Gender Classification Of Mixing And De-Mixing Speech

Sawsan Hadi Abed¹, Nidaa A. Abbas²

¹,² Department of Software, College of IT, University of Babylon, Babil, Iraq.

Abstract
Gender classification is growing in popularity due to the variety of fields in which it can be used. It can be employed in various fields, including criminal investigations and security and authentication services. Gender classifying speech for different speakers is still a demanding and challenging task for recognizing overlapped speech and building a robust prediction model. The paper provides a gender classification system that uses Independent Component Analysis (ICA) and several machine learning algorithms to identify mixing and de-mixing speech signals. ICA is employed to separate the mixed signal into their source signals. The system consists of two stages: the first stage is the mixing and separating process for signals. The second stage involves combining feature extraction and constructing a classification model to determine whether a signal is male or female based on its acoustic attributes. The system will evaluate the efficacy and significance of machine learning algorithms for selecting the optimal method to identify the speaker's gender with the most excellent efficiency and accuracy. Experimentation shows that the best accuracy value for an SVM model with mixing speeches is 87.1 %, and the best accuracy value for a Neural Net and SVM model with de-mixing speeches is 97.8 %.

Keywords: Acoustic properties, Gender classification, ICA, Machine learning techniques, speech processing.

1. Introduction
Nowadays, the classification of gender is one of the essential processes in speech processing. The technique of gender classification aims to define the gender of the speaker (male or female) through voice signals analysis. It is a significant function that will improve the efficacy of many applications such as speech recognition, human-to-machine interaction, speech emotion recognition, investigating criminal voice, and sorting telephone calls by gender categorization. It can enhance the systems of human-computer interaction, particularly dialogue systems, by
customizing services that depend on the voice of gender and enhancing the level of user acceptance (Shivani. S. Tejale and Tushar B. Kute, 2020). Human speech is an effective tool for communication, which includes several unique features like gender, age, and emotive situation. Each sound wave carries a particular frequency, making a human voice distinct from the voices of others. The acoustic properties of voice signals, such as time, strength, frequency, and filtering, can be obtained by analyzing the speech signals.

Regarding gender classification, features extraction is important since it helps identify variations between different speech signals. The speech signal includes vital information about the speaker, like the speakers' gender and age. In order to classify a speech, it is necessary to highlight the most distinguishing features of the speech signal. Person voice must be transformed to digital form first so that beneficial attributes may be extracted.

The human ear is a natural voice classification system. It has a mechanism to recognize gender by speech based on different factors. Similarly, machine learning methods will build and train a model to identify gender-based on voice. It is possible to classify gender based on speech signals using a variety of algorithms, including K-nearest neighbors (K-NN), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), ANN (Artificial Neural Network), and Deep Learning (DL). However, selecting the best algorithm is a difficult task. They use classification learning methods, a test set of features, and unknown labels to determine the best classifier model. The quality of features plays an essential role in enhancing the robustness and effectiveness of classifiers learned from a training set using ML techniques (Rami S. Alkhawaldeh, 2019).

This paper used a fast ICA algorithm to separate the mixing signals and recover the original signals. Then various machine learning techniques (Decision Tree, SVM, RF, K-NN) and acoustic properties of speech signal were utilized to build the classification models. Details of the techniques used and the results are shown throughout the paper.

2. Related Work

Since speech recognition technology is utilized in conjunction with various machine learning methods, there have been numerous advancements in this flexible sector. More and more scientists have been working in the last several years to develop a comprehensive collection of characteristics and models that highlight the differences between the male and female voices.

The approach provided by (Shivaji and Ramesh, 2015) enables identification, classification, and recognition of the exact gender and age of the speaker while also showing speaker profiles. The approach improves identifying gender, age group, and even emotions. Using Mel Frequency Cepstral Coefficient feature extraction and the Gaussian Mixture Model, the proposed system is an excellent vector for selecting system features and modeling feature sets. The classification of a speaker's speech is done using SVM and feature matching techniques. There is an average accuracy and performance result of 80.3% for gender identification and an average accuracy and
performance result of 89.3 % for age recognition in the system (Shivaji J Chaudhari and Ramesh M Kagalkar, 2015).

E. Yucesoy and V. V. Nabiyev (2016) suggested a system for classifying speakers based on a mixture of seven subsystems with attribute vectors of MFCC, PLP, and depend on three separate classifiers at the score level: GMM, SVM, and GMM-SV-based SVM. The Gender database has a 90.4 % categorization success rate for gender identification (E. Yucesoy and V. V. Nabiyev, 2016).

(Mucahit and Ali, 2016) proposed the deep learning model to recognize the gender of a voice. The model demonstrates that gender may be deduced from the acoustic features of voice and speech. In order to obtain the classification model from a dataset containing parameters for speech samples, an Multilayer Perceptron method was applied, and a model with an accuracy of 96.74 % was proposed (Mucahit Buyukyilmaz and Ali Osman Cibikdiken, 2016).

(Saptarshi et al., 2017) proposed developing a system based on perceptual audio features and training a model of classifiers to distinguish between the two types of gender. Two processes are involved in an automatic speech discrimination system: Extraction of voice features from the input speech signal like tempo-based features, pitch, short-time energy, etc. After extracting the process, use supervised classifiers (K-NN and SVM). The researchers used Principal Component Analysis to select features and reduce feature dimensions. The SVM technique surpasses the K-NN model, increasing accuracy by up to 28 % (Saptarshi Sengupta and Ghazaala Yasmin, 2017).

(Kudakwashe and Oludayo, 2018) suggested a gender speech recognition technique that combined feature selection with gender classification using the Random Forest Recursive Feature Elimination (RF-RFE) algorithm and Gradient Boosting Machines (GBMs). Acoustic characteristics were gathered from 1584 males and 1584 females in a public gender voice dataset. After using the RF-RFE with the dataset, the GBMs' classification accuracy improved. Without feature selection, the GBMs algorithm had a precision of 97.58 %, whereas, with feature selection, it had a precision of nearly 100 % (Kudakwashe Zvarevashe and Oludayo O. Olugbara, 2018).

(Gaurav and Rekha, 2019) suggested an emotion and gender recognition system using two classifiers, Support Vector Machine and Nave Bayes, using four-voice variables for classification: shimmer, energy, and pitch. The emotions and gender are detected using attributes extracted from speech signals. These input sequences are fed into the classifiers, which properly detect the gender and emotion of the speech stream. In terms of accuracy, the SVM classifier surpasses the Naive Bayes classifier by about 10% in gender identification and 35% in emotion detection (Gaurav Aggarwal and Rekha Vig, 2018).

Gyanendra and Shuchi (2020) proposed an SVM model with PCA for gender recognition using a person's voice. PCA and an SVM classifier are combined in the proposed model. A small dataset was utilized for training and testing the classifier. The features were collected from voice samples with a maximum duration of 3 seconds. During validation, the prediction accuracy using simply SVM was equal to 91%. However, the accuracy attained using PCA and SVM with linear kernel about 98.42 % with good precision and recall (Gyanendra Sharma and Shuchi Mala, 2020).
3. Independent Component Analysis (ICA)

ICA is a machine learning technique to estimate independent sources from a mixed-signal. It is a statistical and computational method used to separate the mixture signals and recover the original signals, depending on the statistical independence of the signals. ICA has many algorithms like Infomax and FastICA. These algorithms aim to estimate sources signal (independent components) by using the maximum likelihood (ML) estimation method, maximizing the non-Gaussianity, or minimizing the mutual information. (A. Hyvarinen, J. Karhunen and E. Oja, 2001)

The mathematical formula of the mixing process can be formed as:

\[ x = A * s \]  

where \( s = [s_1, s_2, \ldots, s_n] \) is \( n \times 1 \) vector of original signals, \( x = [x_1, x_2, \ldots, x_m] \) represent \( m \times 1 \) mixed vector and \( A \) is an \( m \times n \) non-singular and full-rank mixing matrix.

The ICA is to discover the inverse of the matrix \( A \), \( W \) that results \( y \), the de-mixing transformation model is as:

\[ y = W \times x \]  

Note that \( y = [y_1, y_2, \ldots, y_n] \) represent \( n \times 1 \) separated vector and estimation of the source signal, and \( W \) is an \( m \times n \) estimated unmixing matrix used in the separation process. The following figure is an illustration of an Independent Component Analysis using the cocktail party. Two speech signals are produced from two persons and then recorded by microphones which mix the two source signals linearly. ICA was recovering the source signals from the mixed signals.

Fig (1): Cocktail Party Problem

To get an optimal estimate of the independent components, it is helpful to do several preprocessing on the data. The first preprocessing step is centering. Centering is performed using the subtraction of the average values from the observed signals to obtain the observed signals with zero mean value. This simplifies the ICA calculation and reduces computational time. (Daniel Palacios, Victoria Rodellar, 2019). This operation is computed as in equation (3).
\[ x' = x - \text{E}[x] \]  

(3)

where \( x \) is the mixed signals, and \( \text{E}[x] \) represents the mean of the mixed signals (Expectation). Adding the expectation vector to the produced source signals.

The second preprocessing method is called whitening. It is the most preprocessing important in the ICA approach. It is a linear transformation model of the centered vectors, producing uncorrelated mixed signals and having unit variance. The whitening transformation can be formed as in equation (4)

\[
\bar{x} = \Lambda D^{-1/2} \Lambda^T x
\]

(4)

Where \( \Lambda \) and \( D \) are two matrices. The \( \Lambda \) columns are the eigenvectors of \( \text{E}[xx^T] \), and the diagonal of \( D \) are the eigenvalues of \( \text{E}[xx^T] \). The main advantage of whitening is producing an orthogonal mixing matrix used to recover the original signals.

### 3.1 The FastICA Algorithm

Standard Fast ICA is one of the most widely used linear ICA algorithms, depending on a fixed-point iteration approach. It supposes that there are major pre-processes (centering and whitening) performed on the mixed data. The basic model of this algorithm is the so-called one-unit model, where the computation unit can update the weight vector \( w \) by a specific learning rule. The Learning rule of this algorithm was used to determine the direction of the unit vector \( w \) and find the transformation \( w^T x \) that maximizes the non-Gaussianity of observed signals [13]. For the whitening process, the one-unit FastICA method has the form as given in equation (5).

\[
w(k) = E\{xg(w(k-1)Tx)\} - E\{g'(w(k-1)Tx)\}w(k-1)
\]

(5)

\( W(k) \) represents the weight vector, \( g \) refers to the contrast function, and \( g' \) is the first derivative of the contrast function.

The FastICA algorithm relies on a fixed-point iteration method to determine a maximum of the non-Gaussianity of \( w^T x \). Also, it can be built as an approximation Newton iteration method (Hussein M. Salman and Nidaa A. Abbas, 2021). The main steps of this algorithm can be brief as:

1. Take the initial value of the weight vector \( w \) randomly.
2. assume \( w^+ = E\{xg(w^T x)\} - E\{g'(w^T x)\}w \)
3. assume \( w = w^+ / \|w^+\| \)
4. If converged exit, otherwise go back to step 2.
Notice that the converged represents where the old values of w point are the same as the new values in the same direction.

There are many properties in this algorithm compared with other linear ICA algorithms:

1- Directly, it splits the components of the non-Gaussian distribution. It can estimate the components one-by-one.

2- The convergence of this algorithm is cubic.

3- It is simple to use in several linear mixing troubles.

4. Feature Extraction

The speech signal contains several useful features. These feature sets, along with the voice gender label, serve as the training data for creating a classifier model that can identify the gender of the speaker. The following table lists some of the acoustic properties that have been used to classify the gender of the person.

Table 1: Acoustic Properties of voice sample

<table>
<thead>
<tr>
<th>Properties</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>meanfreq</td>
<td>mean frequency (in kHz)</td>
</tr>
<tr>
<td>Sd</td>
<td>standard deviation of frequency</td>
</tr>
<tr>
<td>median</td>
<td>median frequency (in kHz)</td>
</tr>
<tr>
<td>Q25</td>
<td>first quantile (in kHz)</td>
</tr>
<tr>
<td>Q75</td>
<td>third quantile (in kHz)</td>
</tr>
<tr>
<td>IQR</td>
<td>interquantile range (in kHz)</td>
</tr>
<tr>
<td>kurt</td>
<td>Kurtosis</td>
</tr>
<tr>
<td>skew</td>
<td>Skewness</td>
</tr>
<tr>
<td>sp.ent</td>
<td>spectral entropy</td>
</tr>
<tr>
<td>sfm</td>
<td>spectral flatness</td>
</tr>
<tr>
<td>mode</td>
<td>mode frequency</td>
</tr>
<tr>
<td>centroid</td>
<td>frequency centroid</td>
</tr>
<tr>
<td>sp.ent</td>
<td>peak frequency (frequency with highest energy)</td>
</tr>
<tr>
<td>meanfun</td>
<td>average of fundamental frequency measured across acoustic signal</td>
</tr>
<tr>
<td>minfun</td>
<td>minimum fundamental frequency measured across acoustic signal</td>
</tr>
<tr>
<td>sp.ent</td>
<td>peak frequency (frequency with highest energy)</td>
</tr>
<tr>
<td>maxfun</td>
<td>maximum fundamental frequency measured across acoustic signal</td>
</tr>
<tr>
<td>-------------</td>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>meandom</td>
<td>average of dominant frequency measured across acoustic signal</td>
</tr>
<tr>
<td>mindom</td>
<td>minimum of dominant frequency measured across acoustic signal</td>
</tr>
<tr>
<td>maxdom</td>
<td>maximum of dominant frequency measured across acoustic signal</td>
</tr>
<tr>
<td>dfrange</td>
<td>range of dominant frequency measured across acoustic signal</td>
</tr>
<tr>
<td>modindx</td>
<td>modulation index. Calculated as the accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range</td>
</tr>
</tbody>
</table>

WarbleR package is designed to simplify acoustic analysis in R and allows users to collect acoustic features. It aims to assist the analysis of the acoustic signals in R. This package presents tasks to discover, manage and process multiple sound files, generate spectrograms of full recordings or specific signals and execute various measures of the acoustic signal (M. Araya-Salas and G. Smith-Vidaurre, 2017).

5. Classification Learning Techniques

5.1 K-Nearest Neighbors (KNN)

KNN is one of the most accessible machine learning algorithms that rely on supervised learning techniques used in classification. When a sample is labeled into the dataset, the classifier computes the distance between the neighboring and unknown samples to determine whose class belongs. The nearest neighbor is data points that are nearest in the distance to the unidentified sample. The distance measured to determine distance consists of Euclidean distance, Hamming distance, Jaccard distance, Manhattan distance, Cosine distance, Minkowski distance, Tanimoto distance, and Mahalanobis distance (Mucahit Buyukyilmaz and Ali Osman Cibikdiken, 2016).

5.2 Artificial Neural Network (ANN)

ANNs are the second most widely used technique in classification. It is very fast in the classification once trained, which is essential for real-time applications. It is simulated to human brain processes. It can learn and trained to discover solutions to classify data. A neural network is an oriented graph consisting of nodes linked and by the strengths of those connections named weights. The weights are modified by training the network corresponding to a specified learning rule until it achieves the required task correctly (Archana.G.S and M. Malleswari, 2015).

5.3 Decision Tree

A decision tree is a supervised classification algorithm. It makes classification models in the form of the tree structure. It consists of nodes, leaf nodes, and edges which each internal node is a
feature. The branch denotes a decision rule, and the leaf node represents the decision. It is constructed by information gain, gain ratio, and Gini index (B. Satya Prasad, 2019).

5.4 Support Vector Machine (SVM)

SVM is a linear classification model that employs a supervised learning approach. It maps data into a high-dimensional space and then creates a hyperplane or decision boundary to partition the data into classes. The hyperplane is employed for identifying unknown data samples. The algorithm is given labeled training data to find the best hyperplane. Using an SVM, the class of unknown samples may be defined by examining which side of the hyperplane it lies on (Saptarshi Sengupta and Ghazaala Yasmin, 2017).

5.5 Random Forest (RF)

Random Forest is a supervised classification learning technique, and it is one of the most accurate learning algorithms available. As the name suggests, a random forest is a combined number of trees to create a forest when each tree depends on the values sampled individually and with a similar distribution for all trees in the forest. This classifier creates decision trees from the training dataset then calculates the votes from several decision trees to determine the final class. Random forest is more efficient working on large data sets, estimates which variables are essential during classification estimates missing data, and maintaining accuracy when a larger amount of data is missing (Shivaji J Chaudhari and Ramesh M Kagalkar, 2015).

6. Proposed Methodology

The methodology of the proposed gender classification system contains a set of stages that start with extracting the relevant features that have distinctive characteristics representing speech signal properties. So, these features are chosen as inputs for building up a classifier model for identifying the gender of a person's voice. The proposed system has two phases:

A. Training phase

B. Testing phase

A. Training phase

Feature extraction was computed as features vector from speech signals. Vector of features used as inputs for constructing a classifier model for detecting the gender of an individual voice. The classification model is based on acoustic properties and multiple machine learning techniques of several families. The training phase used UTI-T audio dataset for training the system.

B. Testing phase

In this phase, a custom dataset was built, which contained mixing and de-mixing (separating) signals of both genders. During the training and testing phases, most processes are the same.

Figure 2 presents steps of the training and testing phase of the proposed system:
6.1 Mixing Signals

Initially, required mixed speech signals under some considerations: as mono sounds, frequencies of signal 16KHz, .wav format, and noiseless. Also, achieve the identical, independent distribution (i.i.d.) as possible. After that, randomly determine the mixed matrix that can achieve the best-mixed case under the well-condition number of the mixed matrix. The mixing process can be summarized in the following steps:

1. Take two mono-speech signals which are noiseless with have same length and frequencies under condition of i.i.d metric.

2. mixture coefficient was prepared, which achieves the condition number factor, so the signals are mixing depending on the equation (1).

6.2 Separating Signals

The separation procedure employs independent component analysis (FastICA algorithm) to recover the original signals from mixtures. After Performing the mixing process, centering and whitening (preprocesses) are implemented on the mixing signals as the preprocessing steps of ICA.
Then, the proposed system is performed the ICA approach to separate (recover) many mixed signals. Each source signal can be extracted from a set of signal mixtures by multiplying it by an unmixing matrix.

6.3 Feature Extraction

Feature extraction can be brief in the following steps:

1-Read content of each of the speech samples, which are in WAV format.

2- Fed content of samples into the warbleR R package for extracting features. WarbleR is designed to simplify acoustic analysis in R and allows users to collect acoustic features.

6.4 Classification

In order to recognize test samples with specified characteristics and unknown labels, learning algorithms try to create the best classifier model. Because the choice of a classifier for the gender classification problem mostly depends on the classification accuracy. A classifier was built by a set of machine learning algorithms of many families to identify the gender of a given speech. At last, the system classifies the speech result from mixing and separate processes and computes the accuracy of each classifier.

7. Performance Measurements

Some measurements are used as evaluation metrics to calculate the system's performance.

- Signal-to-Noise Ratio (SNR): The most popular objective measurement used to evaluate speech quality (K. Kondo, 2012). The better range of SNR is between 0 to 1, the best value of SNR is close to 0 (Hussein M. Salman, Nidaa A. Abbas, 2021). It is computed as:

\[
\text{SNR} = 10 \log_{10} \frac{\sum_{n=-\infty}^{\infty} s^2(n)}{\sum_{n=-\infty}^{\infty} (s(n) - \hat{s}(n))^2} \text{ (dB)}
\]

The parameter s represents the source signals, and \( \hat{s} \) denote the recovered signals.

- Kurtosis: It is the classical method used as a measurement of a non-Gaussianity. It represents fourth-order cumulant. Kurtosis could be defined in three signs: + super gaussian, 0 Gaussian, and- subGaussian(A. Hyvarinen, 1999).

\[
\text{Kurtosis} = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2\right)^2} - 3
\]

Where \( x_i \) denotes \( i^{th} \) variable, \( \bar{x} \) is the mean, and \( n \) is the sample size

- Accuracy: accuracy of classification is equal to the proportion of correct predictions to the whole of input samples.

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP}
\]
Where:
TP: true positive.
FP: false positive.
TN: true negative.
FN: false negative.

- Precision: it is the proportion of correctly positive predictions concerning positive predictions.
  Precision=TP/ (TP+FP)  

- Recall: It is the proportion of true positives to overall actual positives in the samples.
  Recall=TP/ (TP+FN)

- Specificity: It is the ratio of correctly true negatives to total actual negatives in the samples.
  It is the opposite of recall.
  Specificity= TN/ (TN+FP)

- F1-score: It is the harmonic average(mean ) of both recall and precision. Overall, it is a measure of the preciseness and robustness of your model.
  F1-score=2TP / (2TP+FP+FN)

8. Experimental Work

The proposed system is evaluated through a series of experiments that test the mixing and separating processes and evaluate various learning methods. ITU-T voice dataset with wave format (mono speech 16KHz) is used for the gender classification process. The dataset includes 20 languages. There are sixteen samples of 8 files for each gender. Many pairs of speech signals were tested in the proposed system. These speeches will token from the dataset. The system simulates cocktailing by two different speeches, then injecting the results into a classification model. Depending on the subjective metric; table 2 illustrates the plotting of waveform for original, mixing, and de-mixing (separated) signals.

Table 2: Original, Mixing, and Separating Speech Signals

<table>
<thead>
<tr>
<th>Signal</th>
<th>Signal1</th>
<th>Signal2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td><img src="image1.png" alt="Waveform" /></td>
<td><img src="image2.png" alt="Waveform" /></td>
</tr>
</tbody>
</table>
The mixed-signal wave contains two voice signals representing two speakers presented in the second column of Table 2. The recovered signals are also shown in the last part of table 2.

In addition to the signals plotting, this paper used objective measurements as SNR and kurtosis to measure the performance of separating speech signals. SNR measures the amount of error (noise) in the recovering signal after the separation process. The best results of the SNR when nearby to 0.

**Table 3: Kurtosis, SNR values**

<table>
<thead>
<tr>
<th>Mixed Case No.</th>
<th>Files Names</th>
<th>Kurtosis of Source Signals</th>
<th>Kurtosis of mixing Signals</th>
<th>Kurtosis of de-mixing Signals</th>
<th>Length (samples)</th>
<th>SNR for de-mixing Signals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A_eng_f</td>
<td>25.724</td>
<td>12.344</td>
<td>25.727</td>
<td>192676</td>
<td>0.018981</td>
</tr>
<tr>
<td></td>
<td>A_eng_m</td>
<td>18.658</td>
<td>11.493</td>
<td>18.655</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>B_eng_f</td>
<td>9.540</td>
<td>3.349</td>
<td>9.543</td>
<td>86169</td>
<td>0.0201</td>
</tr>
<tr>
<td></td>
<td>B_eng_m</td>
<td>8.729</td>
<td>4.539</td>
<td>8.728</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Ch_f</td>
<td>9.4608</td>
<td>5.556</td>
<td>9.4604</td>
<td>105237</td>
<td>0.0301</td>
</tr>
<tr>
<td></td>
<td>Ch_m</td>
<td>22.2226</td>
<td>5.898</td>
<td>22.2228</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Da_f</td>
<td>8.4487</td>
<td>3.868</td>
<td>8.452</td>
<td>210531</td>
<td>0.0140</td>
</tr>
<tr>
<td></td>
<td>Da_m</td>
<td>10.9717</td>
<td>6.433</td>
<td>10.9723</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Fi_f</td>
<td>7.215</td>
<td>2.578</td>
<td>7.215</td>
<td>140283</td>
<td>0.0089</td>
</tr>
<tr>
<td></td>
<td>Fi_m</td>
<td>7.871</td>
<td>3.308</td>
<td>7.8742</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Ger_f</td>
<td>5.3920</td>
<td>2.0954</td>
<td>5.3929</td>
<td>83144</td>
<td>0.0206</td>
</tr>
<tr>
<td></td>
<td>Ger_m</td>
<td>6.897</td>
<td>2.7407</td>
<td>6.8966</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
As shown in Table 3, the SNR was nearby to 0 in all mixed cases. The columns of the kurtosis of the source signals and recovered signals show that the selected sources achieved the i.i.d. condition.

The performance reports of gender classification for the proposed system are given in table 4 and table 5. The performance can be evaluated by computing parameters like precision, accuracy, F1 score, recall, and specificity.

### Table 4: The performance metrics of classification for mixing speeches

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Precision</th>
<th>Recall</th>
<th>Specificity</th>
<th>F1_score</th>
<th>Accuracy on testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-NN</td>
<td>80.8%</td>
<td>86.9%</td>
<td>82.7%</td>
<td>83.7%</td>
<td>84.6%</td>
</tr>
<tr>
<td>SVM</td>
<td>82.9%</td>
<td>90.3%</td>
<td>84.4%</td>
<td>86.5%</td>
<td>87.1%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>80.9%</td>
<td>75.9%</td>
<td>85.0%</td>
<td>78.3%</td>
<td>80.8%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>81.7%</td>
<td>80.0%</td>
<td>85.0%</td>
<td>80.8%</td>
<td>82.7%</td>
</tr>
<tr>
<td>Neural Net</td>
<td>82.6%</td>
<td>82.1%</td>
<td>85.5%</td>
<td>82.4%</td>
<td>84.0%</td>
</tr>
</tbody>
</table>

From table 4, it is observed that both the SVM and K-NN give better results compared with other machine learning models. K-NN is giving 84.6% and SVM gives 87.1% accuracy with good precision and recall values on test sets in gender classification of mixing speeches.

### Table 5: The performance metrics of classification for de-mixing speeches

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Precision</th>
<th>Recall</th>
<th>Specificity</th>
<th>F1_score</th>
<th>Accuracy on testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-NN</td>
<td>95.1%</td>
<td>98.1%</td>
<td>95.0%</td>
<td>96.6%</td>
<td>96.5%</td>
</tr>
<tr>
<td>SVM</td>
<td>95.8%</td>
<td>100.0%</td>
<td>95.6%</td>
<td>97.8%</td>
<td>97.8%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>97.4%</td>
<td>95.6%</td>
<td>97.5%</td>
<td>96.5%</td>
<td>96.5%</td>
</tr>
</tbody>
</table>
According to the experimental results shown in table 5, the classification accuracy is above 96% for most machine learning models except SVM and ANN, which have better accuracy sets in gender classification of de-mixing speeches. The accuracy value of both models is equal to 97.8%.

9. Conclusion

This paper describes an experiment using the machine learning models to classify the speaker gender and evaluate the success of mixing and de-mixing speech classification. Classifying the gender of a speaker has been implemented based upon acoustic features of the speech and machine learning algorithms. The results obtained from this system clearly show that SVM is performing better on mixing speeches, and SVM and Neural Network are given better performance compared with all machine learning algorithms to classify gender using acoustic properties of de-mixing speeches.

References


