Speech Gender Recognition Using a Multilayer Feature Extraction Method

Husam Ali Abdulmohsin, Belal Al-Khateeb, and Samer Sami Hasan

Abstract Human speech contains paralinguistic properties used in automatic speech recognition (ASR) systems. These properties are used in many ASR applications such as gender recognition, which is the main goal of this paper. Gender recognition has been the target of many researchers since recognizing the human gender (female or male) is essential in many applications especially in security applications. Through this work, an ASR has been proposed and implemented. The main goal of any ASR system is to determine the best features that can address the required recognition. The features deployed in this work are smoothness, pitch, the first two formants and spectral centroid variability (SCV). The new approach proposed in this work was using the analysis of variance (ANOVA) as a feature selector to choose the best combination of features that can lead to the best classification accuracy, and then apply the decision tree feature selection algorithm to choose the best group of features. Then use backpropagation neural network (NN), Gaussian mixture models (GMM) and SVM as separate classifiers. The common voice dataset was used as benchmark dataset through all experiments of this work. The best result gained with respect to the three genders was 74.87% using the pitch and the first two formant features and classified by NN. The best result gained with respect to the two genders (female and male) was 97.71% using the pitch, and the first two formant features are classified by NN.

Keywords Automatic speech recognition · Speech gender recognition · Backpropagation NN · GMM · Common voice dataset

H. A. Abdulmohsin (✉) · S. S. Hasan
Computer Science Department, Faculty of Science, University of Baghdad, Baghdad, Iraq
e-mail: husam.a@sc.uobaghdad.edu.iq

S. S. Hasan
e-mail: ssami@uob.edu.iq

B. Al-Khateeb
Computer Science Department, College of Computer Science and Information Technology, University of Anbar, Anbar, Iraq
e-mail: belal-alhkhateeb@uoanbar.edu.iq

A. K. Bashir et al. (eds.), *Proceedings of International Conference on Computing and Communication Networks*, Lecture Notes in Networks and Systems 394,
https://doi.org/10.1007/978-981-19-0604-6_10

1 Introduction

When understanding the speech production process in female and male, it can be noticed that formants of females are higher in frequency than those of their male counterparts and the spectrum of female voiced sounds are lower than in male sounds, since the spectrum usually decreases in amplitude with increasing the frequency. All these acoustic effects are caused by the production of speech. Therefore, it is possible to find gender-specific features represented in acoustic speech signals [1].

Due to the transgender phenomena that appeared in the last years, and the surgeries applied to the vocal cords to adapt with the new gender, it is essentially required to determine the gender of the human being, regardless of its new gender for security identification reasons.

Acoustic voiced sounds are generated through the vibration of the vocal cords that in turn generates the periodic behavior of voice. This oscillation in frequency is called pitch. The pitch feature and other frequency-related features used in this work have a clear relationship with gender, and especially the low frequencies that contain the most important speech properties are useful in ASR. It is important to mention that male speech has a low-frequency behavior compared to female speech. The pitch in particular depends on the glottis physical characteristics, which are mass and elasticity [2, 3]. Speech is always represented as a discrete signal $x(n)$; therefore, the pitch represents the fundamental frequency (f_0) where the signal is repeated. The inverse is the fundamental period (T_0). Statistically, there are certain frequency intervals for men regarding each language, for example, the pitch of the Spanish men lay in the frequency interval, 50–300 Hz, and the Spanish women and children can reach to 500 Hz frequency [4].

Many researches have been published in the field of gender recognition, but none has studied the third gender's psychological effect on voice. Through this study, we tried to analyze the effect of the third gender on the human voice, and this study is considered a new field of study, since there are no researches published under this problem statement. Through our work, we were able to recognize the third gender voice, but as can be seen from the results and discussion section, that we did not achieve high classification accuracies. The limitation of our work lays in the features extracted. More features need to be extracted and more related to the third gender. The other limitation is the need for a dataset that shows third gender voice samples that have been under vocal tract transformation surgery, to change the voice from one gender to the other, in order to study the original voice features in sound that can differ the third gender from the other two genders.

2 Related Work

Many researches have confirmed the existence of unique acoustic and physiologic features in each of the female and male voices that can be used to recognize male

from female voice, but till now, they have not reached the required classification accuracy [5, 6]. In the past few years, gender recognition has gained the interest of many researchers.

In 2015, Faek [7] used the first four formants and twelve MFCC's as features and used the SVM as a classifier. A special feature selection is used based on the frequency range that the feature represents. A total of 114 speech samples uttered in Kurdish language were used in this work. The model of two classes (adult males and adult females) of gender recognition reached 96% recognition accuracy.

In 2019, Shaqra et al. [8] designed four emotion recognition models to present the relationship between gender/ age and emotion. The results showed that when using the same classifier for all four models on each gender and on different age limits, the results were higher compared to the performance of the system on all samples without categorization. This proves that gender and age affect the emotion recognition accuracy for its direct effect on voice.

In 2019, Alkhawaldeh et al. [9] provided an analysis about the gender-related features in speech and experimented three feature selection algorithms to find the best features. Then studied different machine learning (ML) models with different theoretical background, to find the best gender-related ML models. The best result gained was 99.7% classification accuracy.

In 2019, Abdulsatar [10] proposed a two-part system. The first part was called pre-processing and feature extraction to select the best feature, which are the first four formant frequencies and twelve MFCCs used to extract relevant features to recognize the gender and K-NN for classification. The highest accuracy classification obtained was 66%.

In 2021, Kwasny [11] applied d-vector and x-vector as deep neural network (DNN)-based embedder architectures to gender classification experiments. Then applied a transfer learning-based training scheme with pre-training the embedder network. The best overall performance achieved gender recognition was 99.60%.

Many challenges were faced; one of the challenges was the psychological situation of the human being and its effect on voice. When the human being is psychologically unstable, his vowels will be differently announced when he is psychologically stable [8, 12]. The other challenge was dealing with similarity in children voices of age 3 to 7 years; therefore, children were avoided in the experiments applied to the method proposed in this work.

The limitation of the state of art works is avoiding working on the third gender (not female neither male). Through this work, many experiments were conducted on the third gender, which is considered the first paper that studies the difference between the three genders (female, male and others).

3 Materials and Methodology

- Dataset
  The dataset utilized in this work is the common voice dataset version en_2637h_2021-07-21. The common voice dataset contains 60 different languages recorded in different percentages as 23% US English, 8% England English, 7% India and South Asia (India, Pakistan, Sri Lanka), 3% Australian English, 3% Canadian English, 2% Scottish English, 1% Irish English, 1% Southern African (South Africa, Zimbabwe, Namibia), 1% New Zealand English and many other languages.

  The common voice dataset was recorded by 75,879 different individual voices with different age limits. 6% of the individuals were younger than 19 years old that means 4552 individuals of age less than 18 have participated in recording this online public dataset and 94% of the dataset is recorded by individuals older than 18 years. All voices related to under legal age (<18) were neglected from our experiments, and only adult voices were included.

  The common voice dataset contains adult voice for three genders, 45% male, 15% female and 40% third gender [13]. The number of samples tested through the experiments of this work was 71,327 samples. A total of 32,098 (45%) of those samples were male voices, 10,699 (15%) of the samples were female voices and 28,530 (40%) of the samples were for other genders.

4 Framework

The block diagram of the proposed method is shown in Fig. 1. In the following three sections, the main steps of the proposed method will be illustrated and discussed in details.

- Pre-processing
  All samples were segmented according to the ratio (0.05%) of the original signal, and the overlap ratio deployed in this work was (0.025%), which provides (50%) overlap, and the total number of segments generated will be $(n - 1)$, where n is the number of original segments.
- Feature Extraction
  A feature is a measurable property established from the material being observed [14]. The most important aspect in feature extraction is extracting the most relevant features to the problem statement. In the case of this work, the features extracted were smoothness, pitch, the first two formants and spectral centroid variability (SCV) features, which are strongly related to human being gender, and were selected according to the experiments conducted.

  Smoothness is defined as the transaction of speech through air, as much as the speech was smooth as much as its transaction was slower, and when speech is rougher, the transaction of speech is faster [6]. The smoothness was calculated

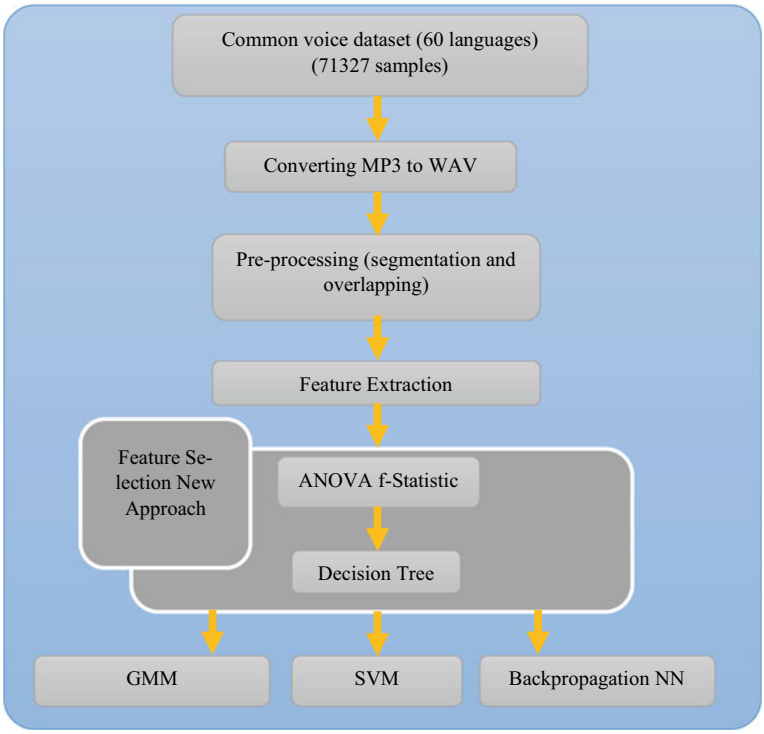

Fig. 1 Block diagram of the method proposed in this work

through two domains, first the time domain, the second, the spectral domain. Smoothness is calculated through Eq. (1 and 2) [15].

$$\mathrm{GV}_t = \frac{1}{P}\sqrt{\sum_{j=1}^{P}(\mathrm{var}_t(j))^2} \tag{1}$$

$$\mathrm{GV}_s = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(\mathrm{var}_s(i))^2} \tag{2}$$

where var_t and var_s represent the variances in time and spectral domain of the spectral feature, P is the dimension of the feature, N is the length in the time domain of the feature

- Feature Selection
  In this work, the ANOVA feature selection algorithm was used to filter features ahead of constructing the decision tree, to remove all irrelevant features, then pass the selected features to the decision tree.

  Decision tree is used for classification purposes or used as a feature selection algorithm of type embedded, and also used in data mining and machine learning [16, 17].

Most of the decision tree implementations in the previous state-of-the-art works such as ID3 [18], C4.5 [19] and CART [20] did not measure the importance of each feature regarding other classes and the final classification results. Therefore, in this work, we will determine the importance of each feature regarding each class. In this work, each feature will be weighted and the weight will be used in feature selection and finally in the decision tree construction. Feature weight will be calculated respectively, the feature with the highest weight will be selected as the root feature of the next layer and so on, the decision tree will be constructed.

A filter ranking method was used through this work for three reasons. First, to filter the less relevant variables. Second, to benefit from the criteria of variable selection by order of the variable ranking techniques. Third and finally, their simplicity and good success are reported from online applications. A ranking criterion is used to score each variable, then a threshold is fixed through experiment, and used to remove variables below that threshold [14].

Feature selection methods that are applied before classification are considered filter feature selection methods, that's why ranking methods are considered filter methods. The main principle of feature selection methods is to select unique features that contain useful information of different classes in the dataset through using a basic property of that feature. This property is called feature relevance that measures the power of that feature classifying different classes [14, 21, 22]. The chi square and analysis of variance (ANOVA) statistical feature selection methods were used in this work to measure the independence of two selected features.

The chi square feature selection method was calculated through Eq. (3) [23].

$$x_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{3}$$

where c is the degree of freedom, $O(s)$ are the observed values, which are s number of values, and $E(s)$ are the expected values.

ANOVA that is developed by the statistician Ronald Fisher [24] is a set of statistical models and their related estimation procedures, like the variation among and between groups of features that are used to analyze differences between means. ANOVA is based on total variance law, where the variance observed in a specific variable is partitioned into attribute components to other sources of variation. ANOVA in the simplest form provides a statistical experiment of whether two or more feature groups means are equal, not as the t-test that involves two means only.

The proportion of variance in ANOVA represented by a feature or groups of features can be found through Eq. (4).

$$\text{Variance} = \frac{\text{SST}}{\text{TotalSS}} \tag{4}$$

where the SST is the treatment sum of squares and the Total SS is the total sum of squares. As much as the higher ratio, the more groups of features can represent

the data. In other words, the groups of features with higher proportion must be selected [25].

- Feature Classification
  The backpropagation NN and the GMM classifiers were selected in this work, to classify the three genders.

5 Results and Discussion

The aim of the experiments deployed in this work was to test which of the features (smoothness, pitch, the first two formants and spectral centroid variability (SCV)) or group of features can act better in gender recognition with respect to each of the two classification algorithms which are backpropagation NN and the GMM classifiers. The experiments also aim to evaluate the performance of the new proposed feature selection algorithm with respect to feature and classifier.

- Gender recognition results with respect to each feature and classifier

The four types of features will be tested individually with each of the two classifiers, after applying the feature selection algorithm that is proposed in this work.

Table 1 shows the female gender classification performance according to feature and classifier. The results show the outperformance of backpropagation NN on GMM through all types of features and also show that the first two formant features gained the highest classification accuracy in female gender speech recognition, more than the other three types of features.

Table 2 shows the male gender classification performance according to feature

Table 1 Female recognition results

Feature type	Backpropagation NN (%)	GMM (%)
Smoothness	53.22	48.39
Pitch	46.34	44.4
First two formants	57.45	52.12
Spectral centroid variability (SCV)	50.89	45.23

Table 2 Male recognition results

Feature type	Backpropagation NN (%)	GMM (%)
Smoothness	33.09	38.5
Pitch	52.41	48.36
First two formants	38.79	40.21
Spectral centroid variability (SCV)	47.34	41.56

Table 3 Third gender recognition results

Feature type	Backpropagation NN (%)	GMM (%)
Smoothness	21.02	12.08
Pitch	32.57	30.88
First two formants	25.55	22.04
Spectral centroid variability (SCV)	16.04	19.89

and classifier. The results show the outperformance of backpropagation NN on GMM through all types of features and also show that the pitch feature gained the highest classification accuracy in male gender speech recognition, more than the other three types of features.

Table 3 shows the third gender classification performance according to feature and classifier. The results show the outperformance of backpropagation NN on GMM through all types of features and also show that the pitch features gained the highest classification accuracy in third gender speech recognition, more than the other three types of features. But as a overall conclusion, the third gender was misclassified, mostly to the male gender, that is why the pitch feature outperformed other features like in male gender speech recognition, that is explained by the similar properties in the speech signal of the male and third gender. Because of the misclassification of the third gender, Table 3 shows the low classification accuracy gained with respect to the other two genders classification accuracy mentioned in Tables 1 and 2.

- Gender recognition results with respect to best feature group and classifier

After going through all six combination possibilities of the four feature types with best features selected, it was found that the highest accuracy results gained were through deploying the pitch and the first two formants in speech, with respect to the backpropagation NN classifier as shown in Table 4. The highest accuracy gained was 74.87% with respect to all three genders. If the third gender was excluded from the experiments, the highest classification accuracy achieved to classify female and male genders (without the third gender) is 97.71% with respect to the backpropagation NN, and 91.03% with respect to the GMM classifier using the pitch and first two formants' features.

6 Conclusion

To design a system that can recognize age through the same setting was challenging, because age is related to language, and each language has a different range of frequencies for the male and female and children.

Table 4 All genders' recognition results

Feature type	Backpropagation NN				GMM			
	Female (%)	Male (%)	Third gender (%)	Total (%)	Female (%)	Male (%)	Third gender (%)	Total (%)
Smoothness, pitch	83.23	77.98	25	62.07	79.05	79.6	23.4	60.68
Smoothness, first two formants	88.34	72.45	24.34	61.71	85.03	75.54	19.21	59.89
Smoothness, spectral centroid variability (SCV)	83.06	79.02	20.26	60.78	81.6	82.45	17.08	60.37
Pitch, first two formants	98.32	97.11	29.2	74.87	92.06	90	24	68.68
Pitch, spectral centroid variability (SCV)	84.3	80.34	22.24	62.29	83.2	81.09	18.2	60.83
First two formants, spectral centroid variability (SCV)	75.55	74.44	21.22	57.07	70	79.33	20.5	56.61

The similar frequency behavior between the male voices and third gender caused a lot of ambiguity to the system designed in this work, regardless of the strong gender related features extracted and the new designed feature selection method.

It is clearly noticed that the third gender recognition classification accuracies achieved are very low with respect to the other two genders, which proves two things, first there are some special properties in the transgender's speech that differs them from other genders, but those properties are weak or were not extracted perfectly. Second, a lot of the third gender speeches were misclassified as male voices and vice versa that led to the low accuracy classification in both genders, the male and the third gender, which either is explained that the original gender of the transgenders was a male gender, and the original speech properties retain in the speech of the transgenders regardless of the new gender selected willingly.

Acknowledgements Many thanks to Mozilla (voice.mozilla.org) for creating such a global speech dataset, with many languages and many specifications with huge details.

References

1. J. Harrington, S. Cassidy, The acoustic theory of speech production, in *Techniques in Speech Acoustics*. (Springer, 1999), pp. 29–56
2. L. Rabiner, R. W. Schafer, *Digital Processing of Speech Signal* (1978)
3. J.R. Deller Jr, J.G. Proakis, J.H. Hansen, *Discrete Time Processing of Speech Signals* (Prentice Hall PTR) (1993)
4. Rabiner, L.R., B.J.E.C.P.-H. Gold, *Theory and application of digital signal processing* (1975)
5. Titze, I.R., *Physiologic and Acoustic Differences Between Male and Female Voices*. J. Acoust. Soc. Am. **85**(4), 1699–1707 (1989)
6. G. Fant, *Acoustic Theory of Speech Production* (Walter de Gruyter, 1970)
7. K.F. Fatima, *Objective gender and age recognition from speech sentences*. ARO Sci. J. Koya Univ. **3**(2), 24–29 (2015)
8. F.A. Shaqra, R. Duwairi, M.J.P.C.S. Al-Ayyoub, *Recognizing emotion from speech based on age and gender using hierarchical models*. Procedia Comput. Sci. **151**, 37–44 (2019)
9. R.S.J.S.P. Alkhawaldeh, DGR: gender recognition of human speech using one-dimensional conventional neural network. Sci Program **2019** (2019)
10. A.A. Abdulsatar, et al., *Age and gender recognition from speech signals*. J. Phys. Conf. Ser. (2019)
11. D. Kwasny, D.J.S. Hemmerling, Gender and age estimation methods based on speech using deep neural networks. Sensors **21**(14), 4785 (2021)
12. H.A. Husam Ali Abdulmohsin, A.M.J.A. Hossen, J. Mech. Continua Math. Sci. Speech Emot. Recogn. Survey **15**(9), 24 (2020)
13. R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F.M. Tyers, G. Weber, Common voice: a massively-multilingual speech corpus, in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)* (2020), pp. 4211—4215
14. G. Chandrashekar, F.J.C. Sahin, A survey on feature selection methods. Comput. Electr. Eng. **40**(1), 16–28 (2014)
15. P.T. Nghia, et al., A measure of smoothness in synthesized speech. Electron Commun **6**(1–2) (2016)
16. H. Zhou, et al., A feature selection algorithm of decision tree based on feature weight. Exp. Syst. Appl. **164**, 113842 (2021)
17. H. Sun, X.J.C. Hu, I.L. Systems, Attribute selection for decision tree learning with class constraint. Chemometr. Intell. Lab. Syst. **163**, 16–23 (2017)
18. , J.R. Quinlan, Induction of decision trees. Mach. Learn. **1**(1), 81–106 (1986)
19. J.R. Quinlan, *C4. 5: Programs for Machine Learning* (Elsevier, 2014)
20. C.-H. Yeh, *Classification and Regression Trees (CART)* (Elsevier, 1991)
21. H.A. Abdulmohsin et al., A new hybrid feature selection method using T-test and fitness function. CMC-Comput. Mater. Continua **68**(3), 3997–4016 (2021)
22. R. Kohavi, G.H John, Wrappers for feature subset selection. Artif. Intell. **97**(1–2), 273–324 (1997)
23. S. Gajawada, Chi-square test for feature selection in machine learning (2019). Retrieved from Towards Data Science: https://towardsdatascience.com/chi….
24. P. Moran, C.J.T.o.t.R.S.o.E. Smith, The correlation between relatives on the supposition of mendelian inheritance. Earth Environ. Sci. Trans. R. Soc. Edinburgh **52**, 899–438 (1918)
25. A. Anderson, *Business Statistics for Dummies* (Wiley, 2013)