Accent and gender recognition from English language speech and audio using signal processing and deep learning

Accent and Gender Recognition from English Language Speech and Audio Using Signal Processing and Deep Learning

Jagjeevan Singh Shergill[1], Chandresh Pravin[1], and Varun Ojha[1]

Department of Computer Science
University of Reading, Reading, UK
{j.s.shergill,c.pravin,v.k.ojha}@reading.ac.uk

Abstract. This research is concerned with taking user input in the form of speech data to classify and then predict which region of the United Kingdom the user is from and their gender. This research was conducted on regional accents, data preprocessing, Fourier transforms, and deep learning modeling. Due to lack of publicly available datasets for this type of research, a dataset was created from scratch (12 regions with a 1:1 gender ratio). In this paper, we propose modeling the human's voice accent and voice gender recognition as a classification task. We used a deep convolution neural network, and experimentally developed an architecture that maximized the classification accuracy of the mentioned tasks simultaneously. We also tested the model on publicly available spoken digit detests. We find that the gender classification is relatively easier to predict with high accuracy than the accent in our proposed multi-class classification model. Accent classification was found difficult because of the regional accent's overlapping that prevents it from being classified with high accuracy.

Keywords: Speech Recognition · Speech Classification · Deep Learning · Convolution Neural Network · Signal Processing

1 Introduction

In the information age, voice recognition is becoming increasingly more important. Users are rapidly adopting voice recognition systems, whether it be within their handheld devices or home assistants such as Amazon's Alexa. Apple released Siri (part of the iPhone 4S release), illustrating how voice recognition's complexity and potential had evolved since its early beginnings and demonstrated its future potential in real-world applications [16]. Since then, multiple voice recognition technologies have been incorporated into many systems, most notably home assistants such as Google Home and Amazon's Alexa.

Machine Learning for Speech (English Language Accent) Classification is a task of generalizing different accents in natural speech audio datasets. To begin, we may first define what an *accent* refers to linguistically; according to [12], an

accent is defined as a manner of pronunciation that is characteristic of a particular location, nation, or individual. This paper investigates if machine learning algorithms can be utilized to analyze speech metadata in order to develop a model able to classify different accents according to geographical location and gender.

This paper aims to leverage machine learning techniques to predict the regional accent (of the United Kingdom) and the gender of a speaker. We firstly begin by identifying regions from around the United Kingdom to be considered by the learning model and compiling a dataset containing audio speech waveforms from people native to each respective region. The condition of selecting individuals for collating audio files is that they are native of their respective regions and exemplify the local dialect and accent. This indeed requires prior research to identify the general linguistic accents based on a given region. The audio samples are then sourced from publicly available sources, i.e., videos found on YouTube. Twelve different regions within the united kingdom were chosen to construct the dataset with a 1:1 ratio of male to female.

The audio files' raw representations were then transformed into the frequency domain using the Short Term Fourier Transform (STFT) and subsequently reshaped into a spectrogram [2]. The individual spectral components provide a breakdown of the audio signal's constituent frequency components as a function of time, thus revealing information about the signal magnitude and phase at different frequencies. The time-frequency spectrogram was then used as an input to a two-dimensional convolution neural network (CNN) to classify gender and linguistic accents from the audio signals. Following are the main focus of this work:

- We compiled a custom-made speech dataset consisting 12 regions of the UK with 1:1 gender ratio.
- We investigated a methodology consisting of the pre-processing of audio data and deep learning classification modeling.
- We experimentally develop a deep learning architecture that maximized classification accuracy. In this case, model was able to classify gender with higher accuracy than accent.

Section 2 reports existing literature, and Section 3 describes the speech classification modeling, including dataset and deep learning model description. The results are discussed in Section 4, followed by conclusions in Section 5.

2 Relevant Literature

Overviews of accents respective to regions in the United Kingdom can be seen in [10, 4]. Understanding the nuances of languages is an important aspect of developing an effective machine learning model and understanding how to preprocess the audio data at hand. Within local dialects and accents, we must also consider the issue of the emergence of convoluted accents, where the increased accessibility to travel around the country has meant that in several regions,

where once a traditional dialect from the region was common, has now been phased out by a more diluted form of the accent [8].

An example of recent advancement in the field of speech and audio classification is Google's Translatotron, a state-of-the-art model that makes use of a spectrogram decoder in the audio translation process for speech classification and language translation [9]. The use of spectrograms for accent classification tasks have indeed been proposed by Ai et al. [1] in their accent recognition and language translation model. The recently proposed approach to machine translation termed Neural Machine Translation aims to address possible performance limitations from the current paradigm of building a translation model from many sub-components, each of which is tuned separately, by building a single deep network that can be tuned to maximize the translation performance [3].

In contrast, with regard to gender classification, we define the model as having two genders to classify; male and female. To generalize, the variation between the genders is thought of as a difference in pitch, the average pitch of a male voice is lower than that of a female voice during a natural speech [15]. The natural speech pitch of an individual is influenced by a variety of factors and may not always follow generalization made regarding males having a lower pitch and females having a higher pitch in comparison, this is investigated by Meena et al. [13] using fuzzy logic and neural networks for gender classification. The pitch frequency is related to anatomy, and the "high" and "low" pitch of the genders may lie with a Gaussian distribution around a mean, such as 110Hz for male speech and 200Hz for female speech. Children and young adults may have a variable pitch as anatomy changes over time [17].

3 Speech Classification modeling

3.1 Problem Statement

The purpose of this study is to establish if one can take speech data from a user and predict which regional accent that user belongs to within the United Kingdom, and also what gender they are. Mathematically, for a given input speech signal $\mathbf{x}_i = (x_0, x_1, \ldots, x_t)$ paired with a target class y_i for $i = 1, 2, 3, \ldots, n$ training examples, where $y = \{\text{Male Voice}, \text{Female Voice}\}$ for speech to gender classification or $y = \{a_{1-F}, a_{1-M}, a_{2-F}, a_{2-M}, \ldots, a_{12-F}, a_{12-M}\}$ for 12 regional accents (each have Female and Male) classifications, the aim is to minimize a classification error rate $\mathcal{L}$ to train a classifier $f(X, \mathbf{w})$ as per the following:

$$\mathcal{L}(X, f(X, \mathbf{w})) = \sum_{i=0}^{n} (f(x_i, \mathbf{w}) \neq y_i),$$

where $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$, $Y = [y_1, y_2, \ldots, y_n]$, and $\mathbf{w}$ is the class parameters. It is hoped that a trained classifier $f(X, \mathbf{w})$ with low error rate $\mathcal{L}$ would correctly classify a user's accent and gender.

3.2 English Language Accent Dataset

We focus on classifying 12 regions in the UK English language accents (Fig. 1). The regions of interest are Northern Ireland, Scotland, Wales, North East England, North West England, Yorkshire and Humber, East Midlands, West Midlands, East of England, Greater London, South East England, South West England.

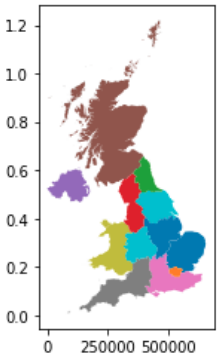

Fig. 1. 12 regions of the UK in our accent classification model.

Indeed, the regions themselves have differing accents and dialects within the area, particularly in areas that share boundaries with other regions, however, for this study, we focus on the general accents from the given region. The dataset was complied depended on the availability of audio data from various regions across the UK and thus, the number of regions, and therefore accents evaluated in this study was limited to 12, although in reality, the number of accents within the UK can be categorized in greater detail [5]. The dataset was compiled based on the availability of audio data containing 12 accents belonging to 12 regions of the UK. However, due to overlapping between the region, accents, it limits the strong and distinct 12 categories. The full dataset is available at following repository[1].

3.3 Feature Extraction Algorithms

Data Pre-processing. Raw audio files were pre-processed prior to being used as training data or prediction. For consistency, the audio samples were resized to be the same length (10 Seconds) and also be sampled at the same frequency (8kHz), finally to be transformed to the frequency domain to result in a spectrogram. Much of the task of preprocessing was carried out using the *librosa*[2] package on python for audio manipulation and analysis.

The audio files sourced from YouTube were approximately 30 seconds long. These waveforms were subsequently split into three segments of length ten seconds, whilst disregarding any additional audio data for regularity. Any audio files less than 10 seconds long after splitting into three segments, indicating that the original audio file was less than 30 seconds long in total were removed.

Finally, all audio files were resampled at 8kHz to reduce the size of the audio files, standardize all audio input and thus ensure the dataset is uniform. The resampling process still retains the frequencies related to speech, whilst having the additional benefit of removing any unwanted noise outside of the

[1] doi: https://doi.org/10.5281/zenodo.4071925
[2] http://librosa.org/doc/latest/index.html

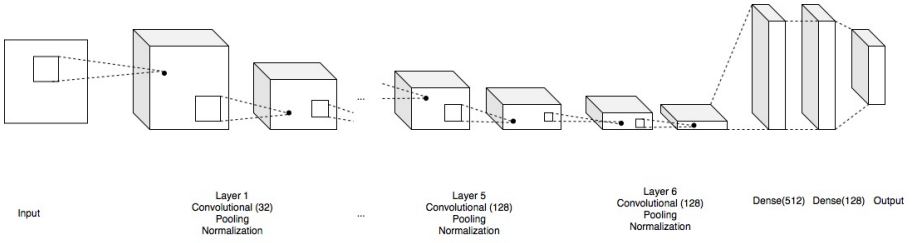

Fig. 2. Proposed DNN architecture used for accent classification algorithm of two-dimensional FFT spectrograms of audio data.

human hearing frequency range. The combination of truncating at 10 seconds and resampling at a common rate ensures all audio samples have the same length, which is particularly useful for applying to the proposed CNN model.

Spectrogram Transformation. An audio waveform is often represented in the time domain, this is a signal amplitude against time mapping and can be intuitively understood by the reader since audible sound corresponds to peaks on the graph and silence to a relatively flat line with a 0 amplitude. To transform audio from the time domain to the frequency domain, the audio is broken into millisecond samples, and for each sample, an STFT operation is computed [11]. The STFT for that sample is plotted as a vertical line as its relative power. The aggregated set of STFT plots for all the millisecond samples that constitute the audio samples is formed, thus resulting in a spectrogram.

An STFT operation isolates frequencies to a short time interval. This technique allows the characteristic decomposition of an audio signal. Transformation of a time-series signal into the Fourier domain provides a representation of the audio signal as a distinct set of frequencies, which consequently improves the ability to generalize the gender and accent of the speaker using machine learning algorithms, as shown by [1, 15]. There have been various instances of CNN models being used for two-dimensional time-series image classification [6, 7]; this work continues from this basis in transforming the audio signal into the Fourier domain and generating a two-dimensional spectrogram to be used as the input of a CNN model. Further details of the spectrogram generation can be seen in Section 4.

3.4 Deep learning architecture search

Fig. 2 shows the CNN model architecture used for the accent and gender classification task experimentally developed in this study (see Fig. 3 and Table 1). The corresponding model parameters can be seen in Table. 2, which details the number, types, configuration, shape, and size of the layers within the model. The model proposed in this paper is a two-dimensional CNN that used the audio file spectrogram images as the input to the model.

3.5 Experimental Setup

The classification model outlined in this paper consists of 12 accent classes and 2 gender classes, resulting in 24 prediction classes. The model was fine-tuned using an iterative experiential procured for neural network design outlined in [14]. Moreover, a decision tree is utilized to arrive at the optimum model.

The initial step in the process was transforming the one-dimensional, time series audio files, an example shown by Fig. 4, into a two-dimensional spectrogram shown in Fig. 5. The spectrogram array was then used as input to train the CNN model shown in Fig. 2 that was arrived at by using decision tree shown in Fig. 3 and decisions shown in Table 1. Model details is in Table 2.

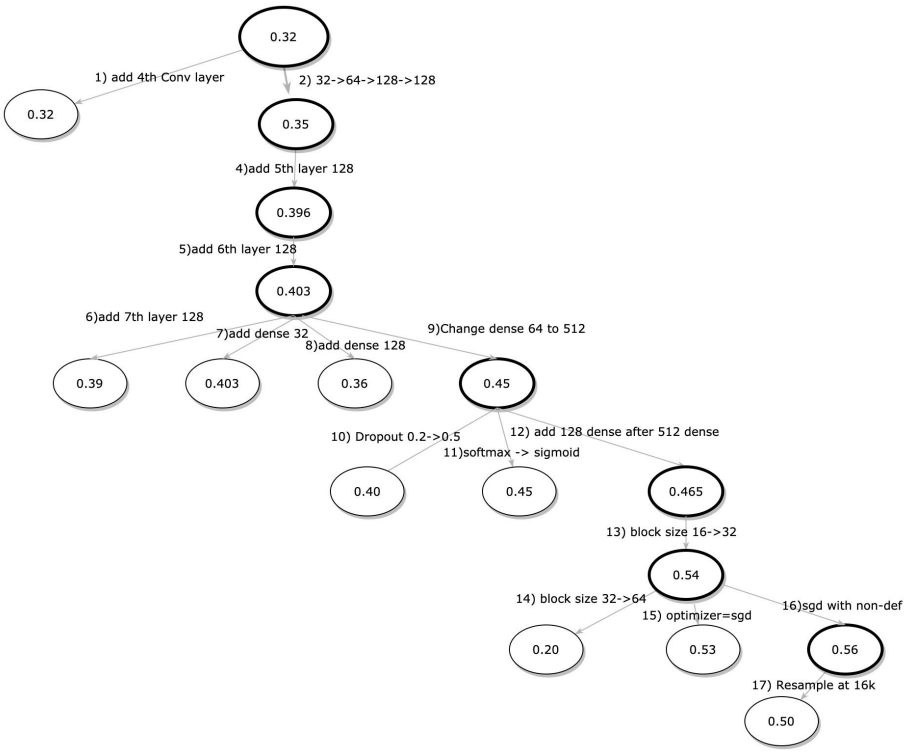

Fig. 3. Experiment decisions for constructing the CNN model.

4 Results and Discussion

The dataset used was firstly pre-processed and then used to build and test a classification model, resulting in **65% accent prediction accuracy** and **85%**

Table 1. CNN model tuning decisions

	Description	Prev	New	Keep
1	Add 4th convolution/maxpooling layer, keeping filter=32	0.32	0.32	N
2	Create 4 convolution layers, with filter sizes (32,64,128,128)	0.32	0.35	Y
3	Add 5th convolution layer, with filter=128	0.35	0.396	Y
4	Add 6th convolution layer, with filter=128	0.396	0.403	Y
5	Add 7th convolution layer, with filter=128	0.403	0.39	N
6	Add an extra Dense layer, with units=32	0.403	0.403	N
7	Add an extra Dense layer, with units=128	0.403	0.36	N
8	Change existing Dense layer, with units change from 64 to 512	0.403	0.45	Y
9	Change existing Dropout layer, with rate change from 0.2 to 0.5	0.45	0.40	N
10	Change existing output Dense layer, with activation function change from softmax to sigmoid	0.45	0.45	N
11	Add an extra Dense layer, with units=512	0.45	0.465	N
12	Change batch size for model learning from 16 to 32	0.465	0.54	Y
13	Change batch size for model learning from 32 to 64	0.54	0.20	N
14	Change the compile model from Adam to SGD, with default SGD parameters	0.54	0.53	N
15	Change the SGD parameters from default to custom	0.54	0.56	Y
14	Change the audio resampling rate from 8kHz to 16kHz	0.56	0.50	N

model accuracy for gender classification (Fig 6). The model was tested on Google speech commands datasets [18] that gave **90% on gender classification**.

Since the model was trained on British English accent datasets, it performs poorly on generic accent classification (accuracy 15%). It should be noted that the gender classification does not suffer from a great loss of accuracy when applied to the Google digit dataset, as the model resulted in **90% accuracy**. Indeed, even with some accent misclassification, the false positives still match the correct gender.

For example, a classification of a female audio file presenting an accent from the east of England has over 75% misclassification rate. However, 100% false positives still classify correctly as female. This indicates that the model has the ability to generalize gender from an audio signal, irrespective of the region of accent that the person exhibits. Thus, further reaffirming the understanding that the feature set extracted for gender classification shows less variability [15] in comparison to accents.

The study has demonstrated some success in the classification of audio by the UK region and significant success in identifying the gender of a speaker. This has justified the approach detailed in Section 3. However, the success of this work has been limited by the size of the custom-made dataset used. It has been shown, by using a different, larger dataset being applied to the same model outlined in this study, that the classification accuracy increases with respect to classifying gender from an audio signal.

Table 2. CNN model hyperparameter configuration

Block	Layer	Configuration	Layer shape	Layer size
1	Convolution	filters=32, kernel=3, strides=1	(1025,157,32)	5,149,600
	Max-pooling	pool_size=2 strides=2	513.79.32	1,296,864
	Normalization	filters=32 kernel=3 strides=1	(513,79,32)	1,296,864
2	Convolution	filters=64 kernel=3 strides=1	513.79.64	2,593,728
	Max-pooling	pool_size=2 strides=2	257.40.64	657,920
	Normalization	-	257.40.64	657,920
3	Convolution	filters=128 kernel=3 strides=1	257.40.128	1,315,840
	Max-pooling	pool_size=2 strides=2	129.20.128	330,240
	Normalization	-	129.20.128	330,240
4	Convolution	filters=128 kernel=3 strides=1	129.20.128	330,240
	Max-pooling	pool_size=2 strides=2	65.10.128	83,200
	Normalization	-	65.10.128	83,200
5	Convolution	filters=128 kernel=3 strides=1	65.10.128	83,200
	Max-pooling	pool_size=2 strides=2	33.5.128	21,120
	Normalization	-	33.5.128	21,120
6	Convolution	filters=128 kernel=3 strides=1	33.5.128	21,120
	Max-pooling	pool_size=2 strides=2	17.3.128	6,528
	Normalization	-	17.3.128	6,528
7	Flatten	-	6528	6528
8	Dense	units=512	512	512
	Normalization	-	64	64
9	Dense	units=128	128	128
	Normalization	-	128	128
	Dropout	rate=0.2	128	128
10	Prediction	units=24	24	24

5 Conclusion

This paper presents a comprehensive study on speech classification, where this study offers a custom-made a dataset of UK regional accent and gender speech dataset. A comprehensive experimental evaluation of a range of convolutional neural network (CNN) model architectures were studies to improve both accent and gender classification form speech data. The model gives very high accuracy on gender classification on both our custom-made dataset and publicly available Google speech command dataset. On the custom-made dataset on which the CNN model was trained does provide a comparatively low accuracy on accent classification compared to gender classification. This signifies that the model's performance for gender classification is easier to predict with smaller dataset training. However, for accent classification, it requires a much larger dataset to obtain very high accuracy. Accent classify may also suffer from overlapping accent in our custom-made dataset as the regions were chosen are close to one.

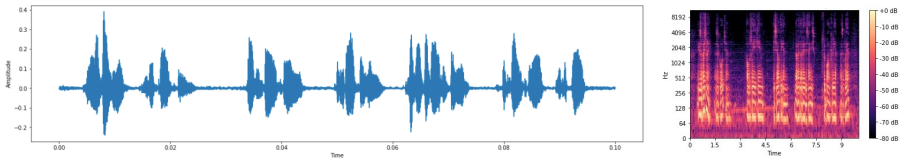

Fig. 4. Example of an audio signal of a native English-speaking person from the east midland region class. Sample audio from the complied dataset with length 10 seconds, sampled at a frequency of 8 kHz.

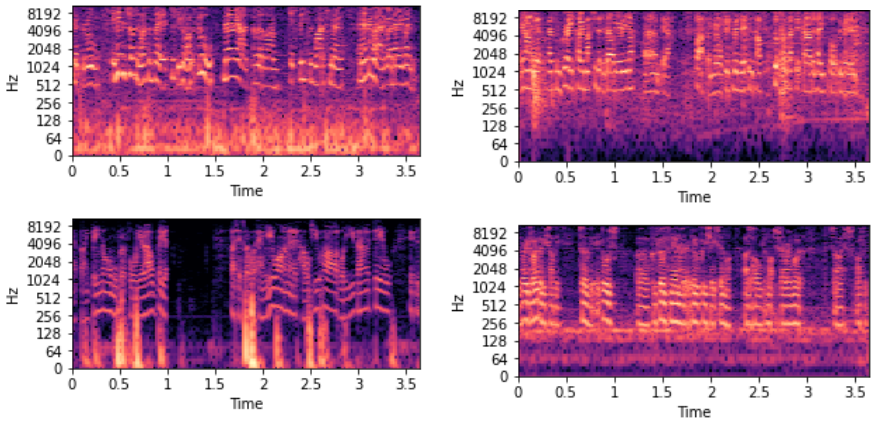

Fig. 5. Two-dimensional spectrogram images of audio signals. The spectrogram shows a frequency against time plot with the relative power of the frequencies presents being shown in dB. A DNN aims classify variations on these spectrogram images.

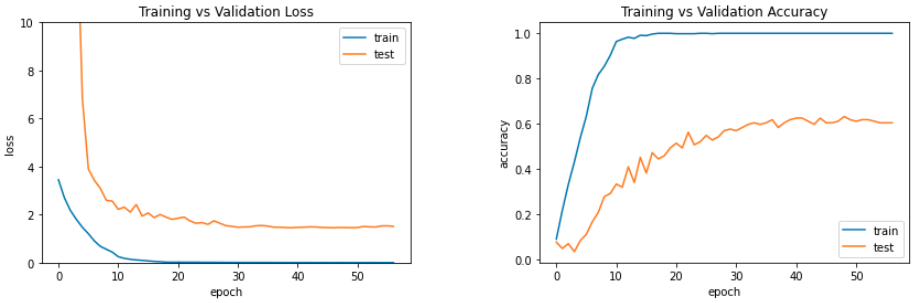

Fig. 6. *Left:* graph showing training (shown by a blue curve) and validation (shown by an orange curve) set loss functions of the model. *Right:* graph showing the training (shown in blue curve) and validation (shown in orange curve) set model accuracy.

In our future work, we aim to investigate audio feature extraction techniques and deep learning model architecture improvement.

References

1. Ai, L., Jeng, S.Y., Beigi, H.: A new approach to accent recognition and conversion for mandarin chinese. arXiv preprint arXiv:2008.03359 (2020)
2. Allen, J.: Short term spectral analysis, synthesis, and modification by discrete fourier transform. IEEE Transactions on Acoustics, Speech, and Signal Processing **25**(3), 235–238 (1977)
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate (2014)
4. Chen, T., Huang, C., Chang, E., Wang, J.: Automatic accent identification using gaussian mixture models. In: IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU'01. pp. 343–346. IEEE (2001)
5. Coupland, N., Bishop, H.: Ideologised values for british accents 1. Journal of Sociolinguistics **11**(1), 74–93 (2007)
6. Fawaz, H.I., Forestier, G., Weber, J., Idoumghar, L., Muller, P.A.: Deep learning for time series classification: a review. Data Mining and Knowledge Discovery **33**(4), 917–963 (2019)
7. Hatami, N., Gavet, Y., Debayle, J.: Classification of time-series images using deep convolutional neural networks. In: Verikas, A., Radeva, P., Nikolaev, D., Zhou, J. (eds.) Tenth International Conference on Machine Vision (ICMV 2017). vol. 10696, pp. 242 – 249. International Society for Optics and Photonics, SPIE (2018)
8. Hughes, A., Trudgill, P., Watt, D.: English Accents and Dialects: An Introduction to Social and Regional Varieties of English in the British Isles, Fifth Edition. The English Language Series, Taylor & Francis (2013)
9. Jia, Y., Weiss, R.J., Biadsy, F., Macherey, W., Johnson, M., Chen, Z., Wu, Y.: Direct speech-to-speech translation with a sequence-to-sequence model. Interspeech 2019 (Sep 2019)
10. Kat, L.W., Fung, P.: Fast accent identification and accented speech recognition. In: 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. vol. 1, pp. 221–224. IEEE (1999)
11. Krawczyk, M., Gerkmann, T.: Stft phase reconstruction in voiced speech for an improved single-channel speech enhancement. IEEE/ACM Transactions on Audio, Speech, and Language Processing **22**(12), 1931–1940 (2014)
12. McKean, E.: The New Oxford American Dictionary. No. v. 2 in The New Oxford American Dictionary (2005)
13. Meena, K., Subramaniam, K.R., Gomathy, M.: Gender classification in speech recognition using fuzzy logic and neural network. Int. Arab J. Inf. Technol. **10**(5), 477–485 (2013)
14. Ojha, V.K., Abraham, A., Snášel, V.: Metaheuristic design of feedforward neural networks: A review of two decades of research. Engineering Applications of Artificial Intelligence **60**, 97 – 116 (2017)
15. Rai, P., Khanna, P.: Gender classification techniques: A review. In: Wyld, D.C., Zizka, J., Nagamalai, D. (eds.) Advances in Computer Science, Engineering & Applications. pp. 51–59. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
16. Sadun, E., Sande, S.: Talking to Siri: Learning the Language of Apple's Intelligent Assistant. Pearson Education (2013)
17. Sigmund, M.: Gender distinction using short segments of speech signal. International Journal of Computer Science and Network Security **8**(10), 159–162 (2008)
18. Warden, P.: Speech commands: A dataset for limited-vocabulary speech recognition. arXiv preprint arXiv:1804.03209 (2018)