Optimisation of Phonetic Aware Speech Recognition through Multi-objective Evolutionary Algorithms

Jordan J. Bird[1], Elizabeth Wanner[2], Anikó Ekárt[3], Diego R. Faria[4]

[1,4]*Aston Robotics, Vision and Intelligent Systems (ARVIS)*
[1,2,3,4]*Computer Science Department*
School of Engineering and Applied Science
Aston University, United Kingdom
{birdj1[1], a.ekart[3], d.faria[4]}@aston.ac.uk

[2]*Department of Computing*
Federal Center for Technological Education (CEFET-MG), Brazil
efwanner@decom.cefetmg.br

Abstract

Recent advances in the availability of computational resources allow for more sophisticated approaches to speech recognition than ever before. This study considers Artificial Neural Network and Hidden Markov Model methods of classification for Human Speech Recognition through Diphthong Vowel sounds in the English Phonetic Alphabet rather than the classical approach of the classification of whole words and phrases, with a specific focus on both single and multi-objective evolutionary optimisation of bioinspired classification methods. A set of audio clips are recorded by subjects from the United Kingdom and Mexico and the recordings are transformed into a static dataset of statistics by way of their Mel-Frequency Cepstral Coefficients (MFCC) at sliding window length of 200ms as well as a reshaped MFCC timeseries format for forecast-based models. An deep neural network with evolutionary optimised topology achieves 90.77% phoneme classification accuracy in comparison to the best HMM that achieves 86.23% accuracy with 150 hidden units, when only accuracy is considered in a single-objective optimisation approach. The obtained solutions are far more complex than the HMM taking around 248 seconds to train on powerful hardware versus 160 for the HMM. A multi-objective approach is explored due to this. In the multi-objective approaches of scalarisation presented, within which real-time resource usage is also considered towards solution fitness, far more optimal

April 20, 2020

solutions are produced which train far quicker than the forecast approach (69 seconds) with classification ability retained (86.73%). Weightings towards either maximising accuracy or reducing resource usage from 0.1 to 0.9 are suggested depending on the resources available, since many future IoT devices and autonomous robots may have limited access to cloud resources at a premium in comparison to the GPU used in this experiment.

Key words: Speech Recognition, Phoneme Classification, Applied Hyperheuristics, Multi-objective Evolutionary Computation

---

1. Introduction

Our modern life is influenced by technological innovations such as Intelligent Personal Assistants (IPAs). An Intelligent Personal Assistant is an intelligent software agent (Russell and Norvig, 2016), combining voice recognition, natural language processing, machine learning, and web semantics, that has been designed with the goal to assist people with basic tasks based on user commands by either text or voices. IPAs can be found in gadgets such as smartphones, tablets, smart watches, and smart speakers. They can, for example, check weather forecasts, remotely switch electrical devices on and off, answer questions, play music, place online shopping orders, provide real-time information, just to name a few tasks. Experts say that by 2021 there will be almost as many IPAs on the planet as people (Lahoual and Frejus, 2019), and more sectors of the economy such as from healthcare to private automotive industries will find uses for those technologies.

Although the most common application of IPAs has been filtering information from the internet, health, and educational applications can also be found in contemporary literature. Verlic, et al. (Verlic et al., 2005) presented iAPERAS, an expert system designed to aid in the lifestyles of non-professional athletes based on scientific research findings. Usually, the non-professional athletes rely on online information about training methods and nutritional recommendation and iAPERAS represented a more reliable alternative. In Wilges, et al., the authors tested a framework aiming to implement a set of resources for developing *Intelligent Learning Objects* (Wilges et al., 2007) which argued that the learning environment implemented accordingly is more flexible and adaptable than current approaches. An environment model coupled with an Animated Pedagogical Agent interacted with two agents of the systems. In Nunes, et al., the authors presented an Intelligent

Virtual Teaching Environment coupled with an Animated Pedagogical Agent aiming to educate children to preserve the environment (Nunes et al., 2002). The Animated Pedagogical Agent monitored, guided and individualised the learning process using a student model and teaching strategies.

Voice assistants, defined as *digital assistants that apply voice recognition as a point of control*, natural language processing, dictation, and synthesis of speech are all used to produce a specific service to a user. Virtual digital assistants are becoming increasingly accessible and available to the general public such as Google Home, Amazon Echo/Alexa, and Apple HomePod (López et al., 2017), as well as dictation for writing documents (Devine et al., 2000). If the home assistant is asked to perform a task, for example, setting an alarm for the next morning. The natural language signal produced by a microphone is converted into data through statistical extraction (Lerch, 2012), and following this, classification is performed (that is, *what did the user say?*). Finally, the answer is produced from a pre-defined database. More combinations for query within the database will improve the voice assistant system but this comes at a computational cost, due to the requirement of a more extensive search.

Home assistants are employed in different situations such as helping elderly people, people with special needs (Shpigelman et al., 2009), and improving educational processes. Furthermore, in rural areas in which access is limited by distance, isolation and lack of transportation, the usage of home assistants can provide medical evaluation and intervention, and enhance quality of life (Hudson and Cohen, 2006). Computer-mediated support interventions for people with special needs have been proved to provide socio-emotional support for those with needs (Shpigelman et al., 2009), so the home assistants may also help promote this inclusion. IPAs can also act as a teacher, learning facilitator or a student peer in collaborative setting, for education.

Speech recognition and voice dictation have become a viable and affordable technology. Systems converting the spoken word to text, faster than typing, have been used in many domains: from apps for mobile phones that allow you to compose text to medical transcription (Devine et al., 2000). Those technologies can create and edit documents, transcribe recordings into text, and use voice command to control day-by-day actions.

There are many language-dependent key issues in speech recognition despite the benefits of its usage. Speech recognition is a pattern recognition task in which a signal, or temporal statistics of the signal, are classified as a sequence of sounds, words, phrases, or sentences. In some phonetic languages, such as those found across some of Europe (for example, Spanish or Italian), speech-to-text is a relatively easy task since written sounds and spoken sounds often correspond in a one-to-one relationship. In the majority of languages, and in the case of this work, English, the conversion of speech to text is a much more complicated procedure due to the differing nature of written text to how it is spoken, something which in many cases is situationally dependent.

Furthermore, with completely global user bases, multitudes of accents and evolving dialect must be considered, especially non-native English speakers. This paper proposes an approach to speech recognition via the phonemic structure of the morphemes to be recognised, rather than classical word and phrase recognition techniques, which could lead to a speech recognition system that requires no retraining when new words are added to the dictionary. Additionally, the multi-objective scalarisation approach allows for the definition of a goal-based approach to the system, through the definition of scores given to the accuracy and resource usage metrics; to give an example of this, an IPA with cloud access to a powerful distributed computing framework could focus on a model which maximises accuracy due to abundance of technical resources, whereas a robot with access to only a CPU may find more success in maximising accuracy whilst minimising resource usage concurrently.

This work provides a large extension to previous study (Bird et al., 2019b). In this previous work, a preliminary approach which succesfully explored the application of the DEvo algorithm for single-objective hyper-heuristic evolutionary optimisation for hyperparameter selection of deep neural networks towards the classification of phonetic sounds, which make up some of the English language. In this present work, the algorithm is repurposed for distributed computing, and furthermore a multi-objective evolutionary approach through scalarisation is also explored and compared.

The main contributions of this work are as follows:

- The generation of a large, publicly-available diphthong vowel dataset sourced from subjects who are both native and non-native English

speakers (United Kingdom and Mexico)[1].

- A benchmark of the most common model used for contemporary voice recognition, the Hidden Markov Model, when training from a spoken set of phonemes.
- The search method for an optimal Artificial Neural Network topology for phoneme classification through an single-objective evolutionary hyperheuristic approach (*DEvo*).
- Extension of the DEvo algorithm towards scalarisation for multi-objective optimisation.
- A detailed comparison of models in terms of both their classification ability and computational resources required, both of which are considered important for real-time training.
- The final comparison of the produced models which puts forward the *DEvo* approach as the most accurate method of classifying spoken phonemes making up the English language.

The remainder of this work is organised as follows. Section 2 presents the motivation of this work and discusses the evolution of English Language and Phonetics. Section 3 discusses some classical approaches to speech recognition. Section 4 explains the fundamental concepts needed in this work. The proposed approach is presented in Section 5 and the results are shown and discussed in Section 6. A comprehensive comparison with some state-of-art approaches is presented in Section 7. Finally, section 8 concludes this work and discusses future direction.

2. Evolution of English Language and Phonetics

The most contemporary revolution of the English language arguably occurred with the invasion of the Norman (French) Duke William the Conqueror into Anglo-Saxon England and his subsequent reign over a united Normandy and England (Baugh and Cable, 1993).

The merging of Frankish and Anglo-Saxon culture, formed a language more reminiscent of modern English (Loyn, 2014). Words such as those in

[1]https://www.kaggle.com/birdy654/speech-recognition-dataset-england-and-mexico

Table 1, among many others which once had similar or identical meanings between the two languages, formed their own separately different meanings encapsulated within a more modern rendition of English. English language in spoken form retained most of its base phoneme structure, with few sounds being lost or gained. This shows that, even through the most extreme revolution of the English language where the entire language changed greatly, the phonetics tends to retain.

Table 1: Comparison of translations between 9th Century Old English and Old French

Old English	**Old French**
Roof	Ceiling
Room	Chamber
King	Royal
Cow	Beef
Swine	Pork

Table 2: The seven diphthong vowels in spoken English language in terms of their phonetic symbols and examples

Symbol	**English Example**
ɪə	N**ear**, **ear**, cl**ear**, f**ear**
eə	H**air**, th**ere**, c**are**
eɪ	F**a**ce, sp**a**ce, r**a**in, c**a**se, **ei**ght
ɔɪ	J**oy**, empl**oy**, t**oy**, **oy**ster.
aɪ	M**y**, s**i**ght, pr**i**de, k**i**nd, fl**igh**t
əʊ	N**o**, d**o**n't, st**o**ne, al**o**ne
aʊ	M**ou**th, h**ou**se, br**ow**n, c**ow**, **ou**t

Phonology is the study of the fundamental components of a spoken language as well as their relationships with one another (Fromkin et al., 2006). As previously mentioned, when it comes to English, spelling does not consistently represent the sound of language, for example: (i) the same sound may be represented by many letters or combination of letters (e.g. h<u>e</u> and p<u>eo</u>ple); (ii) the same letter may represent a variety of sounds (e.g. f<u>a</u>ther and m<u>a</u>ny); (iii) a combination of letters may represent a single sound (e.g. <u>sh</u>oot and <u>ch</u>aracter); (iv) a single letter may represent a combination of sounds (e.g. xero<u>x</u>); (v) some letters in a word may not be pronounced at all (e.g. s<u>w</u>ord

and psychology), and (vi) there may be no letter to represent a sound that occurs in a word (e.g. cute).

Most speech sounds are created by pushing air through the vocal cords. The phonetic alphabet of a language considers the biological source of the sound (*Labial, Dental, Alveolar, Post-alveolar, Palatal, Velar, or Glottal*) and a further biological affect upon the sound (*Nasal, Plosive, Fricative, or Approximant*), which overall make up every universally spoken sound found within a dialect, ie. all sounds enabled by the human vocal system. Consonants are sounds produced with some restriction or closure in the vocal tract, while vowels are classified by how high or low the tongue is, the position of the tongue inside the mouth and whether or not the lips are rounded. Diphthongs represent a sequence of two vowel sounds and require two muscular movements to produce. Table 2 shows each of the seven diphthong vowels in the spoken English language by way of their phonetic IPA symbols and examples of spoken words which contain them.

Based on evolution of language, it is theoretically possible through phoneme recognition to not only classify phonemes in speech, but to consider their temporal occurrence and transcribe the speech even with unseen words (*or as yet to be invented and defined*) by the model due to the retention of the word's phonetics. Phoneme errors seriously degrade the intelligibility of speech (Rogers and Dalby, 1996) and thus a classification approach based on phoneme recognition is an important step in improving speech recognition systems.

3. Related Work

Early research into the speech processing and recognition fields started in 1952 at Bell Labs, where single spoken digits were processed and classified (Juang and Rabiner, 2005). Statistical features of the power spectrum were observed towards classification of the spoken digits, power spectrum features are a notable step in modern voice recognition as one of the stages of Mel-frequency Cepstral Coefficient (MFCC) analysis (Muda et al., 2010). In this work, MFCC features are considered as static representation of the temporal wave-data gathered in the form of speech.

Many methods of statistical classification (Rabiner, 1989) have been attempted in speech recognition. For example, many of the state-of-the-art methods have employed Hidden Markov Models (HMM) to create speech

recognition models that are accurate enough for keyphrase communication with automated call-centre voices (Baum and Petrie, 1966; Huang et al., 1990). For example, those used when calling a bank in order to direct a customer's call to the correct department. Researchers noted that the success achieved was case-specific and that complex applications of the HMM for transcription of speech-text may not experience the same level of success.

More contemporaneous work focused on consonant, vowel, and limited word recognition. Studies found that Similar Pattern Analysis (SPA) could classify a very limited set of sounds with a 90% accuracy (Shannon et al., 1995), also noting the application in the domain of human-machine interaction in terms of aiding children with temporal processing disorders who have difficulty discerning sounds produced in a short time frame, i.e., those that occur often in natural speech. Research focusing on the classification of acoustic events such as keywords in speech achieved 80.79% accuracy using a Random Forest of decision trees (Xue and Zhao, 2008). A similar approach was employed with an accuracy of 81.5% for a set of 14 sound effects (Phan et al., 2015), though it is worth noting that the acoustic events in the last aforementioned study were not produced by humans. A particularly powerful method of machine learning approach for speech recognition, *Connectionist Speech Recognition*, was noted to be an ensemble fusion of predictions between a Multilayer Perceptron (MLP) and a Hidden Markov Model due to their largely statistical differences in prediction and yet high accuracies in terms of classifying audio data (Bourlard and Morgan, 2012). A recent work found that generalisation between language is difficult in (Pipiras et al., 2019), noting the scarcity of data available for Lithuanian speech recognition systems, researchers found high classification ability of spoken Lithuanian phonetics via a sequence-to-sequence approach through encoder-decoder models, achieving +99% over 10-fold cross-validation.

A related benchmarking study of the Random Forest classifier found that language based speech recognition became most optimal, and accurate, at a forest of 50 random decision trees all voting by average probability in a simple ensemble, the error rate of the multi-language corpus data for classification was found to be a relatively low 13.4% (Su et al., 2007). The Random Forest classification method was also used in classifier feature selection, from a dataset of acoustic audio features, to select an apt set of attributes for emotion classification from spoken audio data at approximately 70% accuracy over all test subject sets (who were divided by gender) (Rong et al., 2009).

US-based systems such as *DARPAs EARS* program and *IARPA Babel* operate a method of speech recognition with the extra step of specific-goal keyword segmentation and isolation (a cost-based machine learning approach), which are then used for security purposes by the National Security Agency (NSA) to autonomously detect high risk organisations via a computer system rather than the classical method of human wiretapping (Margulies, 2016).

Limited work on phoneme-based voice recognition has been performed. A bidirectional Long Short-term Memory neural network was tuned to an accuracy of 87.7% (Graves et al., 2013b) on a dataset of phonetic sounds. It is worth noting the usefulness of temporal-considerate machine learning techniques (inputs as batches/streams of data vectors). A limited dataset of the sounds *"B"*, *"D"*, and *"G"* was classified with an overall accuracy of 99.1% using a Time Delayed Neural Network, outperforming a Hidden Markov Model by 1.9% (Waibel et al., 1990). This study suggests the promising capabilities of a temporal-considerate neural network for speech recognition. However, the study was performed on a limited dataset that was not an accurate reflection of the multitude of phonemes found in human language, specifically in spoken English.

A Deep Learning approach through use of a Convolutional Neural Network (CNN) offered a preliminary solution to the spoken accent problem in speech recognition (Fang et al., 2019). The approach can derive a matrix which would be applied to the Mel-Frequency Cepstral Coefficients (MFCC) of a sound which would effectively attempt to mitigate differences in spoken accent by translating between them, with promising preliminary experiments resulting in success.

It is observed that all of the related works made use of temporal statistical analysis of soundwaves in order to create stationary data for classification, rather than attempting to simply classify the continuous sound. Furthermore, the most commonly observed method of generating this mathematical description is to analyse patterns found in short term Mel-Frequency Cepstral Coefficient (MFCC) data at 100-500ms. This is further explained in Section 4. It is also worth noting that a large majority of the studies discussed present results based on train-test split of the data, which is prone to overfitting (Tetko et al., 1995; Moore, 2001) and thus fail to generalise to unseen data (which is in itself the point of speech recognition). In this study, we perform 10-fold cross validation and present the mean accuracy over the folds as well as standard deviation of the results in order to avoid over-fitting

and better model out-of-sample data, which is important for generalisation and also for the genetic search itself, since we do not want the algorithm to simply search for network hyperparameters that best over-fit the validation data.

4. Background

4.1. Mel-Frequency Cepstral Coefficient

Due to the complexity, randomness, and non-stationary aspects of sound waves, classification of raw sound is very difficult. The method introduced in many studies is to apply a sliding time window to the wave data, and perform mathematical analysis in order to produce a dataset of audio-describing statistics. It is these statistics, then, that are classified as the spoken sound. In this work, the Mel-Frequency Cepstral Coefficients (MFCC) (Muda et al., 2010; Sahidullah and Saha, 2012) of the audio are extracted. To produce MFCC datasets, the following process is followed for each sliding time window:

1. The Fourier Transform (FT) of the time window data ω is derived via:
$$X(j\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt. \tag{1}$$
2. The powers from the FT are mapped to the Mel scale, the psychological scale of audible pitch (Stevens et al., 1937). This occurs through the use of a triangular temporal window.
3. The Mel-Frequency Cepstrum (MFC), or power spectrum of sound, is considered and logs of each of their powers are taken.
4. The derived Mel-log powers are treated as a signal, and a Discrete Cosine Transform (DCT) is measured. This is given as:
$$X_k = \sum_{n=0}^{N-1} x_n cos\left[\frac{\pi}{N}(n+\frac{1}{2})k\right] k = 0, ..., N-1. \tag{2}$$
where x is the array of length N, k is the index of the output coefficient being calculated, where N real numbers $x_0...x_{n-1}$ are transformed into the N real numbers $X_0...X_{n-1}$ by the formula.

The MFCCs are finally considered as the resultant amplitudes of the spectrum generated through the above process. MFCCs thus provide a detailed mathematical description on the behaviours of audio data and form a dataset of numerical attributes for use in learning processes.

4.2. Machine Learning Background

4.2.1. Hidden Markov Models

A Markov Chain is a model that describes a sequence and probability of events occurring based on those that have previously occured, that is, a branched and ordered sequence (Gagniuc, 2017). Hidden states within a Markov model describe a previously occurring data object (event) and thus predict the next event in the sequence, the number of hidden states required is therefore largely data dependent in terms of event length but also predictable event precursor length. A Hidden Markov Model's probability calculations and subsequent classification decision are given as follows:

$$Y = y(0), y(1), ...y(L-1), \tag{3}$$

where Y is the probability the sequence of length L occurring. Secondly,

$$P(Y) = \sum_{X} P(Y|X)P(X), \tag{4}$$

describes the probability of Y where the sum runs over all of the generated hidden node sequences, given as X:

$$X = x(0), x(1), ..., x(L-1). \tag{5}$$

The classification is finally chosen based on highest probability on previously studied data sequences within the hidden model, and is thus inherently Bayesian in nature (Bayes et al., 1763).

Due to the temporal nature of speech, it is worth noting that a Zoughi et al. (Zoughi et al., 2020) found success in performing an adaptive sliding window and Convolutional Neural Network approach to various datasets including the TIMIT phoneme classification dataset, where the window would adapt to the specific duration of the phonetic sound.

4.2.2. Artificial Neural Networks

An Artificial Neural network is a computational system of classification or regression inspired by the biological brain (Rosenblatt, 1961), in that a brain's neuron (nerve cell) takes input data to the dendrites and produces an output at nerve endings (Hopfield, 1984). A Multilayer Perceptron (MLP) is a form of Artificial Neural Network (ANN) which can be used to approximate (regression) or classify (in the case of classification) a data point given

training data. For an MLP, a structure of neurons are formed (input = attributes, output = classes) including varying interconnected hidden layers, which are user defined and data dependent. The so-called input layer takes as input the given attributes. The data are processed and finally an output prediction is given, which is either a class or real value. If there are more than one hidden layer, this constitutes a *deep neural network*.

Learning is performed for a defined period of time through the process of *backpropagation* (Bengio et al., 2015). Backpropagation is, in effect, a form of automatic differentiation in which errors are passed backwards from the final layer, to derive a gradient which is then used to calculate neuron weights within the network, dictating their activation and therefore the entire behaviour and decision making processes of the network, given input. Through this, a gradient descent optimisation algorithm is employed to derive neuron activation weight distribution by computing the gradient of the so-called loss function, which is the error rate. After this process of learning, a more optimal neural network is generated.

Errors for regression can be calculated in numerous ways, for example, using the Euclidean distance from the expected value. In classification, a measure of entropy is often used, ie. the level of randomness or predictability for the classification of a set P_j with solution s:

$$E(s) = -\sum_j p_j \times log(p_j). \tag{6}$$

Comparing the difference of two network's entropy gives the Information Gain (*relative entropy*). Kullback-Leibler (KL) divergence, or *information gain*, when a univariate probability distribution of a given attribute is compared to another (Kullback and Leibler, 1951). The calculation with the entropy algorithm in mind is thus simply given as:

$$InfoGain(T, a) = E(T) - E(T|a), \tag{7}$$

that is, with E of Equation 6 in mind, the observed change in entropy. For instances of original ruleset *H(T)* and comparative ruleset *H(T | a)*. A positive Information Gain denotes a lower error rate and thus arguably a better model[2].

[2]When a balanced dataset is considered.

4.2.3. Optimisation of ANN Topology

The optimal number of hidden layers and neurons (topology structure) for a given network is largely data dependent. A combinatorial optimisation problem occurs, and there is no simple linear algorithm to derive the optimal solution - there is *no free lunch* (Wolpert and Macready, 1997). Since fully connected neural networks produce a relatively small search space as connections themselves are assumed, an optimisation approach for the network topology is a realistic search problem.

EvoDeep (Martín et al., 2018) is an evolutionary algorithm used to derive a deep neural network for deep learning (eg. LSTM). Martin, et al. found Roulette Selection (random) for each population member to be best in the solution breeding process, therefore this method was chosen for this study's evolutionary search; in this study, each solution is in turn treated as $parent_1$ and a random second solution from $solutions - 1$ is chosen as $parent_2$.

Denser is an alternative novel method of evolutionary optimisation of an MLP (Assunçao et al., 2018). In addition to the number of hidden layers and neurons within fully connected neural networks, Denser also considers the type of layer itself. This increase of parameters results in a very complex search space, and is subsequently a very computationally intensive algorithm. However, it achieves very high accuracy results, for example 93.29% on the CIFAR-10 image recognition dataset.

Evolution of Neural Networks through Augmenting Topologies (NEAT) is an algorithm for the genetic improvement of neural networks which are not necessarily fully connected between layers (Stanley and Miikkulainen, 2002). The algorithm has been observed to be effective in learning from user input type problems, such as playing games, most notably for an evolving an ANN that learns to play Super Mario[3] in real time (Togelius et al., 2009).

The *Deep Evolutionary* or *DEvo* approach (Bird et al., 2019a), is an evolutionary search method focused on the tuning of the ANN topology of a fully connected, or dense, neural network. Specifically, the algorithm optimises both the number of hidden layers as well as their respective neuron counts with a single objective approach of classification accuracy, thus, equal weighting or distribution of classes are suggested. Diversity is promoted through offspring being strictly heterogeneous to the current population; a new offspring is generated if an example already exists before any fitness

[3] ©Nintendo

Table 3: Gender, Age, and Accent Locale of each of the test subjects

Gender	**Age**	**Accent Locale**
M	22	West Midlands, UK
F	19	West Midlands, UK
F	32	London, UK
M	24	Mexico City, MX
F	58	Mexico City, MX
M	23	Chihuahua, MX

measurement is taken. There are two original contributions of (1) a steadily increasing parameter maximum is suggested in order to promote early simple networks against their more complex but fitness-identical counterparts, as well as (2) a growing maximum population size in order to search the problem space more evenly over the entire simulation. The DEvo approach is extended towards multi-objective optimisation in this study in order to consider resource usage in addition to model accuracy, through the exploration of fitness scalars, as described in Section 5.

5. Method

In this section, the methodology of all experiments is described from data collection to generation of final results.

5.1. Data Collection and Attribute Generation

For recording an audio dataset, subjects were all asked to pronounce the sound as if they were speaking English although not all of the subjects were native English speakers. All of the seven diphthong vowels sounds were recorded ten times each by the six subjects as can be seen in Table 3. The three subjects from the United Kingdom are native English speakers whereas the Mexican subjects were native Spanish speakers but had a fluent proficiency in English. The resultant dataset of 420 individual sound clips were processed in order to remove silence and then produce an MFCC dataset by a sliding window of length 200ms. Ultimately, this produced a large dataset of 32,398 data objects for classification.

5.2. Machine Learning

The training and prediction process applied in this study can be summarised in Figure 1. Input parameters are considered to be the set of recorded

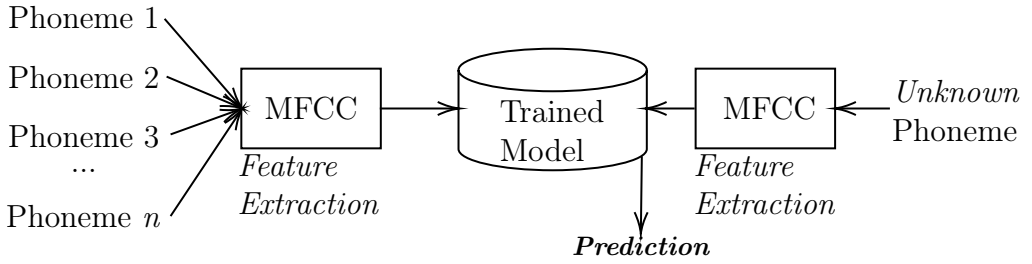

Figure 1: Description of training and prediction process applied in this study. Initial training happens to the left of the trained model where phonemes are used as data objects for learning and validated through 10-fold cross-validation; prediction of unknown phonemes from sound data occurs to the right of the model. (Bird et al., 2019b)

phonemes to train the selected model. Data is converted to a relational time-series for HMM whereas data is randomised for the MLP. A dataset is then generated from the phonemes recorded via statistical extraction by way of their Mel-frequency Cepstral Coefficients, which are then normalised. Machine learning models are trained and validated using 10-fold cross validation, and measured by their overall accuracy. The solution MLPs are given a standard 500 epochs of training time, learning rate of 0.3 and a momentum of 0.2 which were chosen manually based on initial exploration. Future work outlines the further optimisation of these parameters.

The chosen approach for the optimisation of ANN topology is the *DEvo* approach given in (Bird et al., 2019a), due to its proven effectiveness with flat datasets as well as temporal attributes extracted from wave-like data. Given the previous work, it was the finding that roulette selection in breeding is best for exploration of the problem space of ANN topology, and this approach is implemented. Evolutionary algorithms were run for 10 generations (this hyperparameter is introduced to satisfy *'while simulating'*) with a population of 5 (which increased by one per generation until 10), and experiments were repeated and recorded three times. For the multi-objective approach, experiments were repeated five times for each set of hyperparameters, thus giving a total of 15 experiments (providing distributions for non-parametric testing). The DEvo Process is shown in Algorithm 1[4]; random solutions are generated and tested, and then during simulation offspring are generated at

[4]Pseudo-code sourced from (Bird et al., 2019a)

```
Result: Array of best solutions at final generation
initialise Random solutions;
for Random solutions : rs do
    test accuracy of rs;
    set accuracy of rs;
end
set solutions = Random Solutions;
while Simulating do
    for Solutions : s do
        parent2 = roulette selected Solution;
        child = breed(s, parent2);
        test accuracy of child;
        set accuracy of child;
    end
    Sort Solutions best to worst;
    for Solutions : s do
        if s index > population size then
            delete s;
        end
    end
    increase maxPopulation by growth factor;
    increase maxNeurons by growth factor;
end
Return Solutions;
```

Algorithm 1: Evolutionary Algorithm for ANN optimisation

each generation and tested, the weakest solutions are culled. In the second experiment, three simulations of the same hyperparameters are run in which both accuracy and time are considered for a multi-objective problem. Scalarisation is introduced in order to explore multiple methods of fitness calculation:

$$\begin{aligned} \max F(s) &= \lambda_1 \frac{A(s)}{100} - \lambda_2 \frac{T(s)}{x}, \\ T &= \begin{cases} x, & \text{if } T > x \\ T, & \text{otherwise} \end{cases}, \end{aligned} \tag{8}$$

where the Function F of topology s is scored for its accuracy $A(s)$ on a

scale of 0.0...1.0, and for its time usage $T(s)$ on a scale of $0...x$, where x is a modified and values of time usage larger than x are kept at x. The selection of weight hyperparameters, λ_1 and λ_2 (negatively weighted) provide a data dependent scalarisation problem. Weights are introduced since the two metrics are vastly unequal in scale, classification accuracy is measured on a scale of $0.0-100.0$ while resource usage has no bounds and is often very large. A preliminary random search is executed prior to selection of normalisation in order to choose a reasonable candidate for parameter x.

The proposed approach is compared to a classical Hidden Markov Model. The HMMs are searched manually from 25 to 175 hidden units, at a step of 25. The upper limit of 175 is introduced as the next step, 200, failed due to the length of the data being considered. The best model is then used as a baseline comparison in both classification ability and resource usage.

In terms of hardware, the models are trained on a GTX980Ti Graphics Processing Unit. All training occurs via a Python 3.7 implementation of Keras with TensorFlow backend, with the software executing on a Windows 10 system isolated from any network. The Operating System is installed as fresh on a formatted drive with no unimportant background processes allowed, in order to prevent any interference of the measuring of time taken to train.

6. Preliminary Results

6.1. HMM Topology Selection

For choosing the best HMM topology, an approach of manual exploration was applied. Hidden Markov Models comprised of a topology of 25 to 175 hidden units were tested, at a step of 25, and were used to attempt to classify the whole dataset. Figure 2 shows the accuracy for the phoneme classification for each HMM tested. Results showed the HMM having 150 hidden units provided the best accuracy result (86.23%) for phoneme classification on this dataset. This model was then used as a baseline for comparison to the proposed approach. 200 Hidden Units extended beyond the majority of data series and thus the model could not be trained without error, therefore, 175 was the upper limit for benchmarking.

6.2. Single Objective Optimisation of Accuracy

When performing an evolutionary search of the Neural Network topology, the decision variables were the number of hidden layers ($[1, 5]$) and number of

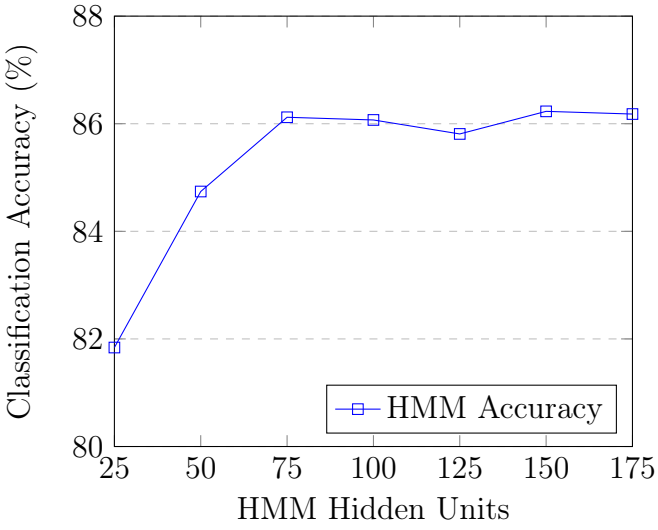

Figure 2: Benchmarking of HMM Hidden Units (Bird et al., 2019b)

neurons in each hidden layer ($[1, 100]$). In the single objective-optimisation, the accuracy was the function to be maximised. Table 4 shows the best accuracy of the strongest neural net solution at each generation of the evolutionary search. Due to the complexity of the search in comparison to the resources available, a relatively limited search was performed but to success. The best results for each search are shown in 4. Their differing areas of the search space suggest that an optimal solution is being converged upon rather than local minima. With more resources available, a more thorough search should be performed in an attempt to derive an even more effective ANN topology than the three layer network suggested.

The first DEvo experiment shows the optimal ANN topology for classification of phonemes is a deep neural network composed of three layers having 30, 7, and 29 neurons respectively. This ANN can accurately classify 88.84% of the MFCC time window data objects.

It is possible to see that an MLP with hyper-heuristically optimised topology has a high classification ability (88.84%) when it comes to the MFCC time windows of audio data in terms of spoken phonemes by both native and non-native English speakers when compared to a classical HMM (86.23%). The advantage of the optimised deep network over the HMM is greater than 2% in terms of simply accuracy alone. If efficiency in terms of time is also of concern, the computational resources required by each of the models for

training can be observed in Figure 4. The time spent in 10-fold Cross validation was measured for each final ANN topology for each simulation run for the evolutionary approach and for the HMM. The best model found was *S4* but it had the highest training time of 248.76. It can be observed that single layer models were competitive but took far fewer computational resources to train; Solution *S1* was the weakest suggestion by the evolutionary approach, and yet still outperformed the best HMM by 1.27% in terms of accuracy while training in less time, a successful reduction of 4.09 seconds compared to the Hidden Markov Model. Although this relatively short decrease in time is observed, an IoT device such as an autonomous robot with access to only a CPU rather than a GPU would experience a far bigger resource advantage.

We note that when the number of layers increases from one to three, the accuracy increases from 88.3% to 88.84% and time spent also increases from 180.34 seconds to 232.9 seconds. Also, comparing the obtained ANN with only one layers, layer size seems more important than depth. One important question that arises is the advantage of deep networks for the phoneme recognition problem using this dataset and thus, exploration of only one hidden layer of size n neurons was performed through the subsequent three simulations. Detailed results from this simulation (S4) are shown in Figure 3 where a single layer of 57 neurons gave the best result of 90.77% at a cost of 248.76 seconds in training time, by far the most computationally complex model produced.

6.3. Multi-Objective Optimisation of Accuracy and Resource Usage

In this section, the aforementioned multi-objective algorithm is explored in three contexts. In two, weights are biased towards each of the objective variables; $\lambda_1 = 0.1, \lambda_1 = 0.9$ and vice versa respectively. A third simulation is also executed, with equally weighted fitness scores, $\lambda_1 = 0.5, \lambda_1 = 0.5$.

Due to the consideration of the optimisation of resource usage, the search space in these experiments is expanded; the maximum number of layers allowed are set to 5, and the maximum number of neurons allowed are set to an increased cap of 2,048. This is due to the simulations having the goal of reduced time and thus relatively simpler solutions are to be expected more often than in single-objective optimisation.

An extremely complex model of the maximum parameters is benchmarked at five hidden layers of 2,048 neurons each. This simulation required 656.27s of computational resources and was introduced as the cap for time in the fitness function in equation 8. Therefore the fitness to a T greater than

Table 4: The best result at each generation for each of the simulations to optimise an MLP ANN

Experiment		**Generation**									
		1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
S1	***Layers***	1	3	4	1	1	3	3	2	1	1
	Neurons	2	3, 5, 9	5, 12, 4, 8	8	8	7, 15, 5	10, 9, 12	12, 10	12	21
	Accuracy (%)	53.67	66.9	70.31	83.37	83.37	81.26	84.3	85.6	87.73	87.5
S2	***Layers***	2	2	3	2	3	3	2	2	3	1
	Neurons	1, 6	19, 1	13, 2, 2	17, 19	19, 6, 10	19, 6, 10	19, 11	19, 7	19, 15, 22	25
	Accuracy (%)	36.9	53.51	78.45	86.81	86.32	86.32	86.88	87.41	87.86	88.3
S3	***Layers***	3	2	2	3	3	3	3	3	3	3
	Neurons	14, 7, 18	11, 27	12, 18	25, 15, 15	25, 15, 15	25, 15, 15	30, 7, 29	30, 7, 29	30, 7, 29	30, 7, 29
	Accuracy (%)	85.18	85.85	86.05	88.45	88.45	88.45	88.84	88.84	88.84	**88.84**

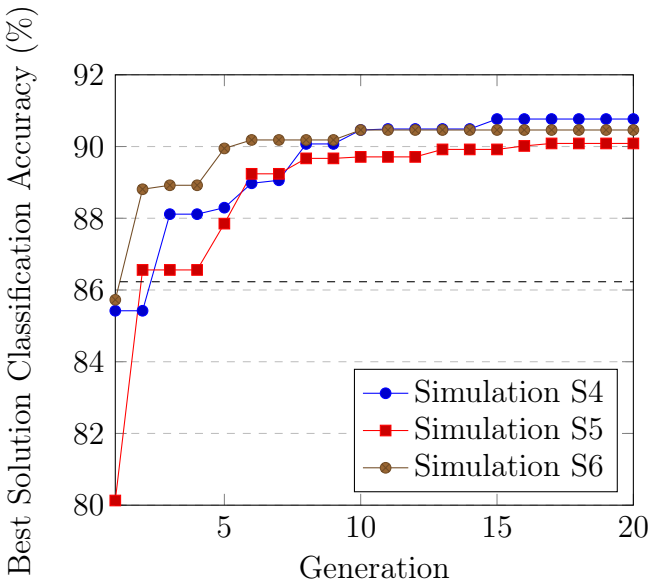

Figure 3: Single-objective Optimisation of Single Hidden Layer Neural Networks. The Dashed Line Denotes the HMM.

Table 5: Final Results for Simulations S4-S6 observed in Figure 3

Solution	**Hidden Layers (Neurons)**	**Accuracy (%)**
S4	1 (57)	**90.77** ±1.7
S5	1 (50)	90.09 ±2.8
S6	1 (51)	90.46 ±2.3

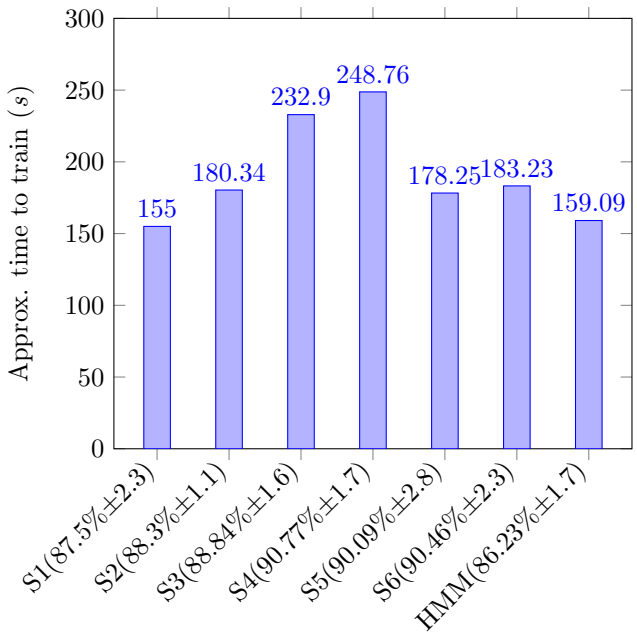

Figure 4: A Comparison of Model Training Time for Produced Models Post-search. S1-S3 are from Table 4 and S4-S6 are from Table 5.

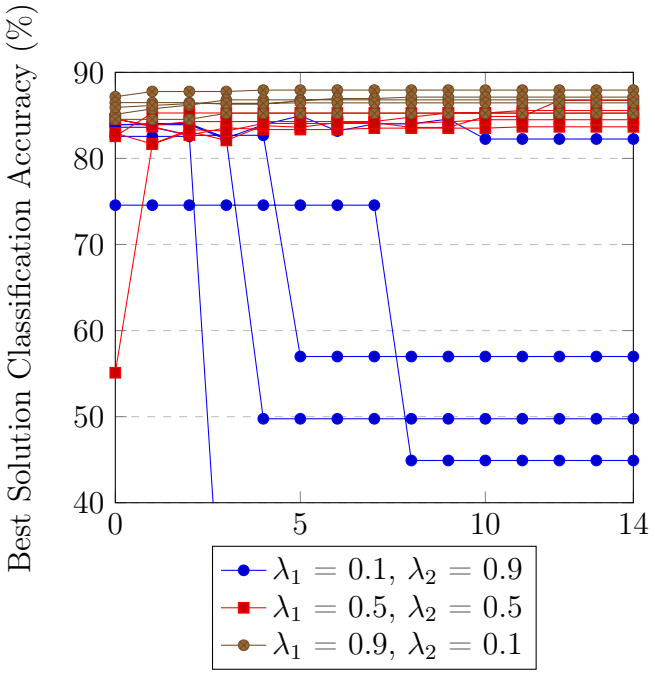

Figure 5: Evolution of Accuracy for Multi-objective Algorithms. A Value of 16.33 is Omitted for Purposes of Readability.

656.27 is simply λ_2, Equation 8 becomes:

$$\begin{aligned} \max F(s) &= \lambda_1 \frac{A}{100} - \lambda_2 \frac{T}{656.27}, \\ T &= \begin{cases} 656.27, & \text{if } T > 656.27 \\ T, & \text{otherwise} \end{cases} \end{aligned} \tag{9}$$

The final results produced can be observed in Table 6. Even with the lowest weighting towards resource usage, time to train was consistently below that of the Hidden Markov Model. In the initial two mult-objective simulations, patterns are as expected; minute differences seemingly contribute towards higher accuracy and lower complexity.In the third multi-objective simulation ($\lambda_1 = 0.1$ and $\lambda_2 = 0.9$) though, an interesting pattern occurs; the weighting towards lower resource usage did not completely perform as would logically be expected. It must be noted that due to the heavy weighting towards minimisation of training time, accuracy suffered heavily, as expected, going so far as to most often produce results that were far below an

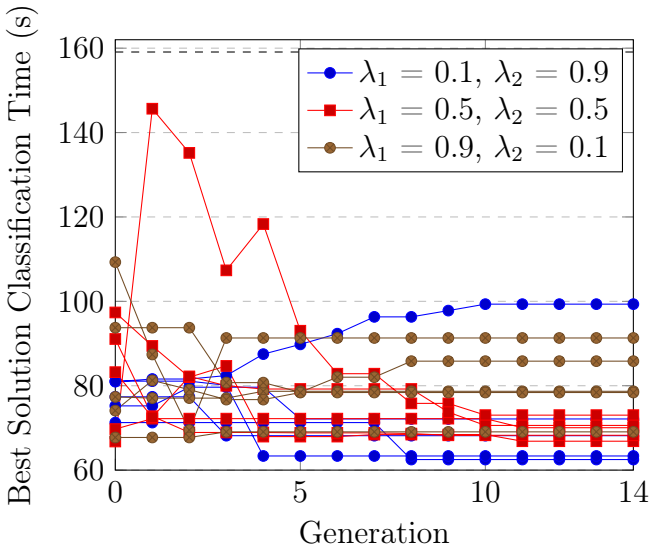

Figure 6: Evolution of Resource Usage for Multi-objective Algorithms. The Dashed Line Denotes the HMM.

Table 6: Comparison of the Results from the Final Parameters Selected by the Multi-objective Simulations. Note: best/worst accuracy are not necessarily of the same solutions as best/worst time and thus are not comparable

Scalars		**Accuracy (%)**			**Time (S)**		
$\boldsymbol{\lambda_1}$	$\boldsymbol{\lambda_2}$	***Mean***	***Best***	***Worst***	***Mean***	***Best***	***Worst***
0.1	0.9	49.75 ±2.3	82.2 ±1.75	16.33 ±3.2	99.33	73.1	62.51
0.5	0.5	85.15 ±1.6	86.73±1.6	83.67±1.6	69.75	66.85	73.03
0.9	0.1	86.7 ±1.3	87.94±1.4	85.25±1.6	80.66	69.07	91.32

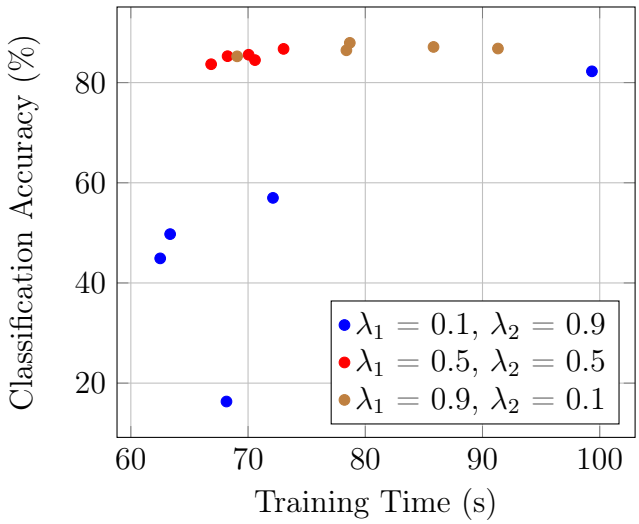

Figure 7: Final results presented by the multi-objective searches

acceptable classification ability - even though this was the case, the mean training time of these simulations were actually higher than those observed when $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$. Interestingly, two of the simulations had a similar spike in resource usage between generations 4 and 8, stabilising to a lower count within a generation of one another. In addition, the $0.5, 0.5$ simulation once experienced a single rising spike at generation 2 which quickly stabilised towards a lower measure soon afterwards. Figure 7 shows a Pareto frontier for the solutions, showing that the red $(0.5, 0.5)$ experience stability as well as strong results for maximising classification accuracy while minimising the training time required.

The fittest result from the $\lambda_1 = 0.9, \lambda_2 = 0.1$ simulations was a two-hidden layer neural network of $471, 1951$ neurons which achieved 87.93% accuracy after resource usage of 78.68 seconds. The fittest result from the $\lambda_1 = 0.5, \lambda_2 = 0.5$ simulations was a two-hidden layer network topology of $218, 1928$ neurons, achieving an 85.57% classification accuracy within 70.05 seconds of training. Finally, the fittest result from the $\lambda_1 = 0.1, \lambda_2 = 0.9$ simulations were two hidden layers of $765, 31$ neurons, which achieved an extremely low 16.33% classification accuracy within 63.35 seconds. As previously described, chosen solutions are dependent on hardware capabilities of the host; discarding $\lambda_1 = 0.1, \lambda_2 = 0.9$ due to weaker results, it is recommended that the weaker yet less complex networks ($\lambda_1 = 0.9, \lambda_2 = 0.1$)

Table 7: Results of the Nemenyi Test for the three sets of accuracy results achieved

	0	**1**	**2**	**3**	**4**
0	-1	0.819	0.9	0.9	0.9
1	0.819	-1	0.9	0.9	0.9
2	0.9	0.9	-1	0.9	0.9
3	0.9	0.9	0.9	-1	0.9
4	0.9	0.9	0.9	0.9	-1

are used for machines with no cloud access or distributed computing, such as an autonomous robot with a CPU (Foster et al., 2016; Qureshi et al., 2016) (since a CPU cannot distribute learning as these experiments did). The more complex networks that achieved higher accuracy at the cost of higher complexity could sensibly be used by a learning machine with access to distributed computing hardware such as a GPU (Steinkraus et al., 2005).

Upon performing the Friedman Test (Friedman, 1937) with an alpha level of 5%, the test statistic for accuracy was 8.4 with a p-value of 0.015, showing statistical difference between the distributions of results. The results of the Nemenyi Post-hoc test (Nemenyi, 1962) can be observed in Table 7.

7. Comparison with State-of-the-art

In this section, we compare our approach to the state-of-the-art in a related dataset. Unfortunately, no competitive datasets for phoneme classification from MFCC data exist in the field. Due to this, we opt for the subset of data extracted from the TIMIT Acoustic-Phonetic Continuous Speech Corpus (Garofolo, 1993) which can be found in (Hastie et al., 1995). The dataset provides a 5-class problem of spoken phonetic sounds from 50 male speakers. For each phoneme, the log-periodogram is calculated at length 256, resulting in a numeric representation of the sound (similar to MFCC).

Table 8 shows our approach is competitive with the state of the art when performed on the TIMIT subset. Our search presented a deep neural network of 580, 36, and 910 hidden neurons which scored 92.85% classification accuracy over 10-fold cross-validation. It is worth noting that the related studies performed a data split approach, and as such, our approach is less prone to overfitting. The average ROC area of this classifier was 0.99 and

Table 8: Comparison of Accuracy and Standard Deviation for the classification of the TIMIT Subset Dataset

Study	**Method**	**Accuracy (%)**	**Std. Dev.**
Cao and Fan (Cao and Fan, 2010)	KIRF	93.1	0.9
Ours (10-fold)	DEvo MLP	92.85	1.3
Cao and Fan (Cao and Fan, 2010)	NPCD/MPLSR	92.8	1.7
Cao and Fan (Cao and Fan, 2010)	NPCD/PCA	92.1	1.2
Cao and Fan (Cao and Fan, 2010)	MPLSR	91.1	1.7
Cao and Fan (Cao and Fan, 2010)	PDA/Ridge	91.1	1.6
Li and Ghosal (Li et al., 2018)	UMP	89.25	N/A
Li and Ghosal (Li et al., 2018)	MLO	85.25	N/A
Li and Ghosal (Li et al., 2018)	QDA	83.75	N/A
Ager et al. (Ager et al., 2013)	GMM	81.5	N/A
Li and Yu (Li and Yu, 2008)	FSDA	81.5	N/A
Li and Yu (Li and Yu, 2008)	FSVM	78	N/A

the F-measure was around 0.93.

8. Future Work and Conclusion

This study showed that hyper-heuristically optimising the topology of an Artificial Neural Network led to a high classification ability of the MFCC data from spoken phonetic sounds by both native and non-native English speakers. In addition to this, in comparison to the Hidden Markov Model, models that required fewer computational resources and yet still outperformed HMM were derived through a multi-objective algorithm. Further work should explore a more fine-tuned minimisation of resources (λ_2) since a value of 0.9 seemed to be too extreme and produced weak results, and thus further pairs weights should be explored towards this end.

Following the success of both the single and multi-objective approaches to hyperheuristic optimisation of phoneme classification, further MLP parameters could be considered as those to be optimised, such as activation, training time, momentum, and learning rate. In addition, further network architectures should be considered for optimisation in order to explore the abilities and effects, such as temporally-aware recurrence through RNN and Bi-directional LSTM which have shown promise in recent advances in speech recognition (Graves et al., 2013b,a) and CNN with Bi-directional LSTM (Passricha and Aggarwal, 2019). Further work should also consider the heuristic optimisation of HMM hidden units, in a one-dimensional problem space,

since this study focused on manual optimisation which was not exhaustive. Since the evolutionary method has shown to be successful, in future, other methods could be explored and compared such as Particle Swarm Optimisation or Ant Colony Optimisation, for example. Additionally, the dataset could be expanded beyond the limited 6-subject data gathered to explore the possibility of generalisation to a large dataset of phonetic utterances.

In terms of the ideal models produced, and with the post-construction of complete words, phrases, and sentences, a speech recognition system could further be produced without the need for retraining in future. That is, should English lexicon evolve, as it does often (in 2018, Merriam-Webster added 800 new words to their dictionary (Merriam-Webster, 2018)), speech recognition models would not require retraining; simply, these words would be constructed from already learnt phonetic sounds. Thus, speech recognition systems would then only be hampered by the evolution of phonetic structure in language; as was previously described, evolution of phonetic language occurs over great lengths of time, compared to which Machine Learning paradigms become obsolete and replaced far quicker.

To finally conclude, in this work, several evolutionarily optimised Neural Network topologies of varying classification ability and computational complexity were presented via both single and multi-objective approaches. A Hidden Markov model was fine-tuned by a brute-force based search producing seven different models, all of varying classification ability, with the strongest for classification being 150 hidden units. All suggested ANN topologies outperformed the Hidden Markov Model in the phoneme recognition problem within single-objective optimisation, whereas multi-objective optimisation presented many solutions that required fewer resources to train, and in many cases, lead to better classification ability also. For real time techniques such as lifelong learning of an autonomous machine, some of the less complex multi-objective solutions are suggested in situations such as the availability of only a single CPU, whereas, in a situation where resources are not at a premium, the single-objective solutions are suggested.

References

Ager, M., Cvetkovic, Z., and Sollich, P. (2013). Phoneme classification in high-dimensional linear feature domains. *Computing Research Repository*.

Assunçao, F., Lourenço, N., Machado, P., and Ribeiro, B. (2018). Denser: Deep evolutionary network structured representation. *arXiv preprint arXiv:1801.01563*.

Baugh, A. C. and Cable, T. (1993). *A history of the English language*. Routledge.

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563.

Bayes, T., Price, R., and Canton, J. (1763). An essay towards solving a problem in the doctrine of chances.

Bengio, Y., Goodfellow, I. J., and Courville, A. (2015). Deep learning. *Nature*, 521(7553):436–444.

Bird, J. J., Ekart, A., and Faria, D. R. (2019a). Evolutionary optimisation of fully connected artificial neural network topology. In *SAI Computing Conference 2019*. SAI.

Bird, J. J., Wanner, E., Ekart, A., and Faria, D. R. (2019b). Phoneme aware speech recognition through evolutionary optimisation. In *The Genetic and Evolutionary Computation Conference*, pages 362–363. GECCO.

Bourlard, H. A. and Morgan, N. (2012). *Connectionist speech recognition: a hybrid approach*, volume 247. Springer Science & Business Media.

Cao, J. and Fan, G. (2010). Signal classification using random forest with kernels. In *2010 Sixth Advanced International Conference on Telecommunications*, pages 191–195. IEEE.

Devine, E. G., Gaehde, S. A., and Curtis, A. C. (2000). Comparative evaluation of three continuous speech recognition software packages in the generation of medical reports. *Journal of the American Medical Informatics Association*, 7(5):462–468.

Fang, F., Wang, X., Yamagishi, J., and Echizen, I. (2019). Audiovisual speaker conversion: jointly and simultaneously transforming facial expression and acoustic characteristics. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6795–6799. IEEE.

Foster, M. E., Alami, R., Gestranius, O., Lemon, O., Niemelä, M., Odobez, J.-M., and Pandey, A. K. (2016). The mummer project: Engaging human-robot interaction in real-world public spaces. In *International Conference on Social Robotics*, pages 753–763. Springer.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701.

Fromkin, V. A., Rodman, R., and Hyams, N. (2006). *An Introduction to Language*. Cengage.

Gagniuc, P. A. (2017). *Markov Chains: From Theory to Implementation and Experimentation*. John Wiley & Sons.

Garofolo, J. S. (1993). Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium, 1993*.

Graves, A., Jaitly, N., and Mohamed, A.-R. (2013a). Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE workshop on automatic speech recognition and understanding*, pages 273–278. IEEE.

Graves, A., Mohamed, A.-R., and Hinton, G. (2013b). Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE.

Hastie, T., Buja, A., and Tibshirani, R. (1995). Penalized discriminant analysis. *The Annals of Statistics*, pages 73–102.

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the national academy of sciences*, 81(10):3088–3092.

Huang, X. D., Ariki, Y., and Jack, M. A. (1990). Hidden markov models for speech recognition.

Hudson, D. L. and Cohen, M. E. (2006). Intelligent agent model for remote support of rural healthcare for the elderly. In *28th Annual International Conference of the IEEE Engineering in Medicine and Biology Science, EMBS 06*, pages 6332–6335. IEEE.

Juang, B.-H. and Rabiner, L. R. (2005). 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.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Lahoual, D. and Frejus, M. (2019). When users assist the voice assistants: From supervision to failure resolution. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, page CS08. ACM.

Lerch, A. (2012). *An introduction to audio content analysis: Applications in signal processing and music informatics*. Wiley-IEEE Press.

Li, B. and Yu, Q. (2008). Classification of functional data: A segmentation approach. *Computational Statistics & Data Analysis*, 52(10):4790–4800.

Li, X. et al. (2018). Bayesian classification and change point detection for functional data.

López, G., Quesada, L., and Guerrero, L. A. (2017). Alexa vs. siri vs. cortana vs. google assistant: a comparison of speech-based natural user interfaces. In *International Conference on Applied Human Factors and Ergonomics*, pages 241–250. Springer.

Loyn, H. R. (2014). *Anglo Saxon England and the Norman Conquest*. Routledge.

Margulies, P. (2016). Surveillance by algorithm: The nsa, computerized intelligence collection, and human rights. *Fla. L. Rev.*, 68:1045.

Martín, A., Lara-Cabrera, R., Fuentes-Hurtado, F., Naranjo, V., and Camacho, D. (2018). Evodeep: A new evolutionary approach for automatic deep neural networks parametrisation. *Journal of Parallel and Distributed Computing*, 117:180–191.

Merriam-Webster (2018). New dictionary words.

Moore, A. W. (2001). Cross-validation for detecting and preventing overfitting. *School of Computer Science Carneigie Mellon University*.

Muda, L., Begam, M., and Elamvazuthi, I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *arXiv preprint arXiv:1003.4083*.

Nemenyi, P. (1962). Distribution-free multiple comparisons. In *Biometrics*, volume 18, page 263. International Biometric Soc 1441 I ST, NW, SUITE 700, WASHINGTON, DC 20005-2210.

Nunes, M., Dihl, L., Fraga, L., Woszezenki, C. R., Oliveira, L., Fransisco, D., Machado, G., Nogueira, C., and da Gloria Notargiacomo, M. (2002). Animated pedagogical agent in the intelligent virtual teaching environment. *Interactive Educational MUltimedia*, (4):53–61.

Passricha, V. and Aggarwal, R. K. (2019). A hybrid of deep cnn and bidirectional lstm for automatic speech recognition. *Journal of Intelligent Systems*, 29(1):1261–1274.

Phan, H., Maaß, M., Mazur, R., and Mertins, A. (2015). Random regression forests for acoustic event detection and classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):20–31.

Pipiras, L., Maskeliūnas, R., and Damaševičius, R. (2019). Lithuanian speech recognition using purely phonetic deep learning. *Computers*, 8(4):76.

Qureshi, A. H., Nakamura, Y., Yoshikawa, Y., and Ishiguro, H. (2016). Robot gains social intelligence through multimodal deep reinforcement learning. In *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, pages 745–751. IEEE.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Rogers, C. L. and Dalby, J. M. (1996). Prediction of foreign-accented speech intelligibility from segmental contrast measures. *Journal of the Acoustical Society of America*, 96(5).

Rong, J., Li, G., and Chen, Y.-P. P. (2009). Acoustic feature selection for automatic emotion recognition from speech. *Information processing & management*, 45(3):315–328.

Rosenblatt, F. (1961). Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, CORNELL AERONAUTICAL LAB INC BUFFALO NY.

Russell, S. J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

Sahidullah, M. and Saha, G. (2012). Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition. *Speech Communication*, 54(4):543–565.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270(5234):303–304.

Shpigelman, C., Weiss, P. L., and Reiter, S. (2009). e-empowerment of young adults with special needs behind the computer screen i am not disable. In *2009 Virtual Rehabilitation International Conference*, pages 65–69. IEEE.

Stanley, K. O. and Miikkulainen, R. (2002). Evolving neural networks through augmenting topologies. *Evolutionary computation*, 10(2):99–127.

Steinkraus, D., Buck, I., and Simard, P. (2005). Using gpus for machine learning algorithms. In *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, pages 1115–1120. IEEE.

Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190.

Su, Y., Jelinek, F., and Khudanpur, S. (2007). Large-scale random forest language models for speech recognition. In *Eighth Annual Conference of the International Speech Communication Association*.

Tetko, I. V., Livingstone, D. J., and Luik, A. I. (1995). Neural network studies. 1. comparison of overfitting and overtraining. *Journal of chemical information and computer sciences*, 35(5):826–833.

Togelius, J., Karakovskiy, S., Koutník, J., and Schmidhuber, J. (2009). Super mario evolution. In *Computational Intelligence and Games, 2009. CIG 2009. IEEE Symposium on*, pages 156–161. IEEE.

Verlic, M., Zorman, M., and Mertik, M. (2005). iaperas - intelligent athlete's personal assistant. In *18th IEEE Symposium on Computer-Based Medical Systems*, pages 134–138. IEEE.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. J. (1990). Phoneme recognition using time-delay neural networks. In *Readings in speech recognition*, pages 393–404. Elsevier.

Wilges, B., Mateus, G., Silveira, R. A., and Nassar, S. (2007). An animated pedagogical agent as a learning management system manipulating intelligent learning objects. In *7th IEEE International Conference on Advanced Learning Technologies, ICALT 2007*, pages 186–188. IEEE.

Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82.

Xue, J. and Zhao, Y. (2008). Random forests of phonetic decision trees for acoustic modeling in conversational speech recognition. *IEEE transactions on audio, speech, and language processing*, 16(3):519–528.

Zoughi, T., Homayounpour, M. M., and Deypir, M. (2020). Adaptive windows multiple deep residual networks for speech recognition. *Expert Systems with Applications*, 139:112840.