ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v18i5.13717

Gender voice classification with huge accuracy rate

Mustafa Sahib Shareef[1], Thulfiqar Abd[2], Yaqeen S. Mezaal[3]
[1,2]Al Muthanna University, Al-Muthanna, Iraq
[3]Medical Instrumentation Engineering Department, Al-Esraa University College, Iraq

Article Info

Article history:

Received Jul 24, 2019
Revised Apr 6, 2020
Accepted Apr 30, 2020

Keywords:

Audacity
Classification accuracy
Machine learning algorithm (J 48)
MFCC
VQ

ABSTRACT

Gender voice recognition stands for an imperative research field in acoustics and speech processing as human voice shows very remarkable aspects. This study investigates speech signals to devise a gender classifier by speech analysis to forecast the gender of the speaker by investigating diverse parameters of the voice sample. A database has 2270 voice samples of celebrities, both male and female. Through Mel frequency cepstrum coefficient (MFCC), vector quantization (VQ), and machine learning algorithm (J 48), an accuracy of about 100% is achieved by the proposed classification technique based on data mining and Java script.

Corresponding Author:

Yaqeen S. Mezaal,
Medical Instrumentation Engineering Department,
Al-Esraa University College,
Baghdad, Iraq.
Email: yakeen_sbah@yahoo.com

1. INTRODUCTION

In speech communication, an auditor doesn’t merely translate a language message from the speech signal, but simultaneously he/she likewise deduces paralinguistic details about gender, age and other features of the talker. This information sort stands for speech information. This gender detection has application in numerous fields like gender categorization of phone calls for gender sensitive investigations and classifying the gender, and eliminating the gender particular constituents gives greater compression rate with improved bandwidth [1].

Based on acoustic features, voice information can be explained by some acoustic factors as in the fundamental frequency (f0) for perceptual relevance and spectral formant frequencies [2]. Male has the frequencies less than the female. Nevertheless, formant frequency is somewhat associated with the vowels and hence it is text reliant. Gender classification can be prepared by pitch/fundamental frequency extracted from diverse approaches. Pitch stands for highly imperative feature that can be gotten from dissimilar methods in time and frequency domains or by the combination of both of them. Time domain approaches have all these methods that can be adopted directly on the speech samples. The speech waveform is straightforwardly investigated based on methods of short time autocorrelation, modified autocorrelation through clipping technique, normalized cross correlation function, average magnitude difference function, square difference function etc. Likewise, in frequency domain approaches for the signal, the frequency content can be primarily processed and then the information can be extracted from the spectrum. These approaches comprise harmonic product spectrum investigation and harmonic to sub harmonic ratio. There are as well several methods that don't come under either time domain or frequency domain as in wavelets [3, 4].

2. LITERATURE REVIEW

Gender based recognition, speech classification and processing were adopted for long time ago. To apply gender recognition, many conceptions have acquired over time. Current papers about gender detection have shown that voice can be altered into diverse parameters. Foremost parameters are pitch and frequency. Classification can be organized for distinguishing female, male and children. Firstly, the recognition system was prepared with the training data. Then, testing data have presented and assessed based on system performance for these data. The acquired consequences have been dissimilar for diverse procedures. They have generated diverse perceptible consequences at diverse periods. Gender based classification based on fundamental frequency [5] and pitch with numerous training and testing data has shown that Logistic regression has been the finest algorithm with accuracy of 92% than random forest. AdaBoost for voice data with identical language performs better in the case of Random Forest algorithm with accuracy of 93%. For speech recognition, random forest outfits agreeably in accordance with fundamental frequency and pitch for categorizing female, male and a child [6]. Added tuning stands for the binning method for developing the efficiency with desired fallouts.

Voice based word extraction lab view [7] operates well for voice classification. It has better results for vowels extracting in male testers. When the testers have been trained and investigated, it capably gives a meaningful outcome. It has as well perceived that by aggregating an unvoiced fragment in speech related to a sound of 's' value of pitch upsurges obstructing a gender detecting for male samples. Correspondingly, by increased voice fragment of speech as in 'a', it drops the pitch value and it is unsuccessful for classifying it as the speaker talks dual dissimilar tones. Speech recognizing systems in adult has impulsive and vocal length changes. They are able to sound as in male and female. Consequently, it has been hard to categorize the both genders.

Several feminine voices have been difficult for investigating based on the pitch [8] as in investigation of single facet of womanly voices that doesn't satisfy the requirements. Based on [8], the feminine voice should be recognized with dissimilar parameters as compared with male such as shrill, high-pitch, emotive and swoopy parameters. These are dissimilar parameters for feminine individuals, and they differ from female to another. Therefore, data set must be sorted out based on this adopted classification of masculine and feminine. In [9], fundamental frequency has a verbal grouping with nonlinguistic and paralinguistic data about speaker. These three issues relate to masculine and feminine, and it as well relies on huge pitch and tone of the talker. It was achieved for setting a frequency irrespective of any acquaintance about range and syllable-external data. Accordingly, a speaker voice differs between high and low pitches among utterers.

Glottal-pulse rate (GPR) and vocal-tract length (VTL) have associated with age, sex and size of a speaker. It has been not certain about the ways for the dual factors to be integrated for affecting the perception of speaker sex, age and size. In [10], experiments have conducted to evaluate the interface influence for GPR and VTL upon findings of speaker sex, size and age. Vowels have been scaled for characterizing individuals with varied GPRs and VTLs, containing numerous amounts over than the standard range of the populace. Spectators have been requested to estimate a size and sex/age of a talker. The discriminations of speaker size explain that VTL holds a resilient impact on apparent speaker size. The fallouts for sex and age category (woman, man, girl or boy) explain that for vowels with GPR and VTL magnitudes in the standard range. Findings of speaker age and sex have affected equivalently by GPR and VTL. For vowels with irregular groupings of small GPRs and diminutive VTLs, the VTL data seem to have impacts on determining the sex/age judgement.

In [11], the automatic voice disorder classification system was projected based on first dual formants of vowels. Five categories of voice syndrome of paralysis, polyp, sulcus and cyst, have been employed in the tests. Voiced Arabic digits from the voice disorderly individuals have verified for an input. The initial formant and 2nd formant have extracted from the [Fatha] and [Kasra] vowels that have been existing in Arabic digits. The four features have been then employed for categorizing the voice disorder by means of dual categories of classification systems: vector quantization (VQ) and neural networks (NNs). In the tests, NN implements superior performance as compared with VQ. For female and male talkers, the classification percentages have been 67.86% and 52.5%, by means of NN. A finest classification rate has been 78.72% for feminine sulcus disorder.

Gender identification by support vector machine (SVM) [12] shows that gender speech is evaluated by numerous speech appliances as in compressed speech, telephone speaking and variance in languages, etc. It transfers that masculine voice from pitch, period and Mel-frequency has been within 100-146Hz range and in the female within 188-221Hz. At this point, the voice has been separated based on the extracted feature as well as frequency.

Gaussian mixture model (GMM) based gender classification [13], suggests that speech can be examined based on age, words, etc. Mixture model has been adopted with recognizing accuracy up to 98%. It stands for the effectual technique for speech detecting and gender investigation. Classification has been dependent on joint factors of pitch and relative spectral perceptual linear predictive (Rasta-Plp) factor for modeling masculine and feminine speech.

This research work examines speech signals to develop a gender classifier by speech analysis for forecasting the gender of the speaker by investigating diverse parameters of the voice sample. By means of MFCC, VQ, and machine learning algorithm (J 48), an accuracy of approximately 100% has realized by the projected classification technique based on data mining and Java script applied on a database with 2270 voice samples of celebrities.

3. MEL FREQUENCY CEPSTRUM COEFFICIENT

In this study, MFCC procedure has been employed for extracting features from a speech signal and compare the unidentified speaker with the current speaker in a database. Figure 1 clarifies a full pipeline of MFCC technique [14]. MFCC technique has frequently employed for generating a fingerprint of the sound files. MFCC has been dependent on recognized dissimilarity of humanoid ear's critical bandwidth frequencies with spaced filters in linear way at lower frequencies. Also, it is employed logarithmically at higher frequencies for capturing the imperative features of speech. Reported researches revealed that humanoid perception of the frequency subjects for speech signals doesn't have the linear scale. Therefore, all tones have an authentic frequency, *f*, determined in Hz, while a specified pitch has determined based on a scale termed as Mel scale. The Mel-frequency scale is linear frequency spacing at lower than 1000 Hz and a logarithmic spacing beyond 1000 Hz. The pitch of a 1 kHz tone, 40 dB higher than the perceptual hearing threshold, is feasibly stated as 1000 Mels as a reference point.

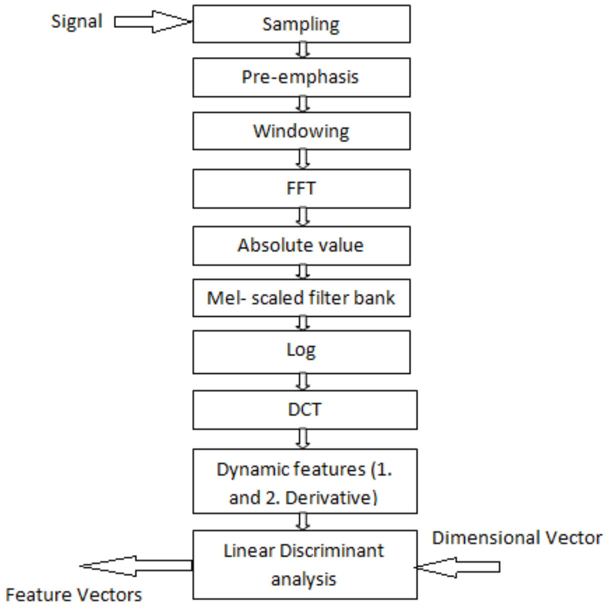

Figure 1. Pipeline of MFCC

The following equation is for calculating the Mels for the specific frequency: $Mel\,(f) = 2595 * log10(1 + f\,/\,700)$. The brief about MFCC procedures has presented in Figure 2. A speech waveform has collected for removing acoustical interference or silence that is feasibly existing in the beginning or ending of a sound file. A windowing block reduces the signal discontinuities through tapering the starting and termination of each frame to zero. The FFT block transforms all frames from time to frequency domain. A signal has schemed in the Mel- frequency wrapping block in contradiction of the Mel spectrum to mirror humanoid hearing. In an ending step, the cepstrum and the Mel-spectrum scale can adapt back to typical frequency scale. This spectrum has a noble depiction of the spectral features of a signal that has been important for signifying and identifying features of a speaker. Once the fingerprint has produced, the acoustic vector can be kept as a reference in a database. Furthermore, its consequential vector can be compared with those in a database. Another time by means of euclidian distance technique and an appropriate matching can be concluded. This process has termed as feature matching.

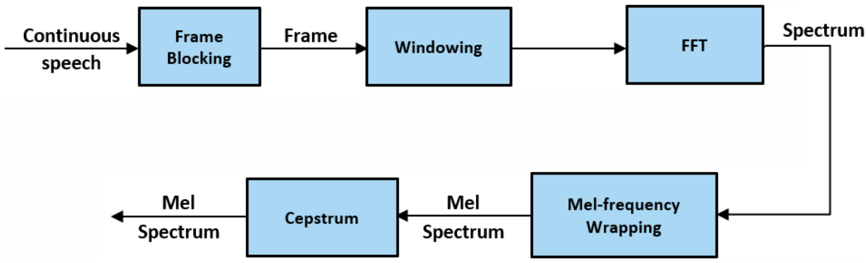

Figure 2. MFCC diagram

4. VECTOR QUANTIZATION

The speaker recognition system would be capable for estimating probability distributions of the processed feature vectors. Storage for each distinct vector that produces from the mode training is intolerable, even as these distributions are definite over a big dimensional space. It has been frequently simpler to begin through quantizing every feature vector to one of a reasonably insignificant amount of template vectors, with vector quantization. VQ processes a huge set of feature vectors and generates a reduced set of measure vectors that characterizes the distribution centroids. The VQ method has a feasibility for extracting a small amount of descriptive feature vectors as the effectual tool for distinguishing the speaker specific features. Accordingly, storage of each generated vector from the training has been intolerable. Based on adopted training data, features have collected for generating a codebook for every speaker. The data, in recognition step from the investigated speaker, will be compared to a codebook of each speaker, and a difference will be measured. These differences can be employed in the recognition decision [14, 15].

5. PROPOSED TECHNIQUE AND RESULTS

By means of data mining and Java Script, for voice gender classification, data set is collected from (CMU_ARCTIC databases); http://festvox.org/cmu_arctic/cmu_arctic/cmu_us_clb_arctic/wav/. Make pre-emphasis for each voice sample in the database. By the way, data set is 2270 (1138 male, 1132 female). Data mining stands for a discovering process for patterns in huge data sets including methods at the intersection of statistics, machine learning and database systems [16-19].

Then, features of data are split to train and test using cross validation algorithm. Then, extract MFCC features for each voice and apply vector quantization on each matrix of MFCC. Then, extract mean (sum of features/their number), standard deviation (STD), zero crossing (ZC), amplitude (AMP*)* features for each voice. Here, the amplitude stands for the energy of voice. Feed train features to machine learning algorithm (J 48) and build classifier. Lastly, feed test features to our classifier to classify voice as male or female. The proposed procedure is depicted in Figure 3.

J48 algorithm stands for an algorithm employed for producing a decision tree established by Ross Quinlan. It has been an extension of Quinlan's earlier ID3 algorithm. The decision trees produced by J48 algorithm can be adopted for classification. Accordingly, J48 algorithm is frequently referred as a statistical classifier. It has been powerful tool in data mining [20]. The detailed accuracy for each class is depicted in Table 1. Table 2 shows the comparisons of our proposed voice recognition method with reported ones in [21-26]. The projected voice recognition method in this study has better than [21-26] with accuracy of about 100 %.

Table 1. Resultant accuracy for each class

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	Class
Weighted	0.998	0.002	0.998	0.998	0.998	0.996	0.998	(Male)1
Avg	0.998	0.002	0.998	0.998	0.998	0.996	0.998	(Female)2
	0.998	0.002	0.998	0.998	0.998	0.996	0.998	

Table 2. The comparisons of proposed voice recognition method with reported approaches in [21-26]

Ref.	Proposed Method	Accuracy
[21]	Pitch frequency and GMM classifier	98.65%
[22]	4 classifiers involving GMM, vector quantization (VQ), multilayer perceptron (MLP), and learning vector quantization (LVQ)	96.4%
[23]	joint estimated voice acoustic level of 5 diverse approaches into single score level	81.7%
[24]	Fusion score of 7 subsystems. The feature vectors include the MFCC, PLP, and prosodic on 3 classifiers of GMM, SVM, and GMM-SV-based SVM combined at the score level	90.4
[25]	Decision tree (DT) and SVM with the MFCC feature, on an isolated Corpus for classifying gender voice	93.16% and 91.45% for MFCC-SVM and MFCC-DT
[26]	Convolutional neural networks and Deep neural networks (DNNs) for MFCC enhancement	58.98% and 56.13%, evaluated on DNN and I-Vector classifiers
This study	MFCC, VQ, and machine learning algorithm (J 48)	99.8 %

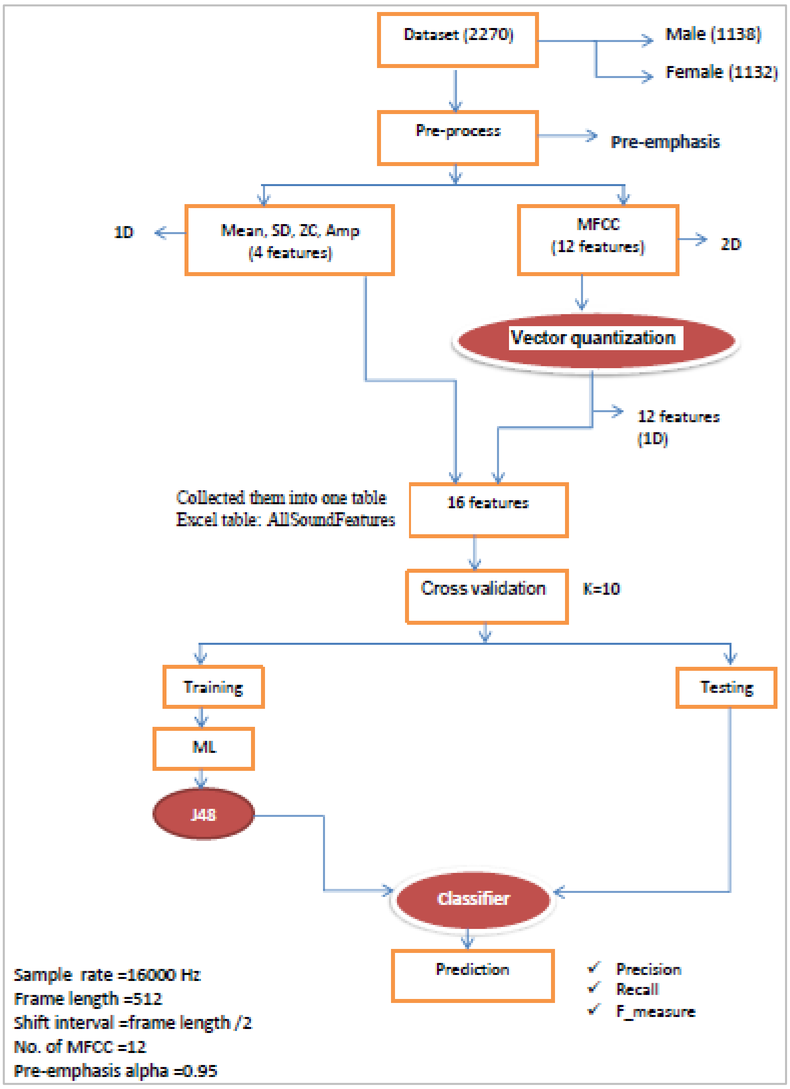

Figure 3. Proposed procedure for voice gender classification in this study

6. CONCLUSION

In this paper, voice gender classification has been implemented based on MFCC, VQ, and machine learning algorithm (J 48). This classification system is tested over voice samples of 2270 celebrities including

1138 males and 1132 females. The classification accuracy of about 100% is achieved by the proposed classification technique that is higher than many reported tecniques in the literature.

REFERENCES

[1] Sukhostat L., and Imamverdiyev Y., "A comparative analysis of pitch detection methods under the influence of different noise conditions," *Journal of voice*, vol. 29, no. 4, pp. 410-417, 2015.

[2] Hautamäki R. G., Sahidullah M., Hautamäki V., Kinnunen T., "Acoustical and perceptual study of voice disguise by age modification in speaker verification," *Speech Communication*, vol. 95, pp. 1-15, 2017.

[3] Eichhorn, Julie Traub, et al. "Effects of aging on vocal fundamental frequency and vowel formants in men and women," *Journal of Voice,* vol. 32, no. 5, 2018.

[4] Phoophuangpairoj, Rong, and Sukanya Phongsuphap, "Two-Stage Gender Identification Using Pitch Frequencies, MFCCs and HMMs," *2015 IEEE International Conference on Systems, Man, and Cybernetics*, IEEE, 2015.

[5] Nasr M. A., Abd-Elnaby M., El-Fishawy A. S., El-Rabaie S., Abd El-Samie F. E., "Efficient implementation of adaptive wiener filter for pitch detection from noisy speech signals," *Menoufia Journal of Electronic Engineering Research*, vol. 27, no. 1, pp. 109-126, 2018.

[6] Ericsdotter C., Ericsson A. M., "Gender differences in vowel duration in read Swedish: Preliminary results Working Papers," *Lund University Department of Linguistics,* pp. 34-37, 2001.

[7] Whiteside S. P., "Temporal-Based Acoustic-Phonetic Patterns in Read Speech: Some Evidence for Speaker Sex Differences," *J International Phonetic Association,* vol. 26, pp. 23-40, 1996.

[8] Henton C. G., "Fact and fiction in the description of male and female pitch," *Language and Communication*, vol. 9, pp. 299-311, 1989.

[9] Bishop J., and Keating P., "Perception of pitch location within a speaker's range: Fundamental Frequency, voice quality and speaker sex," *The Journal of the Acoustical Society of America,* vol. 32, no. 2, pp. 1100-1112, 2012.

[10] Smith D. R., and Patterson R. D., "The interaction of glottal-pulse rate and vocal-tract length in judgments of speaker size, sex, and age," *The Journal of the Acoustical Society of America*, vol. 118, no. 5, pp. 3177-3186, 2005.

[11] Muhammad G., AlSulaiman M., Mahmood A., and Ali Z., "Automatic voice disorder classification using vowel formants," *2011 Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '11),* pp. 1-6, 2011

[12] Gaikwad S., Gawali B., and Mehrotra S. C., "Gender identification using SVM with combination of MFCC," *Advances in Computational Research,* vol. 4, pp. 69-73, 2012.

[13] Zeng Y. M., Wu Z. Y., Falk T., and Chan W. Y., "Robust GMM based gender classification using pitch and RASTA-PLP parameters of speech," *2006 Proceedings of the International Conference on Machine Learning and Cybernetics*, pp. 3376-3379, 2006.

[14] Patel Kashyap, R. K. Prasad, "Speech recognition and verification using MFCC & VQ," *Int. J. Emerg. Sci. Eng.,* vol. 1, no. 7, pp. 333-37, 2013.

[15] Alkhawaldeh Rami S., "DGR: Gender Recognition of Human Speech Using One-Dimensional Conventional Neural Network," *Scientific Programming,* vol. 2019, pp. 112, 2019.

[16] Roiger R. J., "Data mining: a tutorial-based primer," *CRC press*; 2017.

[17] Dutt A., Ismail M. A., Herawan T., "A systematic review on educational data mining," *IEEE Access*, vol. 17, no. 5, pp. 15991-6005, 2017.

[18] Sammut, Claude, and Geoffrey I. Webb, "Encyclopedia of machine learning and data mining," *Springer Publishing Company, Incorporated,* 2017.

[19] Ge Zhiqiang, Zhihuan Song, Steven X. Ding, and Biao Huang, "Data mining and analytics in the process industry: The role of machine learning," *IEEE Access,* vol. 5, pp. 20590-20616, 2017.

[20] Bhargava N., Sharma G., Bhargava R., Mathuria M., "Decision tree analysis on j48 algorithm for data mining," *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering,* 2013.

[21] Y. Hu, D. Wu, and A. Nucci, "Pitch-based gender identification with two-stage classification," Security and Communication Networks, vol. 5, no. 2, pp. 211-225, 2012.

[22] R. Djemili, H. Bourouba, and M. C. A. Korba, "A speech signal-based gender identification system using four classifiers," *Proceedings of the 2012 International Conference on Multimedia Computing and Systems,* pp. 184-187, 2012.

[23] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language*, vol. 27, no. 1, pp. 151-167, 2013.

[24] E. Yucesoy and V. V. Nabiyev, "A new approach with scorelevel fusion for the classification of a speaker age and gender," *Computers & Electrical Engineering*, vol. 53, pp. 29-39, 2016.

[25] M. W. Lee and K. C. Kwak, "Performance comparison of gender and age group recognition for human-robot interaction," *International Journal of Advanced Computer Science and Applications*, vol. 3, no. 12, pp. 207-211, 2012.

[26] Z. Qawaqneh, A. A. Mallouh, and B. D. Barkana, "Deep neural network framework and transformed MFCCs for speaker's age and gender classification," *Knowledge-Based Systems*, vol. 115, pp. 5-14, 2017.