Journal of Engineering Science and Technology
Vol. 17, No. 2 (2022) 1119 - 1134
© School of Engineering, Taylor's University

SPEAKER GENDER IDENTIFICATION IN MATCHED AND MISMATCHED CONDITIONS BASED ON STACKING ENSEMBLE METHOD

AMEER A. BADR[1,2,*], ALIA K. ABDUL-HASSAN[2]

[1]College of Managerial and Financial Sciences, Imam Ja'afar Al-Sadiq
University, Salahaddin, Iraq
[2]Department of Computer Science, University of Technology, Baghdad, Iraq
[*]Corresponding Author: cs.19.53@grad.uotechnology.edu.iq

Abstract

Identifying the gender of the human voice has been considered one of the challenging tasks because it acts as a pre-processing ingredient for enhancing speech analysis applications. In this work, an automatic system is proposed to identify the speaker's gender without depending on the text in matched and mismatched conditions. Firstly, three groups of features are extracted from each utterance using Fundamental Frequency (F0), Fractal Dimensions, and Mel-Frequency Cepstral Coefficient (MFCC) methods. Then, the extracted feature dimensions are reduced using Linear Discriminant Analysis (LDA) method. Finally, the speaker's gender is identified based on proposed stacking ensemble classifier when Logistic Regression (LR), K-Nearest Neighbours (KNN) and Gaussian Naïve Bayes (GNB) are used as base classifiers, while Support Vector Machine (SVM) is used as meta classifier. Four experiments are conducted on two datasets: TIMIT, and Common-Voice. In matched conditions (i.e., same-language), the proposed system accuracy is 99.74%, 87.28% for the TIMIT, and the Common-Voice dataset, respectively. In mismatched conditions (i.e., cross-language), the proposed system shows a high ability to generalize, taking advantage of using the LDA method, where the system accuracy is 81.19%, 97.78% for the (TIMIT\Common-Voice), and (Common-Voice\TIMIT) datasets, respectively. The results also showed a clear superiority for the proposed system in comparison to related works that utilized the TIMIT dataset.

Keywords: Cross-language, Fractal dimensions, Features fusion, LDA, Speaker gender detection.

1. Introduction

The human voice consists of a human being's sound using the vocal cords to talk, laugh, cry, shout, and sing. Since it is the essential source of a sound, the human vocal cords play a major role in the conversation [1]. In addition to the content of a speech, a listener can understand several characteristics of a speech such as identity, accent, gender, emotional state, and age range [2]. Automatic recognition of this kind of speech characteristics can guide Human-Computer Interaction (HCI) systems to automatically adapt to different user needs [3].

Identifying a speaker's gender information, given a short speech, is a difficult task and is a rapidly emerging field of research due to the increasing interest in interaction applications, such as natural spoken dialog systems and HCI. Moreover, such information may also serve as an important analytical feature for decision making. Speaker gender determination can be helpful in many applications, including healthcare, Human-Robot Interaction (HRI), and education [3-5]. Gender identification can play a major role in pre-processing methods by enhancing the accuracy of some speech-based recognition models [2, 6]. An example of such methods is the use of the voice-based gender detection model as a pre-process to voice-based age estimation systems, where it is almost impossible to estimate a human age range through his/her voice unless gender is identified first, as in [7-9].

The performance of gender identification systems is influenced by many factors. The content for the input speech can be text-independent or text-dependent. Selecting the feature sets to use in the model representation is another factor [5]. On the other hand, the use of suitable classifiers is an essential part of the speaker's gender identification systems. An appropriate selection of a classifier is as important as the feature extraction [4]. Combining classifiers to achieve higher accuracy is an important research topic with different names, such as a combination of multiple classifiers, classifier ensembles, and classifier fusion. The combination of different base classifiers can provide additional information regarding unknown examples in the ensemble learning process. It is known that this kind of solution can be used from the accuracy perspective as well as a generalization to improve the overall classification. Wolpert [10] proposed a layered architecture, named as Stacking Ensemble. Lowest level classifiers get the training data as their input to output their subproblem prediction. Successive layers gather predictions of the previous layer as input and a single top-level classifier forms the final prediction [11, 12].

The primary contributions of the present work are highlighted and summarized as follows:

- Build a robust classification method by using the Stacking Ensemble technique in which Logistic Regression (LR), K-Nearest Neighbour (KNN), and Gaussian Naïve Bayes (GNB) are the base classifier and Support Vector Machine (SVM) is the final classifier.
- Combining three features extraction methods which are Fundamental Frequency (F0), Fractal Dimensions, and Mel-Frequency Cepstral Coefficient (MFCC) to extract 60-dimensional informative feature vectors.
- Explore the role of using Linear Discriminant Analysis (LDA) as a dimensionality reduction method to enhance the performance of the proposed system especially in mismatched conditions.

The remainder of this work is organized as follows. Related works of the proposed system are presented in Section 2. The theoretical background of the Stacking ensemble concept and feature extraction methods are reviewed in Section 3. In Section 4, the proposed methodology is discussed. Section 5 demonstrates the simulation results and experiments. Finally, the work conclusions and future works are shown in Section 6.

2. Related Works

There is a lot of previous works that concern the study of speaker gender identification in addition to ensemble-based classification methods.

Zeng et al. [13] presented a speaker gender classification system based on Gaussian Mixture Model (GMM) as classifier, voice pitch and Relative Spectral Perceptual Linear Predictive (RASTA-PLP) as features. Their proposed system performance evaluated on the conditions of clean speech, noisy speech, and multi-language by using the TIMIT dataset. Their simulations show that the proposed system performance was excellent with about (98%) recognition rate; it is very robust for noise and completely independent of languages. Sedaaghi [14] presented a comparative study of gender and age classification in speech signals. The performance evaluation has been conducted on two datasets using five machine learning techniques. The best recognition rate (95%) in gender identification has been achieved for the SVM classifier with the polynomial kernel. Yücesoy and Nabiyev [2] developed a text-independent speaker gender identifying system. The proposed system was based on the classification of MFCC coefficients obtained from speech signals with the GMM; their experiments were conducted on the TIMIT dataset. Their system achieved about (97%) recognition rate, to ensure high accuracy, they increase the number of features as well as the number of mixtures, used in their proposed system. Chen [15] examined the applicability of standard machine learning techniques to the voice-based gender identification problem. Their system achieved about (88%) recognition rate. Only clear utterances from the TIMIT dataset have been used in this work.

Alhussein et al. [16] stated that the contribution of the vocal folds is very vital in the human voice production system. Gender is dependent on the vocal folds' length; a female speaker has shorter vocal folds than a male speaker. The voice of a male gets heavier because of longer vocal folds and thus contains more vocal intensity. Based on this idea, a new type of acoustic time-domain feature has been proposed for the speaker gender recognition system. Their experiments were demonstrated on TIMIT and Arabic datasets achieved (98%, 100%) recognition rate, respectively. Gupta et al. [17] proposed an ensemble Stacked machine learning algorithm to determine speaker gender using the acoustic parameters of voice sample and compares its performance with existing classifier techniques. Their system achieved a (96%) recognition rate. Kanani et al. [18] presented attempts to observe the impact on gender recognition systems of various short-term spectral features with varying dimensions. Their experiments were conducted on ELSDSR and SITW datasets. In matched conditions, their system achieved good recognition rate. While, in mismatched conditions, their results were degraded, and their system need to be enhanced further. Livieris et al. [19] presented a speaker gender recognition system using an ensemble semi-supervised self-labelled algorithm. Their experiments proved the classification efficiency of the proposed algorithm in terms of accuracy (97%), conducting robust and stable predictive models.

Through reviewing related works, there are several topics such as creating systems obtaining high accuracy, creating a noise-robust system, and creating systems can work in cross-corpus scenario, but with limited accuracy. Unlike all related works, this work will focus on several topics at the same time, which are ensuring high accuracy through the use of Stacking Ensemble method, the use of cross-language datasets with a high system ability to generalize and clarify the role of using the LDA method in results improvement.

3. Theoretical Background

In this work, several methods and techniques have been used to extract the features form each speech utterance, reduce the extracted features dimensionality, and classify these features to identify the gender from speech. these methods and techniques are described briefly below.

3.1. Features extraction methods

Speech signal contains various types of information, like speaker identity, speaker age, speaker gender, and speaker emotional state. Features are determined at the first stage of all identification systems, where the speech signal is transformed into measured values with distinguishing characteristics.

Among all types of speech-based feature extraction domains, Cepstral domain features are the most successful ones, where a cestrum is obtained by taking the inverse Fourier transform of the signal spectrum. MFCC is the most important method to extract speech-based features in this domain [2, 20]. MFCCs greatness stems from the ability to exemplify the spectrum of speech amplitude in a concise form. A speaker's voice is filtered by the articulator form of the vocal tract, such as the nasal cavity, teeth, and tongue. This shape affects the vibrational characteristics of the voice. If the shape is precisely controlled, this should give an accurate depiction of the phoneme being formed. The vocal tract shape is reflected in the shorter time power spectrum envelope, and the MFCCs aims to present this envelope accurately. The calculation steps for MFCC features are shown in Fig. 1. At the end of these steps, one energy and 12 cepstral features are obtained [5, 20].

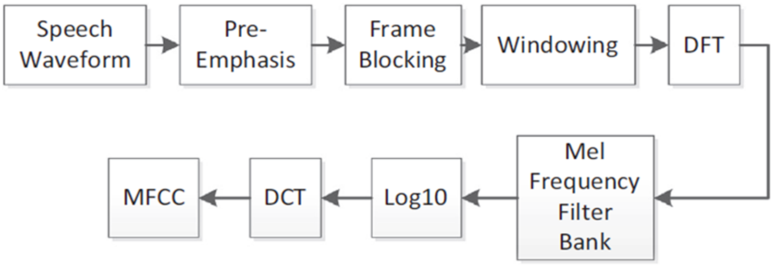

Fig. 1. Mel frequency cepstral analysis [5].

The F0 feature is proven biologically and perceptually as a good discriminator between the voices of males and females. In general, female speech has a higher pitch than male speech, which could, therefore, be used to differentiate male speech from female speech if a precise pitch estimate could be calculated. There is, however, a natural overlap of the pitch values between males and females' voices, thus intrinsically limiting the capacity of the pitch feature for gender identification [21, 22]. F0 used for gender identification in [22]. Among all methods proposed for F0 estimation, the YIN algorithm presented a promising result. The algorithm details can be found in [23].

The fractal dimension is a measure for analysing the complexity of the data. The fractal dimension, unlike the Euclidean dimension, which is a natural number, can be a real number. Fractal geometry expresses an iterative model in objects (audio signal). In other words, if an object is divided into smaller parts, every smaller part is a copied version of the original object, this attribute is called self-similarity. Fractals are mathematically based on frequent permutations in a recursive mathematical formula that generates a geometrically fractal model through numerous iterations [24]. Since the fractal dimension can describe the fragmentation, irregularity, and self-similarity of the speech signal, different signal patterns should yield different fractal dimension values. Fractal dimensions-based features have been used in speech processing lastly, like in speech emotion recognition [25], and speech activity detection [26]. In this work, three different algorithms have been used to extract features based on fractal dimensions, which are: Katz [27], Higuchi [28], and Goh et al. [29]. More details about these algorithms can be found in [25].

3.2. Dimensionality reduction

Predominantly, there is some redundancy in the extracted high-dimensional features. Subspace learning can be used to eliminate these redundancies by further processing extracted features to reflect their semantic information better. Among all dimensionality reduction techniques, LDA is a very common supervised technique for dimensionality reduction problems as a pre-processing step for machine learning and pattern classification applications. The LDA technique aims at projecting the original data matrix in a lower-dimensional space. Three steps were needed to achieve this aim. The first step is to calculate the between-class variance (i.e., the distance between the means of different classes). The second step is to calculate the within-class variance (i.e., the distance between the mean and the samples of each class). The third step is to construct the lower-dimensional space which minimizes the within-class variance and maximizes the between-class variance [30]. The supervised LDA algorithm can be found in [30].

3.3. Classification schemes

In pattern recognition tasks, the classification methods make its decision depend on the number of classes and the similarity of features. For a given input, these classification algorithms build mathematical models to generate a desirable set of outputs. By using a subset of data with valid class labels, the model is trained to generate predicted outputs for specific inputs. Then, by using the test data, model performance is validated [4].

SVM has recently become a prevalent classifier due to its promising performance in various studies. Finding a classifier that minimizes the expected error limits is the major aim of SVM. SVM uses a two-step classification process. A kernel function performs a transformation of the feature from low to high dimensions in the first step. This transformation allows for linear separation at a higher dimension of non-linearly separable data. Secondly, it forms an optimum hyperplane to draw the boundary of the decision between classes. SVM is considered as one of the most accurate and robust methods among classification algorithms due to its use of optimum separation to prevent misclassification of outliers [4, 5, 31]. SVM has been identified as the top 10 classification algorithms in [32]. To guarantee that hyperplanes with the maximum margin are found, an SVM classifier tries to maximize the Eq. (1). in terms of $\vec{w}$ and b [32]:

$$L_P = \frac{1}{2}\, \|\vec{w}\| - \sum_{i=1}^{t} \alpha_i\, y_i (\vec{w}.\vec{x} + b) + \sum_{i=1}^{t} \alpha_i \tag{1}$$

where t is the training examples number, and α_i, i = 1,... ., t, are non-negative numbers such that the L_P derivatives as regard to α_i are zero. L_P is called the Lagrangian where α_i are the Lagrange multipliers. In Eq. (1), the hyperplane defined by the vectors $\vec{w}$ and constant b.

KNN, one of the simplest non-parametric classifiers in machine learning yet, provides good performance in classification; it is based on examples of the closest training in the feature space. KNN classification tries to find in the training set a group of k objects which are closest to the test object. Based on the distance of this object to the labelled objects, KNN uses the Euclidean distance metric to classify an unlabelled object [33, 34]. KNN is one of the top 10 classification algorithms as identified in [7]. Let on has a D training set and a $z = (x', y')$ test object, the distance between z and all the training objects $(x, y) \in D$ is calculated by the algorithm to determine its nearest-neighbour list, D_z. (x is the training object data, whereas $\boldsymbol{y}$ is its class. Equally, x' is the test object data and y' is its class). After the list of nearest neighbors is gained, according to the majority class of its nearest neighbor, the test object is classified as seen in Eq. (2). [32]:

$$Majority\ Voiting\colon\ y' = argmax_v \sum_{(x_i, y_i) \in D_z} I(v = y_i) \tag{2}$$

where I (·) is an indicator function that returns 1 if its argument is true and 0 otherwise, v is a class label, and y_i is the class label for the i^{th} nearest neighbours.

By using the framework of Bayes' theorem, GNB tries to classify observations into one of a pre-defined set of classes based on information supplied by predictor variables. GNB classifier does not take into account the covariance among the predictor variables under the assumption that the predictor variables are class-conditionally independent. GNB tries to estimate a separate Gaussian distribution for each predictor class, and observations are allocated to the class with the maximum posterior probability [35, 36]. Suppose the training data contains a continuous attribute, x. one can first segment the data by the class, and then compute the mean (μ_k) and stander deviation (σ_k) of x in each class (C_k). If some observation value v has been collected, Then, the probability distribution of v given a class C_k can be computed using Eq. (3) [37]:

$$p(x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}}\, e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}} \tag{3}$$

LR has been proven to be highly reliable and accurate among most statistical methods. In this model, the dependent variables are predicted by the independent variables. The dependent variable is in a binary format, while independent variables can be measured on a nominal, ordinal, or ratio scale. Despite LR being based on different assumptions as to the relationship of the dependent-independent variables, LR is considered as a special case of a linear regression model. LR conditional distribution is a Bernoulli distribution rather than a Gaussian distribution since the dependent variable has the form of a binary variable [38, 39]. The relationship between the occurrence and its dependency on several variables in logistic regression analysis can be expressed by the Eq. (4). below [38]:

$$P = \frac{1}{1+ e^{-(b_0+b_1x_1+b_2x_2+\cdots+b_nx_n)}} \tag{4}$$

where p is the occurrence probability, b_0 is the intercept of the model, the b_i (i = 0, 1, 2, ..., n) is the slope coefficients of the logistic regression model, and x_i (i = 0, 1, 2, ..., n) are the independent variables.

3.4. Stacking ensemble

The main goal of the Stacking Ensemble method is the production of a strong, high-level learner with high generalized performance. The stacking ensemble learning consists of two levels: base-classifier and meta-classifier as seen in Fig. 2. In the base-classifier level, the training set is adopted for training models and making predictions. The individual outputs of base-classifier, along with the training labels, are used to train a second level meta-classier. This second meta-classifier serves to predict the final decision for an input, given the decisions of the base-classifiers [12, 40, 41].

Stacking Ensemble meta-classifier tries to learn and exploit patterns and regularities from the collective knowledge represented by Stacking Ensemble base-classifier outputs. Hence, the meta-classier attempts to correct base-classifiers biases by learning how they commit errors. For training, the base-classifiers outputs under cross-validation can be the input for the meta-classifier [41].

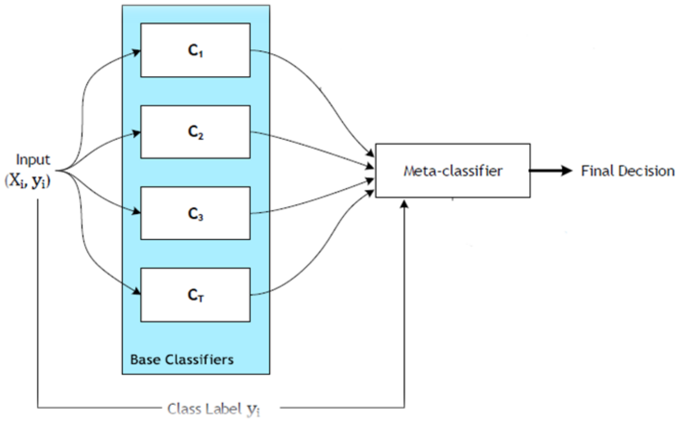

Fig. 2. Stacking ensemble classifier with two-level learning [41].

4. The Proposed Speaker Gender Identification System

As can be seen from Fig. 3, the methodology of this work consists of three main stages: features extraction and normalization, dimensionality reduction, and gender identification. Initially, five groups of features will be extracted from each speaker's utterance, followed by features scaling to fall within a smaller range using the z-score techniques. Then, by using the LDA method, the high dimensional features will be transformed into more meaningful low dimensional features. Finally, a staking ensemble classifier is used to predict the speaker's gender.

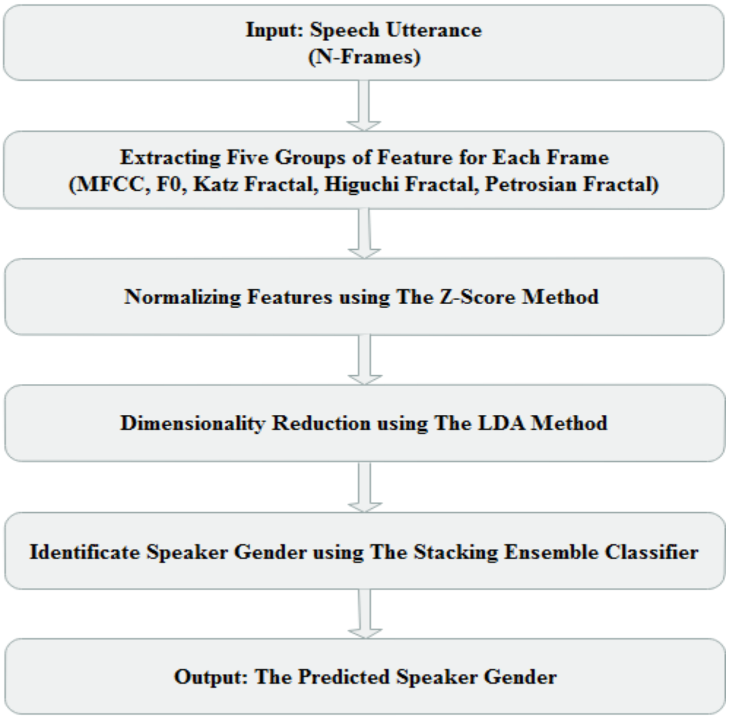
Input: Speech Utterance
(N-Frames)
Extracting Five Groups of Feature for Each Frame
(MFCC, F0, Katz Fractal, Higuchi Fractal, Petrosian Fractal)
Normalizing Features using The Z-Score Method
Dimensionality Reduction using The LDA Method
Identificate Speaker Gender using The Stacking Ensemble Classifier
Output: The Predicted Speaker Gender

Fig. 3. The general framework of the proposed system.

4.1. Utterance based features extraction and normalization

Five groups of features were incorporated in this work in which the errors of the prediction from the variety of the feature groups are complementary. In the beginning, each speaker's utterance is split into frames with a window size of 25 milliseconds and a frameshift of 10 milliseconds to ensure that each frame contains robust information. Then, five groups of features are extracted from each utterance frame, which are MFCC, F0, Katz fractal, Higuchi fractal, and Petrosian fractal. Each group includes twelve feature dimensions, which are in total 60-dimensional features per frame as describe in Table 1.

The expression of features in smaller units will result in a wider range for these features and will, therefore, tend to give these features greater effect. Normalization involves transforming the data to fall within a smaller range, such as (-1, 1) [42]. Therefore, due to the great usefulness of the normalization process in classification methods, the 60-dimensional extracted features will be normalized by using the *z*-score method as given by Eq. (5) [42]:

$$z = \frac{x_i - \mu}{\sigma} \tag{5}$$

where, x_i is the feature vectors, μ , σ are the mean and standard deviation, respectively, of the x_i.

Table 1. The proposed feature dimensions.

Features Group	Feature Dimensions	No. Dimensions
MFCC	Arithmetic means for MFCC1 to MFCC12	12
F0	Arithmetic mean, geometric mean, harmonic mean, maximum, minimum, median, first quantile, third quantile, variance, standard deviation, skewness, kurtosis	12
Katz Fractal	Arithmetic mean, geometric mean, harmonic mean, maximum, minimum, median, first quantile, third quantile, variance, standard deviation, skewness, kurtosis	12
Higuchi Fractal	Arithmetic mean, geometric mean, harmonic mean, maximum, minimum, median, first quantile, third quantile, variance, standard deviation, skewness, kurtosis	12
Petrosian Fractal	Arithmetic mean, geometric mean, harmonic mean, maximum, minimum, median, first quantile, third quantile, variance, standard deviation, skewness, kurtosis	12

4.1. Dimensionality reduction using the LDA method

At this stage, LDA takes as its input a set of 60-dimensional normalized features grouped into classes. Then, it is finding an optimal transformation that maps these input features into a lower-dimensional space while preserving the class structure. The output of this stage is a 1-dimensional feature vector contain the most important information to distinguish between male and female from their voice.

4.2. Gender identification using stacking ensemble classifier

In stacking ensemble methods, all machine learning techniques can be a base-classifier. The selection of base-classifier is based on two principles: high diversity between different base-classifiers and low complexity of base-classifier. Based on the above two principles, in the proposed stacking ensemble classifier, three state-of-the-art classifiers are considered as base classifiers: KNN (i.e., *k*=5), GNB, and LR (i.e., 5000 iterations). On the other hand, and due to its great importance and superiority in the field of classification, SVM (i.e., RBF kernel) will be used as a meta-classifier. The proposed stacking ensemble classifier is shown in Fig. 4.

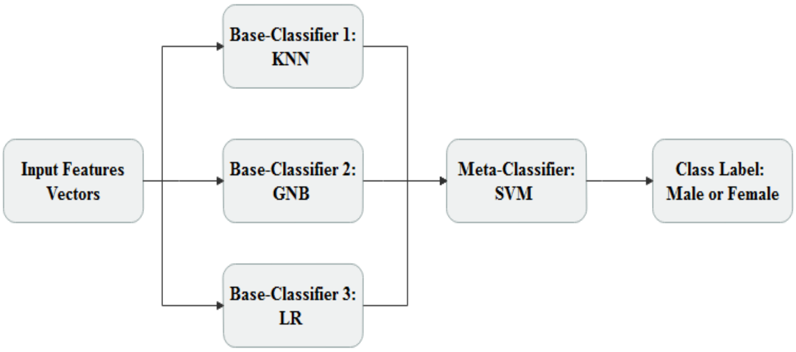

Fig. 4. The proposed stacking ensemble classifier.

5. Experimental Results and Discussions

The datasets used in this work are described in detail in this section, and also the results of the experiments are explained and discussed. The performance of the proposed gender identification system is carried out by using the following parameters and equations [16, 42]:

- True positive (TP): the male speaker detected by the system as a male.
- True negative (TN): the female speaker detected by the system as a female.
- False positive (FP): the female speaker detected by the system as a male.
- False negative (FN): the female speaker detected by the system as a male.

$$precision = \frac{TP}{TP+FP} \tag{6}$$

$$recall = \frac{TP}{TP+FN} \tag{7}$$

$$F\_score = \frac{2 \times precision \times recall}{precision+recall} \tag{8}$$

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \tag{9}$$

5.1. Dataset discerption

In this work, two datasets have been used, which are The Common Voice corpus [43], and The DARPA TIMIT acoustic-phonetic continuous speech corpus [44]. One of the aims of this work is to evaluate the performance of these two datasets on matched and mismatched conditions.

The TIMIT corpus of reading speech was designed to develop and evaluate automatic speech recognition systems. Text corpus design was a joint effort among the Stanford Research Institute (SRI), Massachusetts Institute of Technology (MIT), and Texas Instruments (TI). The speech was recorded at TI and transcribed at MIT. The sampling frequency of recoded utterances is chosen to be 16-kHz with a 16-bit rate [44]. TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. In this work, the same number for male and female speakers, which is 192 from all 8 regions has been chosen. In total, the number of utterances has been used in this work is 1920 for males, and 1920 for females.

The Common Voice corpus is a massively multilingual collection of transcribed speech intended for speech technology research and development. Common Voice is designed for Automatic Speech Recognition purposes but can be useful in other domains like language identification, age-group recognition, and gender identification because of its wide range labelling metadata. The most recent release includes a total of 38 languages collecting data. Over 50,000 individuals have participated so far, resulting in 2,500 hours of collected audio [43]. In this work, the Arabic language corpus has been chosen. The number of text-independent utterances used is 5,885 for males, and 2,843 for females. The sampling frequency of recoded utterances is chosen to be 48-kHz with a 16-bit rate.

5.2. Results and discussions

A set of experiments are conducted to show the performance of the proposed speaker gender identification system, which includes performance evaluation in matched as well as mismatched conditions, studying the influence of using LDA as supervised dimensionality reduction technique on system accuracy in cross-language datasets, and compare results with other works. All experiments are conducted on two datasets which are TIMIT and Common voice. The time complexity of the system when the training part (80%) and the testing part (20%) is 0.34 sec and 2.86 sec for TIMIT and Common voice datasets, respectively.

A. Performance evaluation in matched conditions

The performance of the proposed system is evaluated in the matched condition in terms of precision, recall, F-score, accuracy and confusion matrix. This experiment is also showing the influence of data training size in the proposed stacking ensemble classifier. Tables 2 and 3 show the results of this experiment on TIMIT and Common voice datasets.

Table. 2. The performance evaluation in matched conditions of the proposed system on TIMIT and common voice datasets.

Dataset	Train Size (%)	Test Size (%)	Gender	Precision (%)	Recall (%)	F-Score (%)	Acc. (%)
TIMIT	80	20	Male Female	100 99.51	99.48 100	99.74 99.75	99.74
	66	33	Male Female	99.84 99.24	99.22 99.85	99.53 99.54	99.53
	50	50	Male Female	99.26 99.41	99.37 99.31	99.31 99.36	99.33
	5-Folds		-	99.34	99.08	99.20	99.21
Common Voice	80	20	Male Female	88.85 83.56	92.78 75.92	90.77 79.56	87.28
	66	33	Male Female	89.21 82.57	92.11 77.05	90.64 79.71	87.18
	50	50	Male Female	88.17 83.76	92.84 74.76	90.45 79.00	86.86
	5-Folds		-	88.55	91.74	90.08	86.30

Table. 3. The confusion matrix (%) of the proposed system in matched conditions on TIMIT and common voice datasets.

TIMIT Dataset			Common-Voice Dataset		
	F	**M**		**F**	**M**
F	100	0	**F**	75.92	24.07
M	0.52	99.47	**M**	6.62	93.37

As seen in Tables 2 and 3, the effectiveness of the proposed system is evaluated using five measures: precision, recall, F-score, accuracy, and confusion matrix. The first four measures are applied to different sizes of training data; they are 80%, 66%, and 50%, in addition to 5-folds cross validation. In the TIMIT dataset, the F-score of the proposed system for males is almost equal to the F-score of the females for all sizes of training data and that because of using a balanced number of utterances. On the other hand, the overall accuracy of the proposed system was affected slightly by the size of training data, as it decreased from 99.74% when the training size is 80% to 99.33% when the training size became 50%. In the Common Voice dataset, the F-score of the proposed system for males is higher than the F-score of the females for all sizes of training data and that because of using an unbalanced number of utterances. On the other hand, the overall accuracy of the proposed system is affected slightly by the size of training data, as it decreased from 87.28% when the training size is 80% to 86.86% when the training size is 50%.

B. Performance evaluation in mismatched conditions

The performance of the proposed system is evaluated in the mismatched condition in terms of precision, recall, F-score, accuracy, and confusion matrix. The mismatched conditions include an unbalanced number of utterances, different languages, and different sampling frequencies. Tables 4 and 5 show the results of this experiment on TIMIT and Common Voice datasets.

As seen in Tables 4 and 5, the effectiveness of the proposed system is evaluated using five measures: precision, recall, F-score, accuracy and confusion matrix. All these measures are applied in a cross-corpus situation when one dataset is used for training, the other is used for testing, and vice versa. In the first situation, the proposed system has been trained on the TIMIT dataset and tested on the Common Voice dataset. The F-score of the proposed system for males is higher than the F-score of the females and that because of using an unbalanced number of utterances in testing. On the other hand, the overall accuracy of the proposed system looking good given the mismatched conditions like language, dataset size, and sampling frequency. In the second situation, the proposed system has been trained on the Common Voice dataset and tested on the TIMIT dataset. The F-score of the proposed system for males is almost equal to the F-score of the females and that because of using a balanced number of utterances in testing. On the other hand, the overall accuracy of the proposed system looking perfect which demonstrates the high ability of the proposed system to generalize.

Table. 4. The performance evaluation in mismatched conditions of the proposed system on TIMIT and common voice datasets.

Train Dataset	Test Dataset	Gender	Precision (%)	Recall (%)	F-Score (%)	Acc. (%)
TIMIT	**Common Voice**	Male	94.52	76.55	84.59	81.19
		Female	65.16	90.82	75.88	
Common Voice	**TIMIT**	Male	95.84	99.90	97.83	97.78
		Female	99.89	95.66	97.73	

Table. 5. The confusion matrix (%) of the proposed system in mismatched conditions on TIMIT and common voice datasets.

TIMIT\ Common-Voice			Common-Voice\ TIMIT		
	F	**M**		**F**	**M**
F	90.81	9.18	**F**	95.66	4.33
M	23.45	76.54	**M**	0.10	99.89

C. The effects of using LDA as dimensionality reduction method

This experiment is designed to assess the effects of using LDA as a dimensionality reduction method on the proposed system accuracy in mismatched conditions. Table 6 shows the results of this experiment on Cross-language datasets.

As seen in Table 6, there is a significant impact of using LDA as a dimensionality reduction method on the accuracy of the proposed system in this experiment. Hence, the accuracy of the proposed system increased from 68% to about 81% in (TIMIT \ Common Voice) cross-language datasets. As well as the accuracy of the system increased from 77% to about 98% in (Common Voice \ TIMIT) cross-language datasets.

Table. 6. The Effect of the LDA on the accuracy of the proposed system in mismatched conditions.

Cross-Language Dataset (Train\Test)	Accuracy (%)	
	Without LDA	With LDA
TIMIT \ Common Voice	68.50	81.19
Common Voice \ TIMIT	77.16	97.78

D. Comparison with related works

A comparative study in terms of overall accuracy between the proposed system and several recent works in the same field has been presented. Table 7 shows the comparison process between the proposed system and four other works that depended on the TIMIT dataset. On the other hand, the comparative study dependent on the Common Voice dataset is unavailable because this is the first use of this dataset in such a system.

As seen in Table 7, the proposed system showed a clear superiority in terms of overall accuracy when compared with several other works for the same field dependent on TIMIT datasets.

Table. 7. Accuracy based comparative study between the proposed system and several works in the same field conducted on TIMIT dataset.

Authors	Accuracy (%)
Chen [15]	88.00
Yücesoy and Nabuyev [2]	97.76
Zeng et al. [13]	98.00
Alhussein et al. [16]	98.27
Propose System	**99.74**

5. Conclusion and Future Works

In this work, an automatic system is proposed to identify the speaker gender in matched and mismatched conditions. Firstly, five groups of features are combined to further improve system performance. After that, the 60-dimensional extracted features are reduced into a 1-dimensional informative feature using the LDA method. Finally, the stacking ensemble classifier gives the proposed system the classification power taking advantage of the base-classifiers as well as meta-classifier strength. The experimental results show the efficiency of the proposed system in both matched and mismatched conditions. The accuracy of the proposed system in matched condition is 99.74%, 87.28% for the TIMIT, and the Common-Voice dataset, respectively. The use of the LDA method has a major impact on the efficiency of the system in mismatched conditions with accuracy of 81.19%, 97.78% for the (TIMIT\Common-Voice), and (Common-Voice\TIMIT) datasets, respectively. For future work, a deep neural network may be utilized for gender identification from speech utterances where the network can be fed with the same reduced feature vectors proposed in this work.

References

1. Kumar, P.; Baheti, P.; Jha, R.; Sarmah, P; and Sathish, K. (2018). Voice gender detection using gaussian mixture model. *Journal of Network Communications and Emerging Technologies* (*JNCET*), 8(4), 132-136.
2. Yücesoy, E.; and Nabiyev, V. (2013). Gender identification of a speaker using MFCC and GMM. 8*th International Conference on Electrical and Electronics Engineering* (*ELECO*). Bursa, Turkey, 626-629.
3. Li, M.; Jung, C.; and Han, K. (2010). Combining five acoustic level modelling methods for automatic speaker age and gender recognition. 11*th Annual Conference of the International Speech Communication Association*. Chiba, Japan, 2826-2829.
4. Ahmad, J.; Fiaz, M.; Kwon, S.; Sodanil, M.; Vo, B.; and Baik, S. (2015). Gender identification using MFCC for telephone applications-A comparative study. *International Journal of Computer Science and Electronics Engineering* (*IJCSEE*), 3(5), 351-355.
5. Barkana, B.; and Zhou, J. (2015). A new pitch-range based feature set for a speaker's age and gender classification. *Applied Acoustics*, 98, 52-61.
6. Alipoor, G.; and Samadi, E. (2018). Robust speaker gender identification using empirical mode decomposition-based cepstral features. *Asia-Pacific Journal of Information Technology and Multimedia*, 7(1), 71-81.
7. Pribil, J.; Pribilova, A.; and Matousek, J. (2017). GMM-based speaker age and gender classification in Czech and Slovak. *Journal of Electrical Engineering*, 68(1), 3-12.
8. Přibil J.; Přibilová, A.; and Matoušek, J. (2016). GMM-based speaker gender and age classification after voice conversion. *First International Workshop on Sensing, Processing and Learning for Intelligent Machines* (*SPLINE*). Aalborg, Denmark, 1-5.
9. Erokyar, H. (2014). *Age and gender recognition for speech applications based on support vector machines*. MS Thesis. Electrical Engineering. University of South Florida.
10. Wolpert, D. (1992). Stacked generalization. *Neural Networks*, 5, 241-259.
11. Hatami, N.; and Ebrahimpour, R. (2007). Combining multiple classifiers: diversify with boosting and combining by stacking. *IJCSNS International Journal of Computer Science and Network Security*, 7(1), 127-131.
12. Jiang, W.; Chen, Z.; Xiang, Y.; Shao, D.; Ma, L; and Zhang, J. (2019). SSEM: A novel self-adaptive stacking ensemble model for classification. *IEEE Access*, 7, 120337-120349.
13. Zeng, Y.; Wu, Z.; Falk, T.; and Chan, W. (2006). Robust GMM based gender classification using pitch and Rasta-PLP parameters of speech. *Proceedings of the Fifth International Conference on Machine Learning and Cybernetics*. Dalian, China, 3376-3379.

14. Sedaaghi, M. (2009). A comparative study of gender and age classification in speech signals. *Iranian Journal of Electrical and Electronic Engineering*, 5(1), 1-12.
15. Chen, K. (2014). Gender identification by voice. Stanford University, CS229.
16. Alhussein, M.; Ali, Z.; Imran, M.; and Abdul, W. (2016). Automatic gender detection based on characteristics of vocal folds for mobile healthcare system. *Mobile Information Systems*, Volume 2016, Article ID 7805217.
17. Gupta, P.; Goel, S.; and Purwar, A. (2018). A stacked technique for gender recognition through voice. *Proceedings of* 2018 *Eleventh International Conference on Contemporary Computing* (*IC*3). Noida, India, 1-3.
18. Kanani, I.; Shah, H.; and Mankad, S. (2019). On the performance of cepstral features for voice-based gender recognition. *Information and Communication Technology for Intelligent Systems*. Singapore, 327-333.
19. Livieris, I.; Pintelas, E.; and Pintelas, P. (2019). Gender recognition by voice using an improved self-labelled algorithm. *Machine Learning and Knowledge Extraction*, 1(1), 492-503.
20. Badr, A.A; and Abdul-Hassan, A.K. (2020). A review on voice-based interface for human-robot interaction. *Iraqi Journal for Electrical and Electronic Engineering*, 16(2), 91-102.
21. Harb, H.; and Chen, L. (2005). Voice-based gender identification in multimedia applications. *Journal of Intelligent Information Systems*, 24, 179-198.
22. Ting, H.; Yingchun, Y.; and Zhaohui, W. (2006). Combining MFCC and pitch to enhance the performance of the gender recognition. 8*th international Conference on Signal Processing*. Guilin, China, 9464328.
23. Cheveigné, A.D.; and Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4), 1917-30.
24. Mohtasham-zadeh, V.; and Mosleh, M. (2018). Audio Steganalysis based on collaboration of fractal dimensions and convolutional neural networks. *Multimedia Tools and Applications*, 78, 11369-11386.
25. Tamulevičius, G.; Karbauskaite, R.; and Dzemyda, G. (2019). Speech emotion classification using fractal dimension-based features. *Nonlinear Analysis-Modelling and Control*, 24(5), 679-695.
26. Shafiee, S.; Almasganj, F.; Vazirnezhad, B.; and Jafari, A. (2010). A two-stage speech activity detection system considering fractal aspects of prosody. *Pattern Recognition Letter*, 31(9), 936-948.
27. Katz, M.J. (1988). Fractals and the analysis of waveforms. *Computers in Biology and Medicine*, 18(3), 145-156.
28. Higuchi, T. (1988). Approach to an irregular time series on the basis of the fractal theory. *Physica D: Nonlinear Phenomena*, 31(2), 277-283.
29. Goh, C.; Hamadicharef, B.; Henderson, G.; and Ifeachor, E. (2005), Comparison of fractal dimension algorithms for the computation of EEG biomarkers for dementia. 2*nd International Conference on Computational Intelligence in Medicine and Healthcare* (*CIMED*2005). Lisbon, Portugal, 464-471.
30. Tharwat, A.; Gaber, T.; Ibrahim, A.; and Hassanien, A.E. (2017). Linear discriminant analysis: A detailed tutorial. *AI Communication*, 30(2), 169-190.

31. Vapnik; and Vladimir, (1995). *The nature of statistical learning theory*. (1st ed.). New York: Springer-Verlag.
32. Wu, X.; Kumar, V.; Quinlan, J.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G.; Ng, A.; Liu, B.; Yu, P.; Zhou, Z.; Steinbach, M.; Hand, D.; and Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14, 1-37.
33. Goua, J.; Mab, H.; Ouc, W.; Zengd, S.; Raoe, Y; and Yan, H. (2018). A generalized mean distance-based k-nearest neighbour classifier. *Expert Systems with Applications*, 115, 356-372.
34. Prasatha, V.; Alfeilat, H.A.; Hassanat, A.; Lasassmeh, O.; Tarawneh, A.; Alhasanat, M.; and Salman, H. (2019). Effects of distance measure choice on KNN classier performance - A review. *Big Data*, 7(4), 221-248.
35. Bi, Z.; Han, Y.; Huang, C.; and Wang, M. (2019). Gaussian naive Bayesian data classification model based on clustering algorithm. *International Conference on Modelling, Analysis, Simulation Technologies and Applications* (*MASTA*), Hangzhou, China, 396-400.
36. Ontivero-Ortega, M.; Lage-Castellanos, A.; Valente, G; Goebel, R.; and Valdes-Sosa, M. (2017). Fast Gaussian Naïve Bayes for searchlight classification analysis. *NeuroImage*, 163, 471-479.
37. Abdulah, H.; and Al-Tuwaijari, J. (2019). Cancer classification using gaussian naive bayes algorithm. 2019 *International Engineering Conference* (*IEC*). Erbil, Iraq, 165-170.
38. Tsangaratos, P.; and Ilia, I. (2016). Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of model's complexity and training dataset size. *Catena*, 145, 164-179.
39. Jacob, A. (2017). Modelling speech emotion recognition using logistic regression and decision trees. *International Journal of Speech Technology*, 20(3), 897-905.
40. Alexandropoulos, S.; Aridas, C.; Kotsiantis, S.; and Vrahatis, M. (2019). Stacking strong ensembles of classifiers. *Artificial Intelligence Applications and Innovations*. Crete, Greece, 545-556.
41. Malmasi, S. and Dras, M. (2017). Native language identification using stacked generalization. *ArXiv abs*/1703.06541.
42. Han, J.; Kamber, M.; and Pei, J. (2012). *Data mining concepts and techniques*. 3rd Edition. USA: Morgan Kaufmann Publishers.
43. Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F.M.; and Weber, G. (2020). Common voice: A massively multilingual speech corpus, *arXiv:*1912.06670*v*2.
44. Garofolo, J.; Lamel, L.; Fisher, W.; Fiscus, J.; Pallett, D.; and Dahlgren, N. (1993). *TIMIT Acoustic-phonetic Continuous Speech Corpus LDC*93*S*1. USA, Philadelphia: Linguistic Data Consortium.