 doi:10.20944/preprints201903.0047.v1

Article

DGR: Deep Gender Recognition of Human Speech

Rami S. Alkhawaldeh

[1] Department of Computer Information Systems, The University of Jordan, Aqaba, 77110, Jordan; r.alkhawaldeh@ju.edu.jo

Abstract: The speech entailed in human voice comprises essentially para-linguistic information used in many voice-recognition applications. Gender voice-recognition is considered one of the pivotal parts to be detected from a given voice, a task that involves certain complications. In order to distinguish gender from a voice signal, a set of techniques have been employed to determine relevant features to be utilized for building a model from a training set. This model is useful for determining the gender (i.e, male or female) from a voice signal. The contributions are involved in two folds: (i) providing analysis information about well-known voice signal features using a prominent dataset, (ii) studying various machine learning models of different theoretical families to classify the voice gender, and (iii) using three prominent feature selection algorithms to find promisingly optimal features for improving classification models. Experimental results show the importance of sub-features over others, which are vital for enhancing the efficiency of classification models performance. Experimentation reveals that the best recall value is equal to 99.97%; 99.7% of two models of Deep Learning (DL) and Support Vector Machine (SVM) and with feature selection the best recall value is 100% for SVM techniques.

Keywords: Gender Recognition; Speech Signal; Deep Learning; Evolutionary Search; PSO search; Wolf Search

1. Introduction

The voice of human speech is an effective communication method consisted of unique semantic linguistic and para-linguistic features such as gender, age, language, accent and emotional state. The sound waves consisting human voice are unique among all creatures producing sound since every single wave carries different frequency. Identifying human gender based on voice has been a challenging task for voice and sound analysts who deploy numerous applications including: (i) effective advertising and marketing strategies in Customer Relationship Management (CRM) systems which depend on gender interoperability such as user interface style as well as preferences of words and colours; (ii) investigating criminal voice in crime scenarios; (iii) Enhancing Human-Computer Interaction (HCI) systems especially dialogue systems by customizing services that rely on gender voice and also improving the level of user satisfaction. Due to the importance of identifying gender through voice-recognition, the human voice should be converted from analogue to digital form to extract useful features and then to construct classification models. The robustness and effectiveness of classifiers are determined by the quality of features that depend on a training set employing Machine Learning (ML) techniques. Therefore, eliciting voice features plays a vital role in improving the efficiency of classifiers since the human voice is liable for non-useful features. Research on improving the efficiency of voice classifiers is copious, particularly studying the process of extracting efficient features from voice including identifying the linguistic content of a speech signal components and disposing of non-useful content such as background noise.

There are a set of features used for recognizing the voice genders. Among the most common features utilized for voice-gender recognition are Mel-scaled power spectrogram; Mel-Frequency Cepstral Coefficients (MFCC); power spectrogram chroma (Chroma); spectral contrast (Contrast) and tonal centroid features (Tonnetz). By getting the extracted features combined with the gender label as a form of a training set, ML techniques are used to build a high-quality model for recognizing the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0047.v1

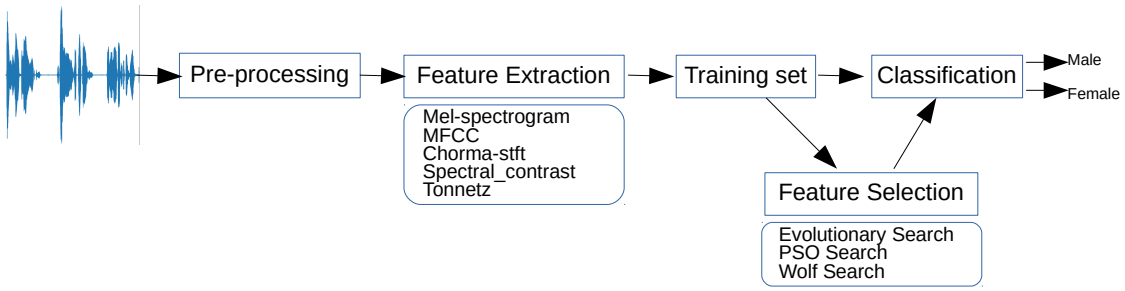


Figure 1. General gender recognition framework.

voice gender as shown in Figure 1. In particular, each classification technique is used to build a set of hypothesis models and selects the most optimal one. This model classifies the unknown voice label by receiving the voice features and categorizing the voice gender.

A multitude corpus of research has been conducted to address the efficiency of voice classifiers aiming to enhance the accuracy of programs being used. The authors in [1] used two level classifiers (pitch frequency and GMM classifier) to recognize speaker gender on TIDIGITS dataset with achievement reached 98.65%. The authors in [2] used four classifiers including GMM, MultiLayer Perception (MLP), Vector Quantization (VQ) and Learning vector quantization (LVQ) to analyze voices taken from IviE corpus. They managed to achieve a 96.4% success rate. The authors in [3] combined the estimated voice acoustic level of five different methods into one score level. The results were obtained on using aGender dataset for gender category of 81.7% success rates. The authors in [4] proposed a system for identifying speakers using a fusion score of seven subsystems where the feature vectors are MFCC, PLP and prosodic on three different classifiers that are GMM, SVM, and GMM-SV based SVM combined at score level. The classification success rate on gender identification using aGender database is 90.4%. The authors in [5] used two classifiers; SVM and Decision Tree (DT) with MFCCs feature on a private corpus identifying gender voice. The overall accuracies using MFCC-SVM and MFCC-DT for gender classification were 93.16% and 91.45%, respectively. The authors in [6] showed an accuracy of 100% gender discrimination on TIMIT and KEL corpus.

The most efficient classifiers and feature extractors of superior accuracy on gender voice-recognition include Deep Neural Networks (DNNs) and Convolutional neural networks. The authors in [7] proposed an adequate technique to enhance the MFCC features and then adjust the weights between DNNs layers. These improved MFCC features are evaluated on DNN and I-Vector classifiers where the overall accuracies are 58.98% and 56.13%, respectively. The authors in [8] compared two classification techniques were compared (DNN and SVM) using single and combined feature vectors for robust sound classifications. The results showed a better performance for DNN technique as a robust and low sensing to noise approach.

In this work, a novel approach is presented for characterizing the voice gender using a different set of features along with different ML algorithms from various families. These features showed their effectiveness in extracting the voice patterns, hence, categorizing the gender. The contributions are demonstrated as follows:

- Studying a set of voice features and examining their effects as possible suitable features for gender classification techniques.
  RQ1: *To what extent is selecting voice signal features useful on building machine learning classifiers?*
- Using different ML techniques of various families on recognizing the speech gender from the extracted and efficient features.
  RQ2: *What is the performance of various ML models on gender voice-recognition applications?*
- Evaluating well-known natural feature selection techniques on choosing the most optimal features.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0047.v1

RQ3: *To what extent using natural feature selection evaluators is useful for enhancing the performance of ML learning techniques?*

The structure of the paper is organized as follows: Section 2 discusses the proposed approach in classifying a given voice gender. It provides a detailed discussion of the classifiers of voice recognition including: the phases of preprocessing, extracting features, ML techniques, evaluation metrics, and feature extraction methods. Section 3 presents the experimental settings and answers the research questions in each subsection. The conclusions and future work are discussed and summarized in Section 4. Notably, the related work is presented in the introduction Section 1.

2. Deep Gender Recognition (DGR)

The proposed methodology for speech gender classification includes a set of stages as briefly discussed. The stages start by converting the voice, from its abstract representation, into a consistent form in order to extract the relevant features. Then the relevant features are selected as inputs for building a classifier model for recognizing the gender of a human voice. In addition, a DL models is being built to automatically extract useful features and feed them into a fully connected Artificial Neural Network (ANN) for classification. However, here, a set of process of extracting features for other models rather than DL and classification techniques are summarized as follows:
Voice Pre-Processing: A transmitted voice is inevitably vulnerable to noise interference and voice attenuation that needs a pre-processing process to purify it for feature extraction. This phase shows a set of steps as follows:

- **A/D signal Conversion:** is used to convert the given voice from analogue to digital signal by common sampling and quantization techniques [9]. The A/D conversion formulates the signal in an understandable form by machine for easy manipulation.
- **Pre-emphasis process:** Due to attenuation at high-frequency segments of the voice signal, there is a necessary need to use a pre-emphasis filter. The pre-emphasis filter flattens the signal (or speech) waveforms. The process filters low-frequency interference- especially power frequency interference at low-frequency segments- and emphases the high-frequency portions in order to produce a high-pass filter to carry out spectral analysis interference. This process occurs after A/D conversion by the first-order digital pre-emphasis filter equation [10]:

$$H(z) = 1 - \mu z^{-1} \tag{1}$$

  where z represents filter, μ is pre-emphasis filter coefficient with the value ranging commonly between [0.9,1].
- **Frame Blocking and Hamming window:** The frame blocking is a process of handling the filtered digital signal into a number of N small frame segments of adjacent frames separated by M ($M < N$). The process of Hamming window minimizes speech signal discontinuities before and after each frame within the window frame. This method is popularly used in MFCC before Mel frequency warping step where Mel scales is calculated. The analytical representation of the Hamming window is given by:

$$w(n) = 0.54 - 4.4cos(\frac{2\pi n}{N-1}) \qquad 0 \leq n \leq N-1 \tag{2}$$

  where $w(n)$ is the window operation, n is the number of each individual sample and N is the total number of the speech samples [10].
- **Fast Fourier Transform (FFT):** The FFT algorithm is in general used for estimating the Discrete Fourier Transform (DFT) of any sequence, or its inverse form. In speech voice signal, the FFT converts each frame of those N samples from the time domain signal into a form of frequency

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0047.v1

domain [11]. The FFT is considered as a computationally efficient implementation of the DFT method, which is defined on the set of N samples $\{x_n\}$ as follows:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{j2\pi kn}{N}}, \quad k = 0, 1, 2, \ldots, N-1 \tag{3}$$

X_k is a complex number considered as its absolute value (frequency magnitudes or modulus). $\{x_k\}$ is resulting sequence interpreting as: the positive frequencies $0 \leq f < \frac{1}{2}F_s$ correspond to the values $0 \leq n \leq \frac{1}{2}N - 1$, while the negative frequencies $-\frac{1}{2}F_s < f < 0$ correspond to $\frac{1}{2}N + 1 \leq n \leq N - 1$. F_s represents the sampling rate. The results obtained are called frequency spectrum of the voice signal.

Extracting features from digital voice signals: There are a set of relevant features that could be inferred from the voice signal. Hence a pre-processing phase is needed to prepare the speech signals as input for a set of feature extraction techniques. These set of features and a voice gender as a label represent the training set for building a classifier model in order to recognize the voice speech gender. For visualization, Figure 2 shows a voice sample, which British English female voice and its features. The features used in this paper are:

- **Mel-spectrogram:** Computes a Mel-scaled power spectrogram coefficients. An object of Mel-spectrogram type represents an acoustic time-frequency representation of sound as an example shown in Figure 2(b). The power spectral density $P(f, t)$ is sampled into a number of points around equally spaced times t_i and frequencies f_j (on a Mel frequency scale). The Mel frequency scale is defined as:

$$mel = 2595 * log_{10}(1 + hertz/700) \tag{4}$$

- **MFCC:** represents accurately the vocal tract that is a filtered shape of a human voice and also manifests itself in the envelope of short time power spectrum as shown in Figure 2(c). In order to compute MFCCs, a set of sequential steps should be handled, which are:
  - **Frame the signal into short frames:** frame the audio signal into 20-40ms (25ms is standard) frames to overcome changing in samples in short time of period as it is constantly changed in a large period of time.
  - **Periodogram of power spectrum:** calculates for each frame the periodogram estimation of the power spectrum, which identifies the frequencies in the frame.
  - **Apply the Mel filterbank to the power spectra (or sum the energy in each filter):** a filter is required for estimating the energies in various frequency regions that appear in a group of aggregated periodogram bins due to unnecessary information in periodogram spectral estimation. Hence the Mel filterbank estimates the energy near 0 Hertz and then for higher frequencies as less concern is considered for variations.
  - **Logarithm of all filterbank energies:** large variations of energies is scaled using a logarithmic scale as there are no different sounds in large energies. The logarithmic scale is a channel normalization technique that is also exploited for cepstral mean subtraction.
  - **DCT of the log filterbank energies:** Due to the correlation in filterbank energies that lead to overlapping, The DCT is used to decorrelates the energies. This generates diagonal covariance matrices as features.
  - **2-13 DCT coefficients:** choose higher DCT coefficients to reduce the fast changes in the filterbank energies and discard the rest.
- **Chorma-stft (Short Time Fourier Transform):** Computes a chromagram from a waveform or power spectrogram as shown in Figure 2(d). Chroma features are powerful representations for music audio in which the entire spectrum is projected onto 12 bins representing the 12 distinct semitones (or chroma) of the musical octave.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0047.v1

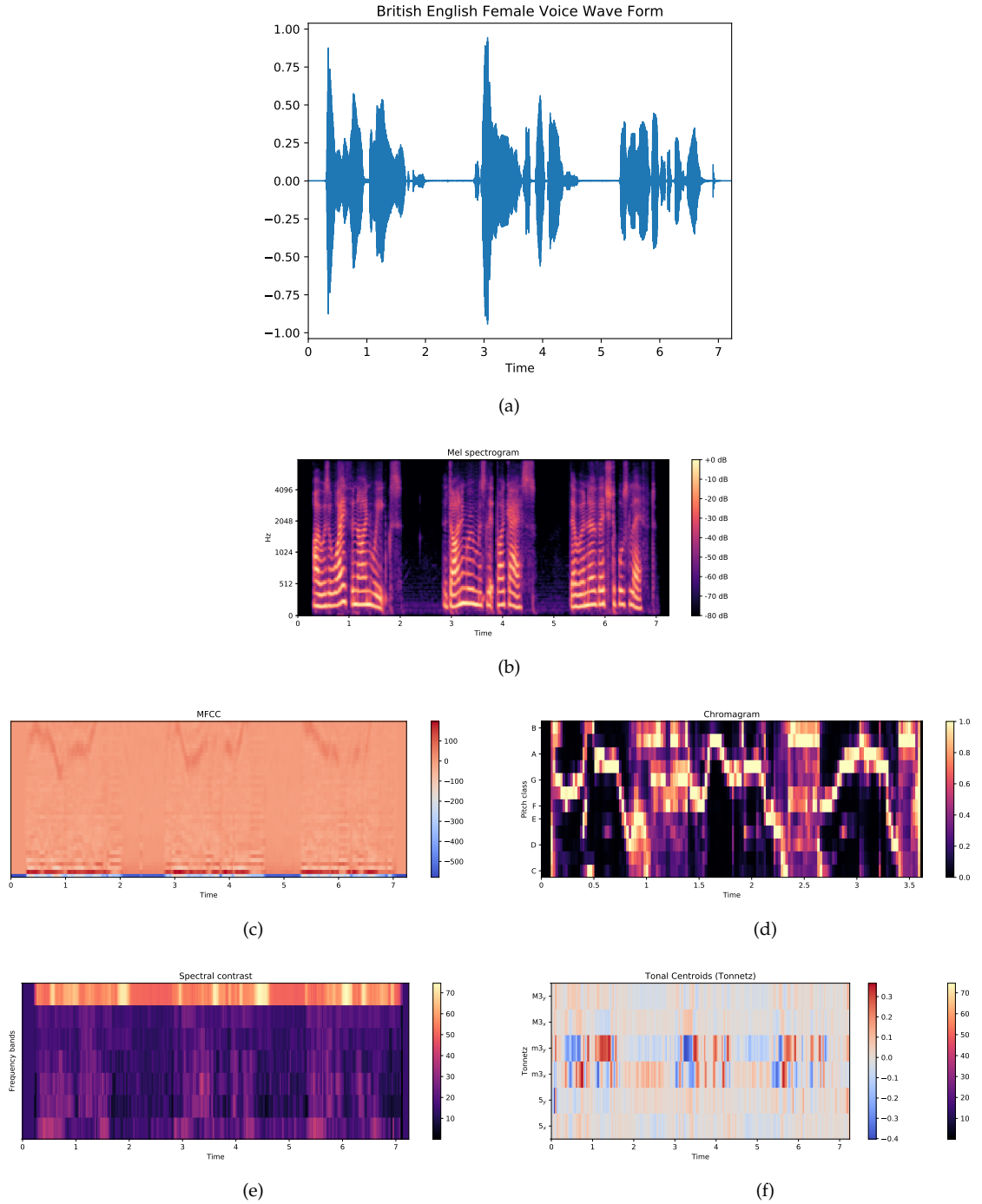


Figure 2. British English Female Sample and its features: (a) The Voice Sample (B_eng_f1.wav), (b) Mel-spectrogram, (c) MFCC, (d) Chromagram, (e) Contrast, and (f) Tonnetz.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0047.v1

- **Spectral_contrast:** Computes spectral contrast, using method defined in [1]. It represents the relative spectral distribution instead of the average spectral envelope.
- **Tonnetz:** Computes the tonal centroid features (or Tonnetz), following the method in [1] that detects changes in the harmonic content of musical audio signals.

Classification learning Techniques: Classification learning algorithms aim to find an optimal classifier model for recognizing test samples of provided features and unknown labels. Several learning techniques fundamentally reveal a philosophical theory in modelling knowledge as mathematical form. In order to cover the diversity in using different forms, a set of classification learning algorithms of various families are used. In particular, the selected classifiers in terms of the family include [12]:

- **Bayes:** is a direct approach that finds the best hypothesis by using the Baye's theorem as a probability theorem for building a rule or graph-based classification models. Two well-known methods are used; which are the Bayesian Network (BN) and Naive Bayes (NB) models.
- **Functions:** In this family, the classifier builds a function (or hypothesis) of input domain (i,e. Features) and maps it into a range of output (i,e. Labels) to form a function for classification. A set of models are used which are Multi-Layer Perceptron (MLP), SMO (Sequential Minimal Optimization for SVM), Logistic (L), Support Vector Machine (SVM linear (S_L), SVM polynomial (S_P), and SVM radial (S_R)), and Latent Dirichlet Allocation (LDA).
- **Deep Neural Network (DNN):** is a framework of two phases; which are feature engineering and classification [13,14]. The feature engineering process automatically extracts useful and non-linear features from the raw data using convolution and pooling layers by optimizing the weights W (or feature maps) between layers [15]. In the classification phase, the useful features are flattened as a vector to be fed into a fully connected ANN. In this work, the architecture of DNN, as shown in Figure 3, receives the MFCC features of input voice as one-dimensional (1D) data. These features are then fed into a convolution layer that consists of three layers of 32, 48 and 120 neurons using ReLu as a non-linear activation function. A pooling (or sub-sampling) layer follows the conventional layers using max function to reduce the size of resulted features. Finally, such features are flattened as the input vector to a fully connected ANN which is three dense layers of 128, 64 neurons using ReLu function and 2 output neurons represent the gender of the input voice using softmax function (i.e, a normalized probability function). Two 1D-DNN models are used which are Normalized Deep Convolution Neural Network (DL_n) and Deep Convolution Neural Network (DL). The parameter settings of DNN are 1000 epochs (or number of iterations), 25% dropout (or regularization), adam optimizer and pooling and feature map size of 2×2.
- **Lazy:** Lazy learners simply classify a new sample by estimating the vector similarity between the sample features and the vectors of samples in the training set and then assign the label of the most similar ones to that test sample. Lazy classifiers differ from other methods known as eager learners. Eager learners construct a machine learning model before testing process as a ready to use classifier models. The lazy learners used in this study are IBk and KStar (k^*).
- **Meta:** the idea is to learn an expert classifier of ensemble weak classifiers combined in a way to predict a label using averaging or voting methods. AdaBoost (Ada) and Bagging (B) are used as well-known algorithms.
- **Trees:** each classifier is a form of a hierarchical tree where a node at each level represents the best attribute at that level while the arcs represent the values of that attributes. The Decision Tree (J48) and RandomForest (RF) models are used.
- **Rules:** traverse each feature values and create a rule by finding the most frequent label. The criterion for selecting features depends on calculating the error rate of the rule. Three techniques are used; which are OneR (1R), Ridor (R), and RoughSet (RS) models.

Feature Selection Techniques: Building an optimal classifier model is affected by no-relevant features used for constructing such a model. These features drive the model to produce low accuracy for

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0047.v1

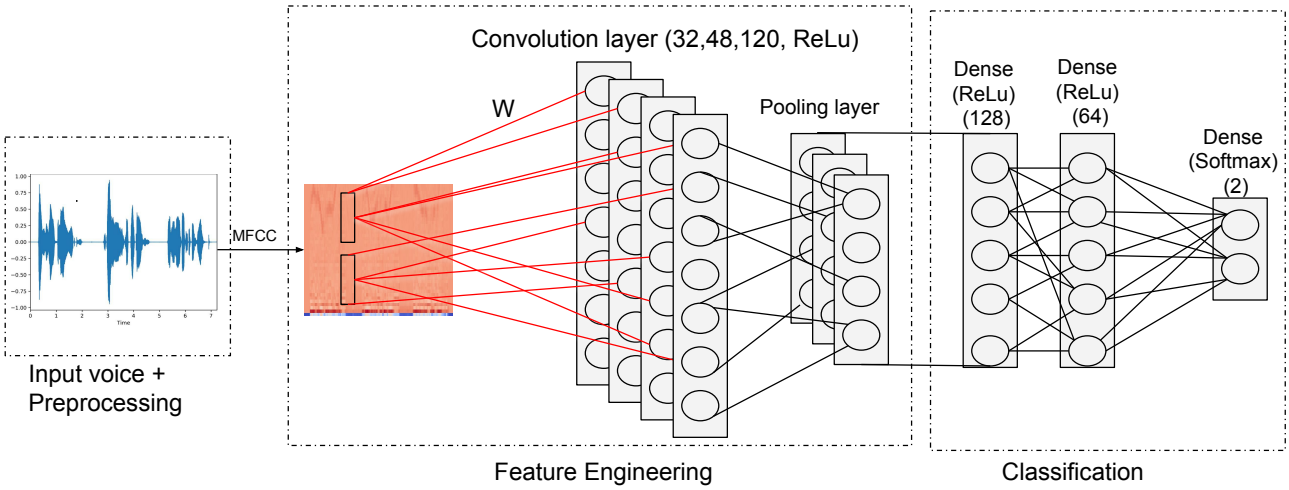


Figure 3. one-Dimensional Conventional Neural Network.

provided labels that lead to under-fitting or over-fitting problem. Therefore, the necessity for selecting relevant subset is needed. Three feature selection optimizers are used which are derived from the natural behaviour which are Evolutionary Search, Particle Swarm Optimization (PSO) search, and wolf search [16–18]. Each algorithm generates a set of individual solutions and then selects the optimal solution based on an evaluation metric and a learner optimizer (or Evaluator). In this work, the evaluation metric used is Area Under the ROC Curve (AUC) to validate whether a classifier can separate positive and negative samples and identify the best threshold for separation [19]. On the other hand, the RF classifier from tree family is used as an evaluator to select the best subset features. **Evaluation:** In this phase, particularly in the training phase, the 10-fold cross-validation method is used for each experiment by repeating it 10 times at each process of building a classifier. The evaluation metrics used are precision and recall [20]. Precision is the ratio of relevant samples among the retrieved ones, while recall is the ratio of retrieved and relevant samples to the total amount of relevant samples.

3. Experimental Results

A set of experiments are conducted for evaluating the contributions, which include studying the efficiency of extracted features, evaluating different learning techniques and analyzing the three natural optimizers used for feature selection. The dataset are demonstrated, the experimental parameters and settings and then presents the evaluation of the presented contributions.

3.1. Experimental Settings

A standard dataset of artificial voices from ITU-T Recommendation P.50 [21] are used. The dataset consists of 20 languages. Each language has 16 voice samples of eight files for each gender. The artificial voice is a signal mathematically produced for regenerating the time and spectral characteristics of the human speech. These artificial voices have over-bandwidth between 100 Hz and 8 kHz, which significantly affect the performance of linear and nonlinear telecommunication systems. The artificial voice is mainly used for objective evaluation of speech processing systems and devices. A single-channel with continuous activity (i.e. without pauses) is sufficient for measuring characteristics. The advantage of generating artificial voice is that it is more easily to be generated and having smaller variability than real voice.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0047.v1

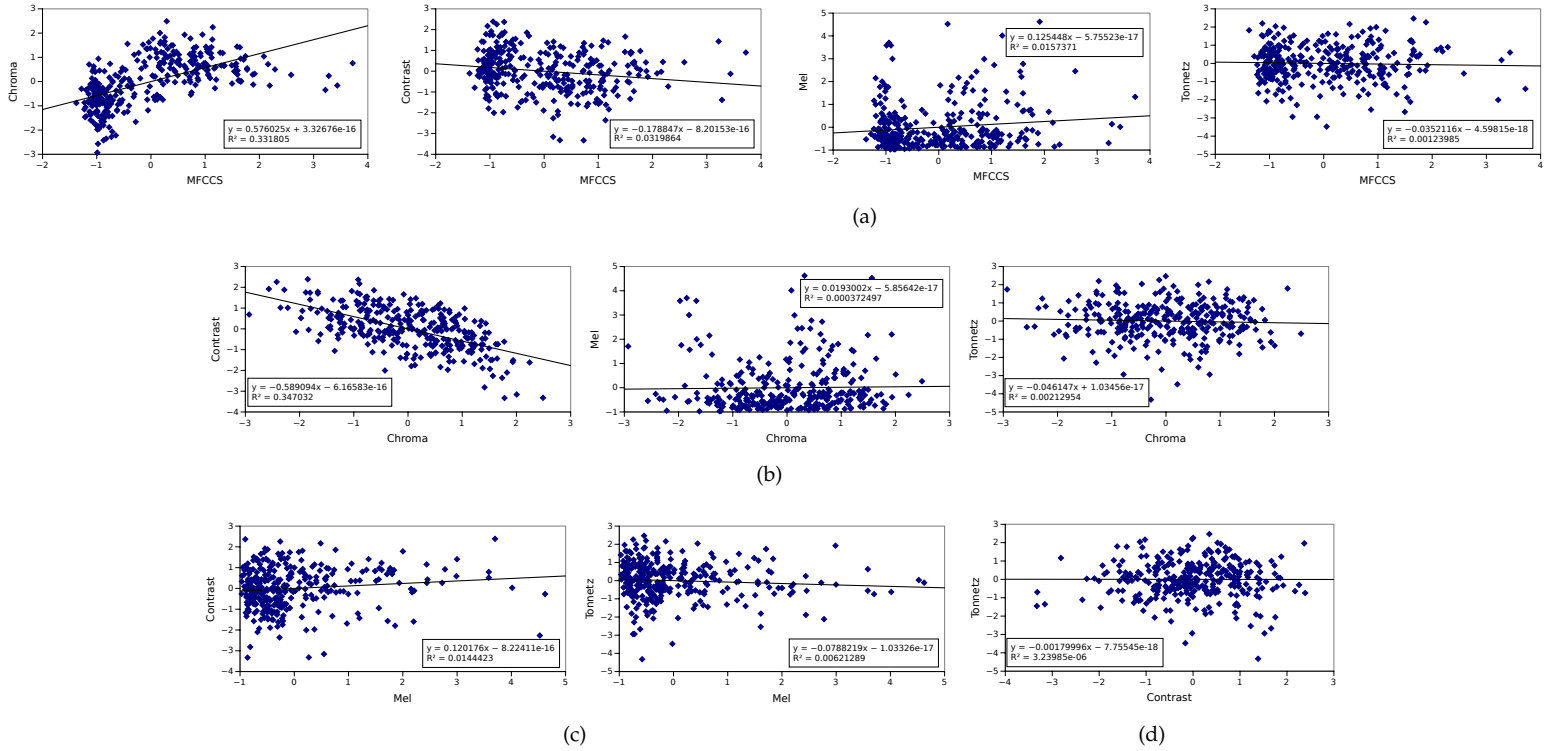


Figure 4. Features correlations and line equations (a) MFCCs .vs Others; (b) Chroma .vs Rest; (c) Mel .vs Rest; and (d) Tonnetz .vs Rest.

The parameter settings of natural feature selection methods in weka[1] tool are used as default settings. Although there are a set of natural methods, this study presents the most common ones, which are the Evolutionary search, PSO search, and Wolf search.

3.2. Voice Features Effect and Correlation

In order to study what features relevant to build an optimal classifier, the correlation between features has to be examined to demonstrate how they are related to each other. Four feature types, in the present work, are considered including MFCCs, Chroma, Mel and Tonnetz. The correlation between features is presented in Figure 4 that shows a scattered plot represents the correlation relationship among different feature types. Each chart contains a linear regression equation that formulates the evaluated feature values. In addition, it clarifies the R^2 correlation coefficient. The R^2 is a statistical measure determines how close the real data points that are fitted by the linear regression model. This means that if the R^2 value is close to 1, the data is highly fitted to the regression line and there is no difference in their effects on tested labels or there is, in contrast, a bad correlation to the labels.

In particular, as shown in Figure 4, the best R^2 is between the MFCCs and Chroma features of 0.332. In contrast, based on the chart, the worst correlation occurs between the Chroma and Contrast features of R^2 equal to 0.35. At each feature category, the MFCCs feature has the best correlation with Chroma feature of 0.332 and worst correlation with Tonnetz feature of R^2 equal to 0.0012. The Chroma feature has the worst correlation with Mel feature of R^2 equal to 0.0004 compared to the rest feature categories. The Mel feature behaves in a worse correlation value in comparing with Tonnetz feature as shown in Figure 4(c) of R^2 equal to 0.0062. The Tonnetz features have the worst correlation value of $3.2e^{-6}$ compared to the Contrast feature category.

[1] https://www.cs.waikato.ac.nz/ml/weka/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0047.v1

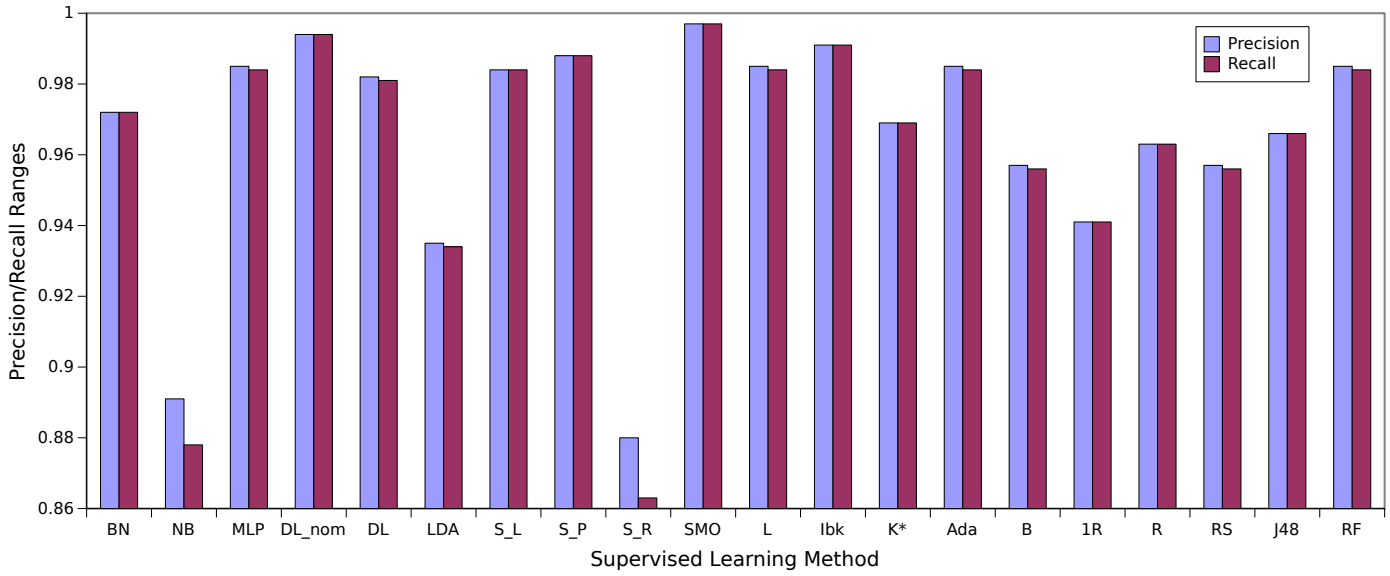


Figure 5. Precision/Recall of Classifier Models.

In summary, the MFCCs, Chroma, and Mel features perform in an efficient performance as they are more related to each other. The reason is that these features extract high energy coefficients from the signal where, in contrast, the other features concern with the tone of the musical signals. This answers the research question **RQ1** that ensures the importance of selecting more suitable features for possibly building more accurate classifiers for voice paralinguistic information aspects.

3.3. Voice Gender Recognition and Classification

This work aims to build a classifier for recognizing the gender of a given human voice. The gender voices, whether it is male or female, they are different from each other by the signal energy and tune. Therefore, it is necessary to construct a classifier model to differentiate the given human voice to male or female due to it is important in many applications as discussed before.

In order to ensure diversity in constructing and using the classification models, different theoretical supervised classifiers are built as shown in Figure 5. The figure shows a bar chart of the x-axis represents the supervised learning methods and the y-axis represents the Precision/Recall evaluation metrics ranged from 0.86 to 1. As shown, the function family experimentally has overall superior performance results compared to the other families especially at the DL_norm and SMO techniques of approximately 99.97% and 99.7%, respectively. However, in particular, the BN technique has better performance than the NB in the Bayes family of approximately 10.2%. This means that the probabilities in the network graph of the BN method are more robust compared to the generated rules in the NB method. In the function families, the DL_norm and SMO techniques locally obtain significant performance as they have in general. The IBk technique gains leading performance compared to K* method of 2.2% as the IBk method has Recall value of 99.1% and K* method has Recall value of 97%. The Adaboost of the meta family gets high values of Precision/Recall compared to the B technique of increasing percentage reached to 13.12%. In incremental Rules family, the R technique gains consistent performance values compared to the 1R and RS methods. The methods in the Rules family have low performance compared to the other families, but the results are presented to explain such a conclusion. In the Tree family, it is clear that the RF gains consistent results in comparison with the J48 technique of increasing the performance value reached 2.1%.

In brief, we discussed constructing classifier models for recognizing the voice gender by using various techniques of different theoretical families. The results showed that the function family gains

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0047.v1

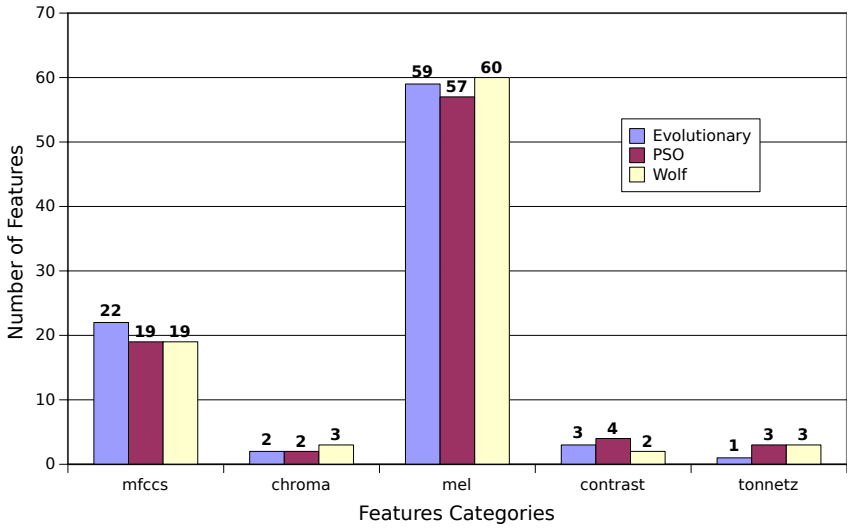


Figure 6. Number of Features at each Category.

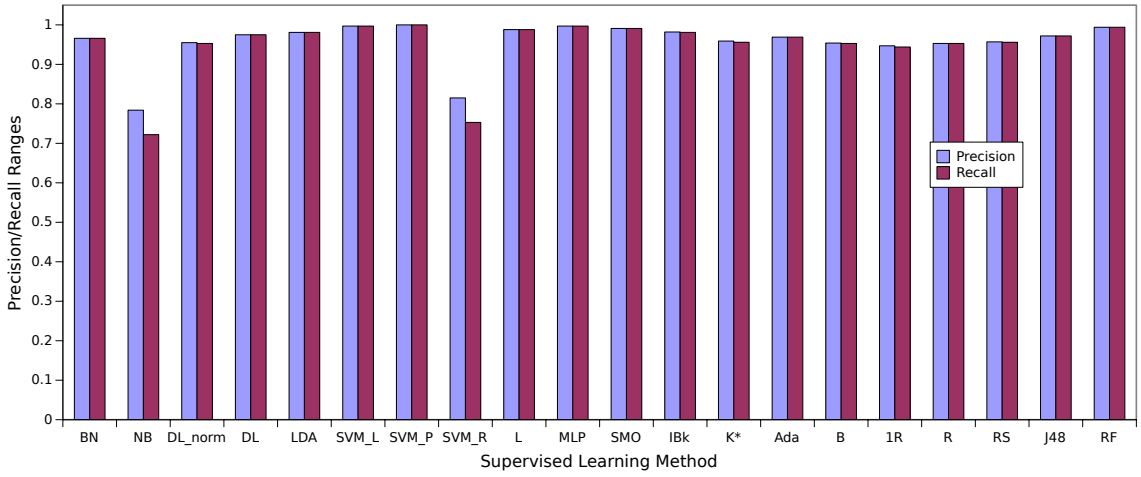
consistent and significant performance values using the DL_norm and SMO techniques. This means that the theoretical methodology for building such models also has an effect in discriminating the human voice gender. Each ML technique algorithmically using good features could be a promising method for the voice gender recognition applications. Thus these results lead us to answer the research question **RQ2** affirmatively.

3.4. Feature Selection Techniques and Results

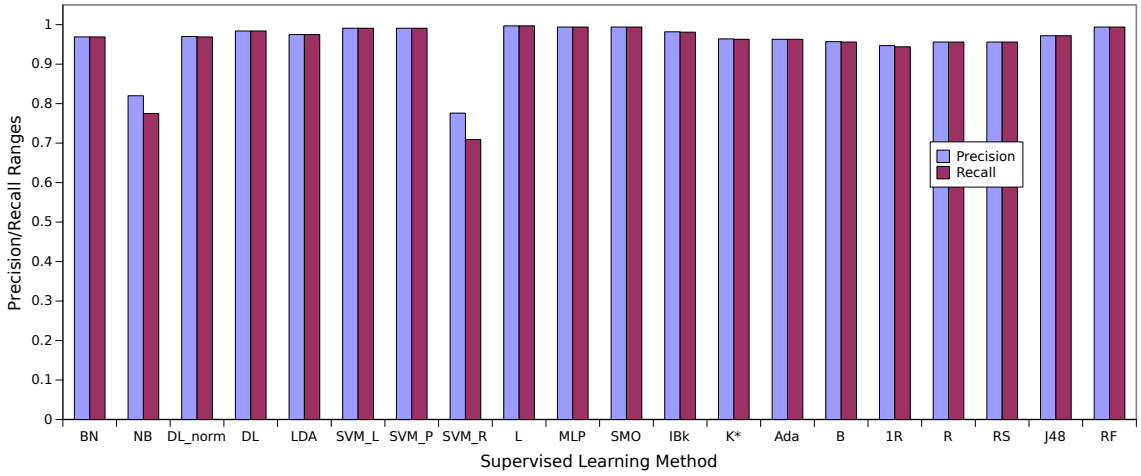
The performance of building classification techniques is affected by the quality of features in the training set. As such, it is necessary to select the most optimal features for enhancing the performance of voice gender recognition models. The three most common optimizers are exploited for feature selection inspired by natural behaviour, which are the EA, PSO, and Wolf techniques as wrapper selection algorithms. These methods technically depend on searching a space of solutions and selecting the most optimal ones in an evaluated classifier such as the RF evaluator.

Five feature categories are discussed with a set of features for each category. There are 40, 12, 128, 7 and 6 subset features for the MFCCs, Chroma, Mel, Contrast and Tonnetz, respectively. However, Figure 6 shows a bar chart that explains the number of selected sub-features at each category using the three selection methods. The x-axis represents the feature categories while the y-axis represents the number of selected sub-features for the selection methods. In particular, the percentage of selected sub-features using the EA, PSO and Wolf techniques are (55%; 47.5%; 47.5%), (16.7%; 16.7%; 25%), (41.1%; 44.5%; 46.88%), (42.86%; 57.1%; 28.6%), and (16.7%; 50%; 50%) on average of 34.5%, 43.2% and 39.6% for MFCCs, Chroma, Mel, Contrast and Tonnetz, respectively. The percentages show that the EA algorithm selects a small number of sub-features at Chroma and Tonnetz categories. The PSO technique has a small number of sub-features only at the Chroma categories while the Wolf method selects a small number of sub-features at the Chroma and Contrast categories. *The question is how the effect would be if using these sub-features on ML techniques for gender voice recognition applications.* In order to evaluate the selected sub-features from the three feature selection techniques on the recognition of the human voice gender, similar experiments are conducted on the same families of ML techniques as shown in Figure 7. The figure shows three bar charts of the three selection algorithms with x-axis and y-axis similar to Figure 5. The results manifest improvements in the performance of classifier models in overall families with approximately the same effect of three techniques on selecting the optimal features according to the performance of ML techniques. In particular, the best performance

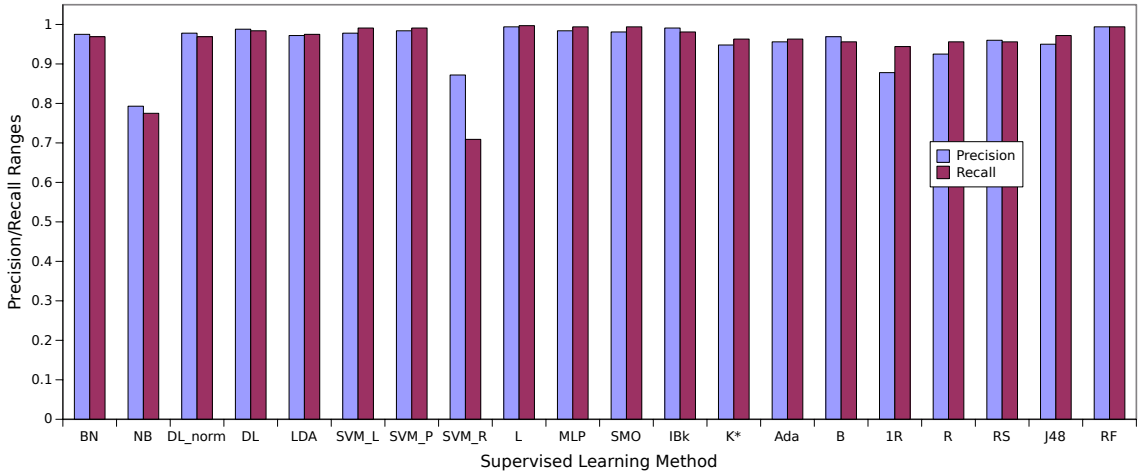
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0047.v1



(a)



(b)



(c)

Figure 7. Swarm Feature Selection Techniques (a) PSO Search; (b) Evolutionary Search; and (c) Wolf Search.

Precision/Recall values after sub-features selection for the Evolutionary, PSO and Wolf methods are for the L method of 99.7%, the SVM_P method of 100% and the RF method of 99.4%, respectively.

In summary, these results ensure that the EA selection algorithm gains a high evaluation performance of 99.7% Precision/Recall value at the L method and also with a small number of sub-features on percentage average of 34.5% compared to the PSO and Wolf techniques. If there is tolerance in the percentage of selecting sub-features, the PSO shows a superior result in classifying the human voice gender reached 100% using the SVM_P ML technique. Hence, using natural feature selection algorithms is useful in enhancing the performance of ML techniques resulting in a small number of relevant features. This accordingly answers the research question **RQ3**.

4. Conclusion and Future Work

Recognizing the gender of human voice has been considered among of the challenging tasks due to its importance in various applications. The contributions are combined into three-folds include: (i) studying the extracted features by examining the correlation between each other, (ii) Building classification models using different ML techniques from distinct families and (iii) evaluating the natural feature selection techniques in finding the optimal subset of relevant features on classification performance. In particular, three feature categories perform in an efficient behaviour due to their theoretical methodology in extracting the relevant coefficient energies in the voice signal, which are the MFCCs, Chroma and Mel that answers the research question **RQ1**. In the performance of classifiers perspective, the ML techniques behave in different ways. The results showed that the function family gained better performance compared to the other families. Although the function family had superior results, the other techniques have promising results and this leads us to answer the research question **RQ2**. Finally, a set of experiments are conducted using three common feature selection techniques inspired by nature, which are EA, PSO and Wolf methods using the RF as an evaluator. These wrapper selection techniques select sub-features from the feature categories which is on average approximately 39.1% on overall features. In spite of a small number of sub-features, the performance of ML techniques was increased as there are some features are not relevant in determining the gender of human voices. This also answers the research question **RQ3**.

In future work, more experiments are being conducted to use many feature categories, ML techniques and try using other natural feature selection techniques. Furthermore, the proposed techniques are being examined on different datasets since just a standard artificial voice is used from ITU-T Recommendation P.50 [21]. The reasons behind using it are that it contains many different languages (i.e, 20 languages), as well as, the voice text is too long.

1. Hu, Y.; Wu, D.; Nucci, A. Pitch-based Gender Identification with Two-stage Classification. *Sec. and Commun. Netw.* **2012**, *5*, 211–225.
2. Djemili, R.; Bourouba, H.; Korba, M.C.A. A speech signal based gender identification system using four classifiers. 2012 International Conference on Multimedia Computing and Systems, 2012, pp. 184–187.
3. Li, M.; Han, K.J.; Narayanan, S. Automatic Speaker Age and Gender Recognition Using Acoustic and Prosodic Level Information Fusion. *Comput. Speech Lang.* **2013**, *27*, 151–167. doi:10.1016/j.csl.2012.01.008.
4. Yücesoy, E.; Nabiyev, V.V. A new approach with score-level fusion for the classification of a speaker age and gender. *Computers & Electrical Engineering* **2016**, *53*, 29 – 39. doi:https://doi.org/10.1016/j.compeleceng.2016.06.002.
5. Lee, M.W.; Kwak, K.C. Performance comparison of gender and age group recognition for human-robot interaction. *IJACSA) International Journal of Advanced Computer Science and Applications* **2012**, *3*.
6. Abdulla, W.; Kasabov, N.; Zealand, D.N. Improving speech recognition performance through gender separation. *changes* **2001**, *9*, 10.

7. Qawaqneh, Z.; Mallouh, A.A.; Barkana, B.D. Deep neural network framework and transformed MFCCs for speaker's age and gender classification. *Knowledge-Based Systems* **2017**, *115*, 5 – 14. doi:https://doi.org/10.1016/j.knosys.2016.10.008.
8. Sharan, R.V.; Moir, T.J. Robust Acoustic Event Classification Using Deep Neural Networks. *Inf. Sci.* **2017**, *396*, 24–32. doi:10.1016/j.ins.2017.02.013.
9. Proakis, J.G.; Manolakis, D.G. *Digital Signal Processing (3rd Ed.): Principles, Algorithms, and Applications*; Prentice-Hall, Inc.: Upper Saddle River, NJ, USA, 1996.
10. Grimaldi, M.; Cummins, F. Speaker identification using instantaneous frequencies. *IEEE Transactions on Audio, Speech, and Language Processing* **2008**, *16*, 1097–1111.
11. Kanatani, K.i. Fast Fourier transform. In *Particle characterization in technology*; CRC Press, 2018; pp. 31–50.
12. Witten, I.H.; Frank, E.; Hall, M.A. *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed.; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2011.
13. Di, W.; Bhardwaj, A.; Wei, J. *Deep Learning Essentials: Your Hands-on Guide to the Fundamentals of Deep Learning and Neural Network Modeling*; Packt Publishing, 2018.
14. Li, Z.; Hoiem, D. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2018**, *40*, 2935–2947.
15. Hinterstoisser, S.; Lepetit, V.; Wohlhart, P.; Konolige, K. On pre-trained image features and synthetic images for deep learning. European Conference on Computer Vision. Springer, 2018, pp. 682–697.
16. Chtioui, Y.; Bertrand, D.; Barba, D. Feature selection by a genetic algorithm. Application to seed discrimination by artificial vision. *Journal of the Science of Food and Agriculture* **1998**, *76*, 77–86.
17. Xue, B.; Zhang, M.; Browne, W.N. Particle Swarm Optimization for Feature Selection in Classification: A Multi-Objective Approach. *IEEE Transactions on Cybernetics* **2013**, *43*, 1656–1671.
18. Emary, E.; Zawbaa, H.M.; Grosan, C.; Hassenian, A.E. Feature Subset Selection Approach by Gray-Wolf Optimization. Afro-European Conference for Industrial Advancement; Abraham, A.; Krömer, P.; Snasel, V., Eds.; Springer International Publishing: Cham, 2015; pp. 1–13.
19. Lusted, L.B. Signal Detectability and Medical Decision-Making. *Science* **1971**, *171*, 1217–1219.
20. Davis, J.; Goadrich, M. The Relationship Between Precision-Recall and ROC Curves. Proceedings of the 23rd International Conference on Machine Learning; ACM: New York, NY, USA, 2006; ICML '06, pp. 233–240.
21. ITU-T Recommendation P.50. Objective measuring apparatus. ***International Telecommunication Union-Telecommunication Standardization Sector (ITU-T)***,*Geneva* **1999**.