# Phoneme Aware Speech Recognition through Evolutionary Optimisation

Jordan J. Bird
Aston University, Birmingham, UK
birdj1@aston.ac.uk

Elizabeth Wanner
Aston University, Birmingham, UK
e.wanner@aston.ac.uk

Anikó Ekárt
Aston University, Birmingham, UK
a.ekart@aston.ac.uk

Diego R. Faria
Aston University, Birmingham, UK
d.faria@aston.ac.uk

## ABSTRACT

Phoneme awareness provides the path to high resolution speech recognition to overcome the difficulties of classical word recognition. Here we present the results of a preliminary study on Artificial Neural Network (ANN) and Hidden Markov Model (HMM) methods of classification for Human Speech Recognition through Diphthong Vowel sounds in the English Phonetic Alphabet, with a specific focus on evolutionary optimisation of bio-inspired classification methods. A set of audio clips are recorded by subjects from the United Kingdom and Mexico. For each recording, the data were pre-processed, using Mel-Frequency Cepstral Coefficients (MFCC) at a sliding window of 200ms per data object, as well as a further MFCC timeseries format for forecast-based models, to produce the dataset. We found that an evolutionary optimised deep neural network achieves 90.77% phoneme classification accuracy as opposed to the best HMM of 150 hidden units achieving 86.23% accuracy. Many of the evolutionary solutions take substantially longer to train than the HMM, however one solution scoring 87.5% (+1.27%) requires fewer resources than the HMM.

## CCS CONCEPTS

• **Computing methodologies → Speech recognition**; Cooperation and coordination;

## KEYWORDS

Computational Linguistics, Phoneme Awareness, Speech Recognition, Evolutionary Optimisation, Artificial Neural Networks

**ACM Reference Format:**
Jordan J. Bird, Elizabeth Wanner, Anikó Ekárt, and Diego R. Faria. 2019. Phoneme Aware Speech Recognition through Evolutionary Optimisation. In *Proceedings of the Genetic and Evolutionary Computation Conference 2019 (GECCO '19)*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3319619.3321951

## 1 INTRODUCTION

In the modern day, voice assistants apply voice recognition as a point of input. Speech synthesis and natural language processing are used to produce a specific service via a particular application. Enabled devices are becoming increasingly available in the home. Smart Home Assistants can help users in different situations such as aiding elderly people since *the older an individual is, the more a long-term care is needed*, people with special needs, and improving educational processes. When a user asks a home assistant to perform a task, such as checking the news, natural language audio signal is first converted into digital data. This is then transcribed by software and compared with a database to find a suitable response. Expanding the number of queries composing the database improves voice recognition at the cost of increasing the response time through the more extensive search that is required.

Despite the variety of possible applications and the benefits of usage, there are several language-dependent key issues in speed recognition. Speech recognition is a multilevel pattern recognition task in which the acoustic signals are analysed and structured into a hierarchy of words and phrases. In some languages, such as Finnish, Italian and Spanish, the speech-to-text conversion is simple because written text almost correspond to its pronunciation, and vice-versa. For most other languages, especially English, the conversion is more complicated. The study of fundamental components of language formed by a system of sounds and their relationships is known as *phonology* [3]. Spelling does not consistently represent the sound of language. In 1888, the *International Phonetic Alphabet* (IPA) was created to have a system in which there was a one-to-one correspondence between each sound in language and each phonetic symbol. Pronunciation of foreign words with a local dialect replaces the natural phonetic structure. It is known that typical pronunciation for a given language differs for speakers of a different native language. As an example, the second syllable in the French word "Monsieur" has its natural post-glottal sound replaced with a post-alveolar "sure" sound, and its phonetics /*"məsjø"*/ become /*"mə'sjə"*/.

## 2 THE SPEECH RECOGNITION PROBLEM

The earliest form of speech recognition began in 1952 in the form of single spoken digits in research performed at Bell Labs [4], p.67. The experiment followed the observation of statistical features of the audio power spectrum, which is still considered today as an effective step of statistical extraction for modern voice recognition [5] (encompassed within Mel-Frequency Cepstral Coefficients

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

*GECCO '19, July 13–17, 2019, Prague, Czech Republic*

© 2019 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-6748-6/19/07…$15.00
https://doi.org/10.1145/3319619.3321951

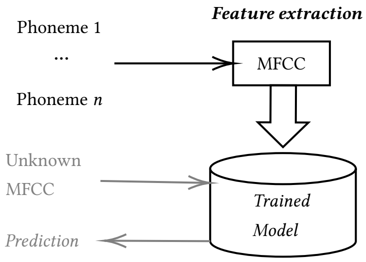

**Figure 1: Diagram of the feature extraction, training, and prediction process**

analysis). Some of the most prominent results on human speech recognition were obtained using statistical methods such as simple Bayesian networks in the form of *Hidden Markov Models* [1]. Due to the complex nature of audio, classification of raw signals is very difficult. Time windowing is introduced and statistical extraction is performed to generate a mathematical description of the data contained within each window via the Mel-Frequency Cepstral Coefficient (MFCC) of the data. The optimal number of hidden layers and neurons (topology structure) for a given network is largely data dependent. For this optimisation problem there is no simple linear algorithm to derive the optimal solution. In this experiment, the Deep Evolutionary (DEvo) algorithm is used to optimise the ANN topology[2].

## 3 METHOD

Audio recordings of diphthong vowels were gathered from test subjects. Subjects were all required to pronounce the sounds as if they were speaking English regardless of their native language. There were six subjects, three from Mexico (Mexico City and Chihuahua) and three from the UK (London and West Midlands). Those from the UK were native English speakers, whereas those from Mexico were native Spanish speakers with fluency in English. Subjects were recorded speaking the seven phonemes, ten times each, producing a dataset of 420 individual clips. Any silence was removed from the clips as to not consider it in rule generation. The subjects who did not speak English as their first language spoke the phonemes as if they were speaking in fluent English and thus their natural accent was retained. A sliding window of 200ms was introduced to extract the MFCC data from audio. Overlapping occurred once at a factor of 0.5, meaning that the first window *w1* measured 0-200ms and the next window *w2* measured 100-300ms and so on. The proposed approach is compared to a classical HMM. The HMMs were searched empirically from 25 to 175 hidden units, at a step of 25. The DEvo algorithms were run for 10 generations with a population of 10, and experiments were repeated and recorded three times. A simple diagram of the training and prediction process can be observed in Fig. 1.

## 4 PRELIMINARY RESULTS

Via exploration of the HMM topology we found that the best model of 86.18% classification accuracy contained 175 hidden units. For the evolutionary training of the Deep MLP ANN, the number of layers ([1, 5]) and the number of neurons in each layer - if the layer exists ([1, 100]) were used as decision variables and the classification accuracy of the MLP ANN was the objective function to be maximised. Three further experiments of only one maximum MLP layer were also performed. Figure 2 shows the accuracy and required resources for all cases. All MLPs were superior to the HMM, but only one required fewer computational resources. Exhaustive search of the full dataset on one hidden layer led to a solution of 91.3% accuracy, with 100 neurons. This solution is relatively close to the DEvo result. The process took 30.85 hours to complete.

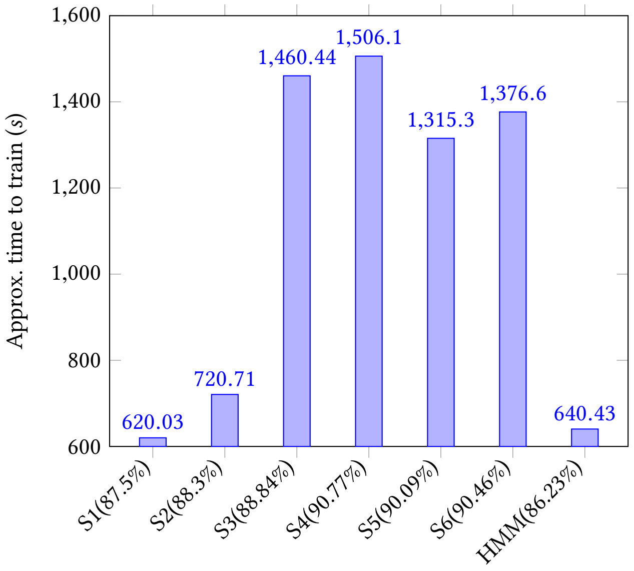

**Figure 2: A Comparison of Model Training Time for Produced Models Post-search. S1-S3 have one hidden layer and S4-S6 have up to five.**

## 5 CONCLUSION

This preliminary study showed that an MLP with hyper-heuristically optimised topology can achieve high classification ability, using MFCC time windows of audio data of phonemes spoken by both native and non-native English speakers. We envisage that with the construction of complete words, phrases, and sentences, speech could be used as an interface for human-robot communication, with a text representation of the spoken language undergoing further analysis such as sentiment or emotional classification.

## REFERENCES

[1] Leonard E Baum and Ted Petrie. 1966. Statistical inference for probabilistic functions of finite state Markov chains. *The annals of mathematical statistics* 37, 6 (1966), 1554–1563.

[2] Jordan J. Bird, Diego R. Faria, Luis J. Manso, Aniko Ekart, and Christopher D. Buckingham. 2019. A Deep Evolutionary Approach to Bioinspired Classifier Optimisation for Brain-Machine Interaction. *Complexity* 2019 (2019).

[3] V.A. Fromkin, R. Rodman, and N. Hyams. 2006. *An Introduction to Language.*

[4] Biing-Hwang Juang and Lawrence R Rabiner. 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 (2005).

[5] Lindasalwa Muda, Mumtaj Begam, and Irraivan Elamvazuthi. 2010. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *arXiv preprint arXiv:1003.4083* (2010).