focus to applications with dataset availability to suit our research needs, or adapt the algorithms to match the new data availability. In the case of neural networks, this would involve adding further layers and model complexity without overfitting according to the new data size constraints.

Medium-level risks

1. High impact: Inability to reproduce impressive performance results of CNNs when working with deep kernel methods, or other probabilistic models

2. Medium impact: Proposed tasks are not met by desired date
3. Low impact: Inability to reproduce performance when working towards autonomous frequency representation discovery

Mitigating measures

1. Work towards a hybrid probabilistic neural network model to leverage gains in both fields
2. Set regular milestones and consider conference deadlines to focus work towards specific scientific contributions
3. Not necessarily a problem, as currently publishing with raw data methods is accepted with poorer performance, as long as an original scientific contribution is made [9].

7 Conclusion

In the fields of Image Recognition, Acoustic Event Detection, and Speech Recognition, there has been a paradigm shift in two general areas.

The first area concerns feature representations. Representations have evolved from meticulously hand-crafted descriptors, such as the Scale-Invariant Feature Transform in images, and Mel-Frequency Cepstral Coefficients in acoustics, to features learned autonomously by neural networks. Autonomous representations are discovered from raw images in image recognition, and from simple spectral representations in acoustics. However, these simple representations are often not as simple as they are portrayed, requiring numerous parameters, and sometimes multiple resolutions, for optimum classification performance. This is most commonly seen with the overwhelming practice of using Short-Time Fourier transforms in conjunction with convolutional neural networks.

The second major evolution has taken place in integrating classification and feature extraction into a single pipeline. In recent years, the most common practice is to use fully end-to-end trained convolutional neural networks, replacing classification algorithms, such as SVMs and random forests, operating on hand-crafted feature representations.

Our research is focused on both improving and automating learning of the feature space, as well as network architectures. Provisionally, this could be achieved with the use of wavelet transformations,

with a focus on discovering bases that suit the data autonomously. Future work will aim to address feature, parameter, and architecture inference jointly to work towards a detector that is able to adapt to its environment dynamically.

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
2. Brown, J.C.: Calculation of a constant Q spectral transform. The Journal of the Acoustical Society of America 89(1), 425–434 (1991)
3. Coello, C.A.C., Lamont, G.B., Van Veldhuizen, D.A., et al.: Evolutionary Algorithms for Solving Multi-Objective Problems, vol. 5. Springer (2007)
4. Dahl, G.E., Yu, D., Deng, L., Acero, A.: Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. IEEE Transactions on Audio, Speech, and Language Processing 20(1), 30–42 (2012)
5. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. vol. 1, pp. 886–893. IEEE (2005)
6. Daubechies, I., Lu, J., Wu, H.T.: Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool. Applied and computational harmonic analysis 30(2), 243–261 (2011)
7. Deng, L.: Achievements and Challenges of Deep Learning–From Speech Analysis and Recognition to Language and Multimodal Processing. In: Fifteenth Annual Conference of the International Speech Communication Association (2014)
8. Denker, J.S., Lecun, Y.: Transforming Neural-Net Output Levels to Probability Distributions. In: Advances in Neural Information Processing Systems. pp. 853–859 (1991)
9. Dieleman, S., Schrauwen, B.: End-to-end learning for music audio. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6964–6968. IEEE (2014)
10. Duvenaud, D.: Automatic model construction with Gaussian processes. Ph.D. thesis, University of Cambridge (2014)
11. Espi, M., Fujimoto, M., Kinoshita, K., Nakatani, T.: Exploiting spectro-temporal locality in deep learning based acoustic event detection. EURASIP Journal on Audio, Speech, and Music Processing 2015(1), 1 (2015)
12. Giannoulis, D., Benetos, E., Stowell, D., Rossignol, M., Lagrange, M., Plumbley, M.D.: Detection and Classification of Acoustic Scenes and Events: An IEEE AASP challenge. In: Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on. pp. 1–4. IEEE (2013)
13. Goëau, H., Glotin, H., Vellinga, W.P., Planqué, R., Rauber, A., Joly, A.: LifeCLEF bird identification task 2014. In: CLEF: Conference and Labs of the Evaluation Forum (2014)
14. Goëau, H., Glotin, H., Vellinga, W.P., Planqué, R., Rauber, A., Joly, A.: LifeCLEF bird identification task 2015. In: CLEF: Conference and Labs of the Evaluation Forum (2015)
15. Graves, A., Liwicki, M., Bunke, H., Schmidhuber, J., Fernández, S.: Unconstrained On-Line Handwriting Recognition with Recurrent Neural Networks. In: Advances in Neural Information Processing Systems. pp. 577–584 (2008)
16. Graves, A., Mohamed, A.R., Hinton, G.: Speech Recognition with Deep Recurrent Neural Networks. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. pp. 6645–6649. IEEE (2013)

17. Graves, A., Schmidhuber, J.: Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks. In: Advances in Neural Information Processing Systems. pp. 545–552 (2009)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
19. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al.: Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Processing Magazine 29(6), 82–97 (2012)
20. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 (2012)
21. Humphrey, E.J., Bello, J.P., LeCun, Y.: Feature Learning and Deep Architectures: New Directions for Music Informatics. Journal of Intelligent Information Systems 41(3), 461–481 (2013)
22. Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W.P., Champ, J., Planqué, R., Palazzo, S., Müller, H.: LifeCLEF 2016: Multimedia Life Species Identification Challenges. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 286–310. Springer (2016)
23. Juang, B.H., Rabiner, L.R.: Automatic Speech Recognition–A Brief History of the Technology Development. Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara 1, 67 (2005)
24. Kiskin, I., Orozco, B.P., Windebank, T., Zilli, D., Sinka, M., Willis, K., Roberts, S.: Mosquito Detection with Neural Networks: The Buzz of Deep Learning. arXiv preprint arXiv:1705.05180 (2017)
25. Korzeniowski, F., Widmer, G.: Feature Learning for Chord Recognition: The Deep Chroma Extractor. arXiv preprint arXiv:1612.05065 (2016)
26. Krizhevsky, A., Hinton, G.: Learning Multiple Layers of Features from Tiny Images (2009)
27. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet Classification with Deep Convolutional Neural Networks. In: Advances in Neural Information Processing Systems. pp. 1097–1105 (2012)
28. Lee, H., Pham, P., Largman, Y., Ng, A.Y.: Unsupervised feature learning for audio classification using convolutional deep belief networks. In: Advances in Neural Information Processing Systems. pp. 1096–1104 (2009)
29. Li, L., Zhao, Y., Jiang, D., Zhang, Y., Wang, F., Gonzalez, I., Valentin, E., Sahli, H.: Hybrid Deep Neural Network–Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition. In: Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on. pp. 312–317. IEEE (2013)
30. Lindsay, B.G.: Mixture models: Theory, Geometry and Applications. In: NSF-CBMS Regional Conference Series in Probability and Statistics. pp. i–163. JSTOR (1995)
31. Lowe, D.G.: Object Recognition from Local Scale-Invariant Features. In: Computer vision, 1999. The proceedings of the seventh IEEE international conference on. vol. 2, pp. 1150–1157. IEEE (1999)
32. MIREX: 2016 results abstracts (2016), `http://www.music-ir.org/mirex/abstracts/2016/FK2.pdf`, accessed: 2017-04-07
33. MIREX: 2016 results summary (2016), `http://www.music-ir.org/mirex/results/2016/mirex_2016_poster.pdf`, accessed: 2017-04-07
34. Mohamed, A.R., Dahl, G.E., Hinton, G.: Acoustic Modeling using Deep Belief Networks. IEEE Transactions on Audio, Speech, and Language Processing 20(1), 14–22 (2012)

35. Mostefa, D., Moreau, N., Choukri, K., Potamianos, G., Chu, S.M., Tyagi, A., Casas, J.R., Turmo, J., Cristoforetti, L., Tobia, F., et al.: The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms. Language Resources and Evaluation 41(3-4), 389–407 (2007)
36. van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K.: Wavenet: A Generative Model for Raw Audio. CoRR abs/1609.03499 (2016)
37. Phan, H., Hertel, L., Maass, M., Mertins, A.: Robust Audio Event Recognition with 1-Max Pooling Convolutional Neural Networks. arXiv preprint arXiv:1604.06338 (2016)
38. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
39. Rabiner, L.R., Juang, B.H.: Fundamentals of speech recognition. PTR Prentice Hall (1993)
40. Radford, A., Metz, L., Chintala, S.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv preprint arXiv:1511.06434 (2015)
41. Rioul, O., Duhamel, P.: Fast algorithms for discrete and continuous wavelet transforms. IEEE transactions on information theory 38(2), 569–586 (1992)
42. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) 115(3), 211–252 (2015)
43. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: ImageNet Large Scale Visual Recognition Challenge. arXiv preprint arXiv:1409.0575 (2014)
44. Sainath, T.N., Kingsbury, B., Saon, G., Soltau, H., Mohamed, A.R., Dahl, G., Ramabhadran, B.: Deep Convolutional Neural Networks for Large-Scale Speech Tasks. Neural Networks 64, 39–48 (2015)
45. Simard, P.Y., Steinkraus, D., Platt, J.C., et al.: Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis. In: ICDAR. vol. 3, pp. 958–962 (2003)
46. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556 (2014)
47. Snoek, J., Larochelle, H., Adams, R.P.: Practical Bayesian Optimization of Machine Learning Algorithms. In: Advances in Neural Information Processing Systems. pp. 2951–2959 (2012)
48. Soltau, H., Saon, G., Kingsbury, B.: The IBM Attila speech recognition toolkit. In: Spoken Language Technology Workshop (SLT), 2010 IEEE. pp. 97–102. IEEE (2010)
49. Stowell, D., Plumbley, M.D.: Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. PeerJ 2, e488 (2014)
50. Swersky, K., Duvenaud, D., Snoek, J., Hutter, F., Osborne, M.A.: Raiders of the Lost Architecture: Kernels for Bayesian Optimization in Conditional Parameter Spaces. arXiv preprint arXiv:1409.4011 (2014)
51. Wilson, A.G., Hu, Z., Salakhutdinov, R., Xing, E.P.: Deep Kernel Learning. In: Proceedings of the 19th International Conference on Artificial Intelligence and Statistics. pp. 370–378 (2016)