Noise-Robust Gender Classification System Through Optimal Selection of Acoustic Features

Puneet Bawa, Vaibhav Kumar, Virender Kadyan, and Amitoj Singh

1 Introduction

A systematic study of the sounds acoustic in speech sounds used for gender detection and their relationship to interpretation is a challenging objective that in many areas, from linguistically to computer recognition, have important implications and applications. Human speakers typically use a normal system in which air is discharged from the lungs and formed into vocal cords and organs, including the tongue, lips, teeth, etc. [1]. Likewise, the acoustic voice analysis depends upon the sample characteristic parameters such as filtering, power, frequency, and duration [2]. These acoustic features have been traditionally defined mainly through the implementation methodologies of linear analytical and visualization approaches. However, in recent years, it is clear that these spectral representations were only very crude approximations to those actually produced by the auditory path in the peripheral and central regions [3–5]. Some of the most important characteristics of the auditory images are due to the asymmetric form of the cochlear filters and the retention of the fine-temporal filter output structure below 3–4 kHz [6]. Likewise, two main reasons for applying reliable biophysical models of the auditory system are

P. Bawa · V. Kumar
Centre of Excellence for Speech and Multimodal Laboratory, Chitkara University Institute of Engineering & Technology, Chitkara University, Chandigarh, Punjab, India
e-mail: puneet.bawa@chitkara.edu.in

V. Kadyan
Speech and Language Research Centre (SLRC), School of Computer Science, University of Petroleum & Energy Studies (UPES), Dehradun, Uttarakhand, India
e-mail: vkadyan@ddn.upes.ac.in

A. Singh (✉)
Maharaja Ranjit Singh Punjab Technical University, Bathinda, Punjab, India
e-mail: amitojsingh@mrsptu.ac.in

V. Kadyan et al. (eds.), *Deep Learning Approaches for Spoken and Natural Language Processing*, Signals and Communication Technology,
https://doi.org/10.1007/978-3-030-79778-2_9

detailed in the concepts relating to auditory systems. First, the relation of the acoustic characteristics of the speech to conventional systems of phonetic classification [7, 8] (as expressed by their output patterns) should be described. Second is the need of defining the practical standards that permits correct encoding of the speech signal in the presence of high background noise and over a wide spectrum of sound pressures [9]. In the same way, a system can also be taught when to use the robust machine learning algorithms to select and incorporate the functionality necessary for mapping voice data.

With technology growing rapidly, machine learning is an area of science that has undergone significant changes and is also a common trend [10]. Machine learning is an artificial intelligence subcategory that uses algorithms and data for computer learning to decide on particular issues in different areas, such as accounting, finance, and medicine, to recognize gender, voicing by means of machine learning and data mining techniques. In addition, the role of gender recognition applies to various digital multimedia applications including speech recognition, intelligent human-computer interactions, speaker diarization, biometrics, and video indexing [11–13]. Likewise, considering the rising demand of machine learning applications in speech recognition topic, the methodology to efficiently recognize gender has played an important role in the field of healthcare systems with existence of some pathologies including vocal-fold cyst. However, it is still regarded as a very difficult and daunting challenge to build a predictive model for gender identification through speech [14]. Also, in such a quickly developing environment of computerization, one of the most vital problems in the developing world is correct gender identification in Indian native languages, which are often termed the low-resource languages [15–17]. However, it is also a costly and time-consuming challenge to find enough marked data for classified training classifiers to make precise predictions, so human work is required, although it is much easier to find unlabeled data in general. Semi-supervised learning (SSL) algorithms are considered to be an appropriate way of exploiting secret data in an unlabeled collection to develop more precise classifications in order to tackle the issue of inadequateness existing in the low-resource data [18, 19]. In a similar manner, many classes of SSL algorithms have been proposed with each being evaluated on different methodologies and approaches with an objective of finding adequate relational difference in the distribution of labeled and unlabeled data.

There are several approaches to speech synthesis that can be used to enhance the incoming speech signal. Similarly, the work to be performed has to result in ground realities to match real-time system implementations and applications. Taking such real-time situations into account, in this chapter, noise data augmentation technique has been applied to introduce into the original dataset using three distinct types of noises, including Babble, Factory, and Volvo at random SNR values and labeled as male and female for classification. Further, this chapter uses a warbleR library package [20] with an objective of performing the acoustic analysis for visualizing the process of gender detection in dialectal Punjabi language. As of our knowledge, some efforts had been made for the development of adequate language resources, but no effort has been made in designing the classifiers corresponding to the Punjabi

children speech. Moreover, the study to access adequate dataset has been performed with the findings comparable with gender detection in order to optimize the selection of required parameters among the extracted 20 acoustic parameters. Finally, the adequate model for recognizing the gender based on the optimal selection of the extracted acoustic features has been made through the comparative analysis of three machine learning algorithms including random forest, SVM, and MLP.

2 Related Work

Analyzing audio and extracting features sometimes can become a significant task when you have to pick certain features and reject in order to perform some tasks. In [21], the authors used machine learning algorithm and computed features that can help to check the authenticity of the audio signal. The experiments were able to distinguish appropriate value for hyper-parameters to be used. Li and Liu [22] experimented with Mel filter energy features and Mel Frequency Cepstral Coefficient (MFCC) features as acoustic criterion for detecting Mandarin vowels with low error rate and high investigation rate. In [23], authors also explored selecting optimal features for accent recognition using MFCC, spectrogram, and spectral centroid features extracted from audio samples and fed the features into two-layer convolutional network. The results depicted that MFCC feature yields the highest accuracy. Likewise, authors in [24] also explored predicting the reason for a newborn baby based on acoustic features. Pitch features and formant frequencies chosen as acoustic features alongside K-means algorithm proved quite handy and provided conclusive results for detecting a "pain" in cry along with reason for the cry. Likewise, the research endeavors for building the state-of-the-art speech recognition model in tonal languages have been analyzed on the basis of the findings relating to native languages [11]. In [25], authors proposed an automated attendance system using audio for gender classification and image for matching the current visual with the stored one in database in order to evaluate whether a student is actually present in the class or not. In [26], investigators explored gender modeling with clean and noisy environments and presented MFCC features alongside Gaussian mixture model (GMM) for audio modeling. The proposed system was capable of gender classification based on either audio or visual feedback whichever is less noisy, although the method is vulnerable to a scenario when both audio quality and visual quality are bad; that is, data is noisy. For simulating male/female detection, in [27], authors investigated GMM modeling along with pitch parameters and RASTA-PLP variables. Both clean and noisy environments were considered while evaluating the generated GMM, which was obtained by varying covariance matrix. The proposed method seems as a step in right direction. Likewise, Copiaco et al. [28] also experimented with multi-channel audio classification with MFCC and Power Normalized Cepstral Coefficients using deep convolutional neural network. The proposed methodology produced 98% accuracy. In [29], authors stacked different machine learning models and tried to use acoustic features to model the data. A

slight improvement in accuracy has been observed with state-of-the-art methods, but it came with a space complexity for such a stacked model alongside time complexity for predicting the gender on one sample. In [30], authors experimented with Mel Frequency Spectral Coefficients (MFSC) rather than MFCC features and used simple neural network for classifying the data based on gender. The selection of optimal features and parameters proved decisive at the end as the results showed substantial improvement in accuracy with smoothing applied.

Using deep learning algorithms dynamically selects essential information in raw language signal for processing of classification layer. Thus, with the proposed algorithm [31], the researcher has avoided the absence of knowledge on feeling, which cannot be modeled mathematically as more of an acoustic feature of voice. In [32], research was conducted in Bahasa Indonesia related to gender identification, and a supervised machine learning algorithm was applied with MFCC features with several modeling algorithms such as SVM, K-nearest algorithm, and artificial neural network. The results paved the way for impact analysis of gender identification for audio recognition. In [33], authors experimented with long short-term memory (LSTM)-based recurrent neural networks for predicting age and gender using audio sample and also reduced the over-fitting problem by using data augmentation and improved the testing accuracy using regularization. The authors also explored bidirectional LSTM alongside MFCC features on low resource dataset and found that more data can yield more accurate results [34]. Also, assembling modeling techniques have been explored using machine learning models like naive Bayes, random forest, and linear regression for hate speech detection by processing Twitter dataset. The study shows that such kind of models can help achieve adequate results [35]. Analysis of audio features was also performed by researchers, and they found that algorithms such as gradient boosting and random forest can help in classifying gender based on acoustic features [36]. The researchers also set up a pipeline for gender-based emotion recognition where MFCC features along with convolutional neural networks were used with an average pooling layer instead of a fully connected layer at the end can achieve accurate results [37].

3 Semi-supervised Classification Algorithms

3.1 Random Forest

Random forest classifier is known for its best use in classification and regression tasks. It is an ensemble algorithm that utilized a stack of decision trees and predicted the class or probability value for every node in the tree. It is often known as random decision tree. On the other hand, the trees can be allotted a certain weight depending upon the importance of node in the decision tree. The node yielding low error rates has the chance of accurate predictions and hence should be allotted higher weight and vice versa. Setting up such pipelines can end up outputting decisive predictions.

3.2 Support Vector Machine

Support vector machine is a supervised modeling approach, which is known as one of the best in classifying or regression problem analysis. SVM models the training samples in such a way that it maximizes the difference between two given classes. A new sample is mapped to a space, and then the modeling algorithm tries to predict whether the sample belongs to the allotted space or not.

For a given training sample

$$(a_1, b_1), (a_2, b_2), \ldots\ldots\ldots\ldots, (a_m, b_m)$$

where a_i indicates m dimensional vector and b_i will be either of -1 or 1 representing the output class to which the sample belongs. The objective will be to find a hyperplane for which distance between the nearest point and hyperplane can be maximized and the classes can be distinguished using the hyperplane.

3.3 Multi-layer Perceptron

Multi-layer perceptron is a feedforward neural network that is used to depict multi-layer neural network or "vanilla" neural network. An elementary MLP will be having an input layer, a hidden layer, and an output layer [38]. It is a supervised learning approach that used back propagation to optimize the random weights which are attached to each hidden layer. In order to distinguish the data, which might not be able to separable using algorithms like SVM and random forest, an activation function is attached to the hidden layer, which is mainly sigmoid activation:

$$f(x) = \frac{1}{1 + e^x}$$

4 System Architecture

The database was created with a collection of 6603 voice recordings from both men and women in nearly equal ratio. This database classified the voice as female or male based on voice and speech acoustic properties. The recordings were done with or without the use of a recorder in both open and closed environment. Each voice sample has been stored with PCM header in .wav format including 3315 male recordings and 3288 female recordings. Further, considering the less amount of existing data, the analysis has further been performed using noise augmentation on both the male and female data such that there exists acoustic mismatch alongside

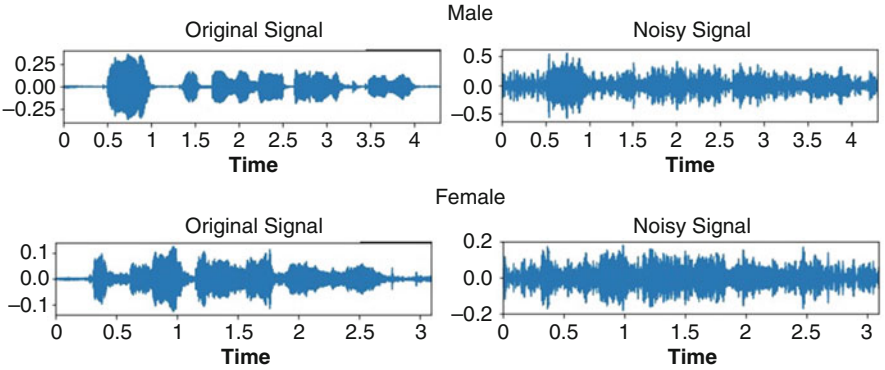

Fig. 1 Visual representation of male and female audio waveform under clean and noisy conditions

variation due to environmental conditions as shown in Fig. 1. Therefore, three different voices—Volvo, Factory, and Babble—from standard NOISEX-92 database [39] has been injected into original dataset at random value which contained much noisy SNR values ranging from −5 dB to 5 dB.

Next, the major focus is on the acoustic feature analysis for the evaluation of the classification performance. The 20 acoustic features—mean frequency (meanfreq), standard deviation corresponding to frequencies (sd), median frequency (median), first quantile (Q25), third quantile (Q75), interquartile range (IQR), skewness (skew), kurtosis (kurt), spectral entropy (sp. ent), spectral flatness (sfm), mode frequency (mode), frequency centroid (centroid), average measure of fundamental frequency (meanfun), minimum measure of fundamental frequency (minfun), maximum measure of fundamental frequency (maxfun), average measure of dominant frequency (meandom), minimum measure of dominant frequency (mindom), maximum measure of dominant frequency (maxdom), range dominant frequency's range across signal (dfrange), and modulation index (modindx) corresponding to male (M) and female (F)—have been extracted using inbuilt R library packages for clean data and noise-augmented data as detailed in Fig. 2a, b respectively.

Perhaps the best and most common machine learning algorithms for classification challenges have been found to be supervised classifiers including random forest, SVM, and MLP. Therefore, the differentiation values of three classification model algorithms utilizing these three classifiers on both clean and noisy datasets as shown in the block diagram in Fig. 3 are being experimented using all 20 features together and three most significant features distinctively.

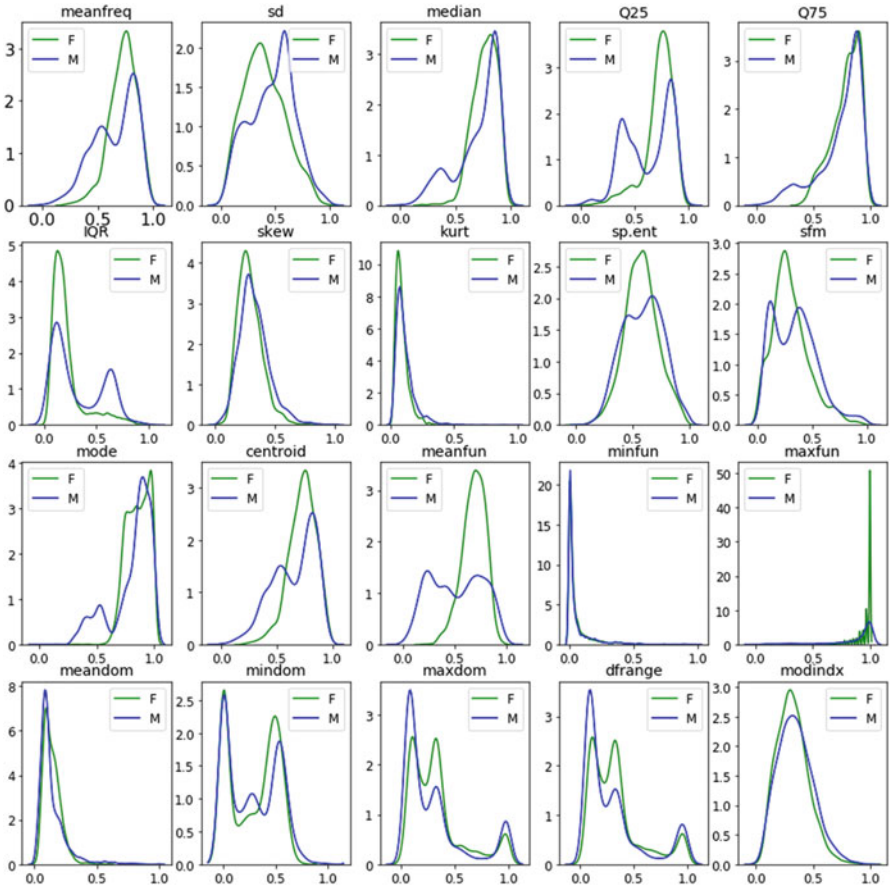

Fig. 2 (**a**) Visualization of acoustic features extracted on clean male and female audio dataset. (**b**) Visualization of acoustic features extracted on noise-augmented male and female audio dataset

5 Results and Discussions

In this section, we address a set of experiments to select the optimal feature parameter model performance for gender recognition from native voice clean dataset. Additionally, the augmented dataset including both noisy and clean dataset has been presented against the classification scheme with an objective of testing the performance of semi-supervised model under degraded conditions.

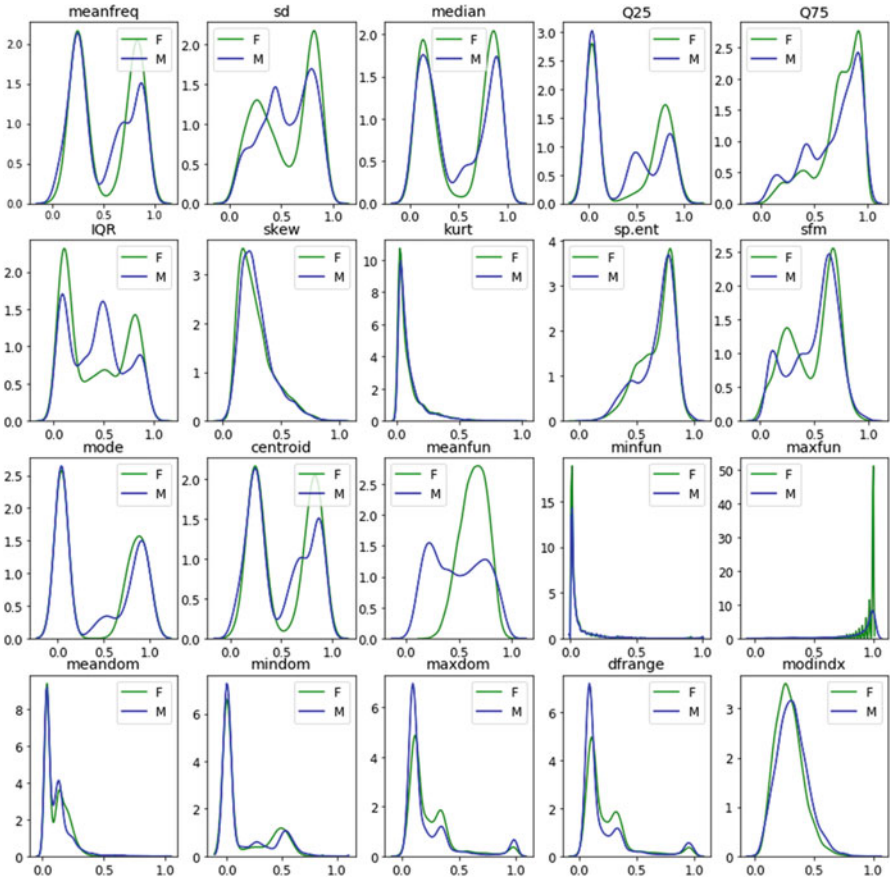

Fig. 2 (continued)

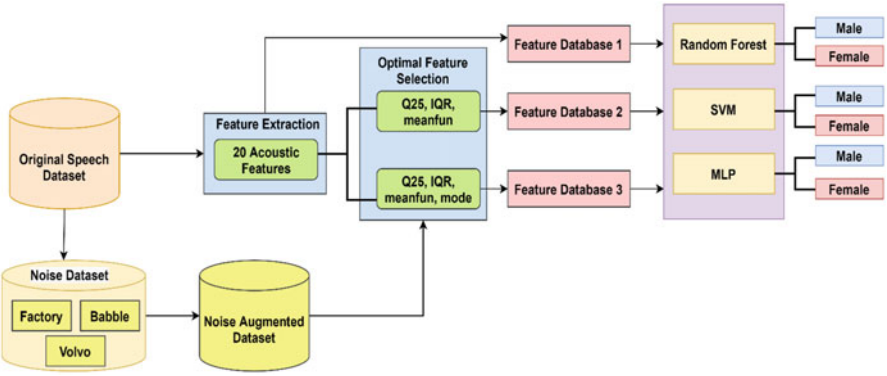

Fig. 3 Block diagram of the proposed gender classification system through optimal acoustic feature selection using noise augmentation

5.1 Performance Evaluation on Clean Dataset

It can be noted that from the outset of extracted features corresponding to the clean dataset as described in Fig. 2a, the three most significant extracted features comparing both male and female audios have been found to be Q25, IQR, and meanfun. Therefore, the comparative analysis for semi-supervised algorithms has been experimented based on the two types of feature selections: one with 20 features all together and the other with three distinctive features as shown in Table 1. In first set of experiments, 20 features resulted into an improved performance than that of ideal selection of three acoustic features. In, order to identify the likelihood of the certain part combinations of all features corresponding to audio set is performed. In addition, given the nonlinear data set corresponding to audio signals, SVM has done no better than the classification method of random forest. However, better performance on the radial basis function (rbf) kernel has been identified with an ideal selection of features with an accuracy of 82.04% over 81.92% with 20 features. Furthermore, 87.28% accuracy utilizing three optimally selected features in case of MLP has outperformed both SVM and random forest classification techniques with an overall RI of 6.54% in case of clean audio dataset.

5.2 Performance Evaluation on Noise-Augmented Dataset

It can be noted that from the outset of extracted features corresponding to the noise-augmented dataset as described in Fig. 2b, the four most significant extracted features comparing both male and female noise-augmented audio sets have been found to be Q25, IQR, mode, and meanfun. Thus, based on baseline results on three preselected optimal feature selections, further experiments on the noise-augmented dataset were conducted with these four most important acoustic features as shown in Table 2. The same spectrum of performances for both random forest and SVM classification techniques is very evident even in case of noisy data. However,

Table 1 Performance evaluation of classification algorithms on clean male and female dataset

	Accuracy (%)	
	20 features	3 features
Random forest	83.04	80.86
SVM	81.92	82.04
MLP	85.56	87.28

Table 2 Performance evaluation of classification algorithms on clean and noise-augmented dataset

	Noise	Accuracy (%)	
		3 features	4 features
Random forest	Yes	86.59	83.28
SVM	Yes	81.80	83.46
MLP	Yes	90.52	92.58
	No	87.28	87.42

random forest classification technique on noise-augmented dataset with 86.59% in Table 2 has outperformed MLP with 85.56% accuracy as in Table 1 utilizing 20 features. Furthermore, two more experiments on MLP classifier with or without noise have shown the relevance of mode frequency parameter such that the classifier utilizing four acoustic features has outperformed the classifier utilizing three acoustic features with an RI of 0.16% on clean dataset and an RI of 2.27% on noise-augmented dataset. Hence, the overall RI of 8.21% in comparison to baseline system has resulted in the development of adequate model classification system for male and female voice.

6 Conclusion

Performing audio analysis can become strenuous while selecting adequate features that can help resolving the cause. Out of numerous features explored, the study found Q25, IQR, and meanfun were able to draw accurate distinction between male and female speakers. Augmentation was applied for creating a noise-robust model alongside adding variability to dataset. After augmenting the dataset, the contour analysis was performed, and this time, mode frequency feature was also included for training of the model and yielded out better performance. MLP outperformed random forest and SVM algorithm, and 8.21% of RI was observed. Using noise-augmented dataset, selection of four features yielded an RI of 6.07%. The research presents opportunity to explore further permutation and combination of feature alongside increasing the corpus. Also, it opens the doors for extending the proposed system for other research areas like age group detection and native and non-native speaker detection.

Conflict of Interest: The authors declare that they have no conflict of interest.

References

1. Z. Zhang, Mechanics of human voice production and control. J. Acoust. Soc. Am. **140**(4), 2614–2635 (2016). https://doi.org/10.1121/1.4964509
2. J.A. Gómez-García, L. Moro-Velázquez, J.D. Arias-Londoño, J.I. Godino-Llorente, On the design of automatic voice condition analysis systems. Part III: Review of acoustic modelling strategies. Biomed. Signal Process. Contr. **66**, 102049 (2021). https://doi.org/10.1016/j.bspc.2020.102049
3. B. Delgutte, N.Y. Kiang, Speech coding in the auditory nerve: I. vowel-like sounds. J. Acoust. Soc. Am. **75**(3), 866–878 (1984). https://doi.org/10.1121/1.390596
4. D.G. Sinex, C.D. Geisler, Responses of auditory-nerve fibers to consonant–vowel syllables. J. Acoust. Soc. Am. **73**(2), 602–615 (1983). https://doi.org/10.1121/1.389007
5. E.D. Young, M.B. Sachs, Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. J. Acoust. Soc. Am. **66**(5), 1381–1403 (1979). https://doi.org/10.1121/1.383532

6. J.K. Bizley, K.M. Walker, Sensitivity and selectivity of neurons in auditory cortex to the pitch, timbre, and location of sounds. Neuroscientist **16**(4), 453–469 (2010). https://doi.org/10.1177/1073858410371009
7. Hermansky H, Sharma S (1999) Temporal Patterns (TRAPS) in ASR of Noisy Speech. In 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258) (Vol. 1, pp. 289–292). IEEE Phoenix, AZ
8. R. Hu, V. Krishnan, D.V. Anderson, Speech Bandwidth Extension by Improved Codebook Mapping Towards Increased Phonetic Classification, in *Ninth European Conference on Speech Communication and Technology*, (Interspeech, Lisbon, 2005)
9. M. Koo, J. Jeon, H. Moon, M.W. Suh, J.H. Lee, S.H. Oh, M.K. Park, Effects of noise and serial position on free recall of spoken words and pupil dilation during encoding in Normal-hearing adults. Brain Sci. **11**(2), 277 (2021). https://doi.org/10.3390/brainsci11020277
10. M.I. Jordan, T.M. Mitchell, Machine learning: Trends, perspectives, and prospects. Science **349**(6245), 255–260 (2015). https://doi.org/10.1126/science.aaa8415
11. J. Kaur, A. Singh, V. Kadyan, Automatic speech recognition system for tonal languages: State-of-the-art survey. Arch. Comput. Method. Eng., 1–30 (2020a). https://doi.org/10.1007/s11831-020-09414-4
12. M.H. Moattar, M.M. Homayounpour, A review on speaker diarization systems and approaches. Speech Comm. **54**(10), 1065–1103 (2012). https://doi.org/10.1016/j.specom.2012.05.002
13. D. Raj, P. Denisov, Z. Chen, H. Erdogan, Z. Huang, M. He, et al., Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis. arXiv preprint **arXiv**, 2011.02014 (2020)
14. S.I. Levitan, T. Mishra, S. Bangalore, Automatic Identification of Gender from Speech, in *Proceeding of Speech Prosody*, (Semantic Scholar, 2016), pp. 84–88
15. P. Bawa, V. Kadyan, Noise robust in-domain children speech enhancement for automatic Punjabi recognition system under mismatched conditions. Appl. Acoust. **175**, 107810 (2021). https://doi.org/10.1016/j.apacoust.2020.107810
16. P. Sarma, S.K. Sarma, Syllable based approach for text to speech synthesis of Assamese language: A review. J. Phys. Conf. Series **1706**(1), 012168 (2020) IOP Publishing
17. A. Singh, V. Kadyan, M. Kumar, N. Bassan, ASRoIL: A comprehensive survey for automatic speech recognition of Indian languages. Artif. Intell. Rev., 1–32 (2019). https://doi.org/10.1007/s10462-019-09775-8
18. Y. Kumar, N. Singh, M. Kumar, A. Singh, AutoSSR: An efficient approach for automatic spontaneous speech recognition model for the Punjabi language. Soft. Comput. **25**(2), 1617–1630 (2021). https://doi.org/10.1007/s00500-020-05248-1
19. S. Thomas, M.L. Seltzer, K. Church, H. Hermansky, Deep Neural Network Features and Semi-supervised Training for Low Resource Speech Recognition, in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, (IEEE, 2013), pp. 6704–6708. https://doi.org/10.1109/ICASSP.2013.6638959
20. M. Araya-Salas, G. Smith-Vidaurre, warbleR: An R package to streamline analysis of animal acoustic signals. Methods Ecol. Evol. .. **8**(2), 184–191 (2017). https://doi.org/10.1111/2041-210X.12624
21. Y. Zhan, X. Yuan, Audio Post-processing Detection and Identification based on Audio Features, in *2017 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)*, (IEEE, Ningbo, China, 2017), pp. 154–158. https://doi.org/10.1109/ICWAPR.2017.8076681
22. G. Li, Y. Liu, The Analysis on the Acoustic Parameters of Distinctive Features for Mandarin Vowels, in *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, (IEEE, Shanghai, China, 2017), pp. 1–5. https://doi.org/10.1109/CISP-BMEI.2017.8302104

23. Y. Singh, A. Pillay, E. Jembere, Features of Speech Audio for Accent Recognition, in *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, (IEEE, Durban, South Africa, 2020), pp. 1–6. https://doi.org/10.1109/icABCD49160.2020.9183893
24. R. Hidayati, I.K.E. Purnama, M.H. Purnomo, The Extraction of Acoustic Features of Infant Cry for Emotion Detection Based on Pitch and Formants, in *International Conference on Instrumentation, Communication, Information Technology, and Biomedical Engineering 2009*, (IEEE, Bandung, Indonesia, 2009), pp. 1–5. https://doi.org/10.1109/ICICI-BME.2009.5417242
25. S. Poornima, N. Sripriya, B. Vijayalakshmi, P. Vishnupriya, Attendance Monitoring System Using Facial Recognition with Audio Output and Gender Classification, in *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP)*, (IEEE, Chennai, India, 2017), pp. 1–5. https://doi.org/10.1109/ICCCSP.2017.7944103
26. D. Stewart, H. Wang, J. Shen, P. Miller, Investigations into the Robustness of Audio-Visual Gender Classification to Background Noise and Illumination Effects, in *2009 Digital Image Computing: Techniques and Applications*, (IEEE, Melbourne, VIC, 2009), pp. 168–174. https://doi.org/10.1109/DICTA.2009.34
27. Y.M. Zeng, Z.Y. Wu, T. Falk, W.Y. Chan, Robust GMM based Gender Classification Using Pitch and RASTA-PLP Parameters of Speech, in *2006 International Conference on Machine Learning and Cybernetics*, (IEEE, Dalian, China, 2006), pp. 3376–3379. https://doi.org/10.1109/ICMLC.2006.258497
28. A. Copiaco, C. Ritz, N. Abdulaziz, S. Fasciani, Identifying Optimal Features for Multi-channel Acoustic Scene Classification, in *2019 2nd International Conference on Signal Processing and Information Security (ICSPIS)*, (IEEE, Dubai, 2019), pp. 1–4. https://doi.org/10.1109/ICSPIS48135.2019.9045907
29. P. Gupta, S. Goel, A. Purwar, A Stacked Technique for Gender Recognition Through Voice, in *2018 Eleventh International Conference on Contemporary Computing (IC3)*, (IEEE, Noida, India, 2018), pp. 1–3. https://doi.org/10.1109/IC3.2018.8530520
30. H. Harb, L. Chen, Gender Identification Using a General Audio Classifier, in *2003 International Conference on Multimedia and Expo.ICME'03.Proceedings (Cat.No. 03TH8698)*, vol. 2, (IEEE, Baltimore, MD, 2003), p. II-733. https://doi.org/10.1109/ICME.2003.1221721
31. T.W. Sun, End-to-end speech emotion recognition with gender information. IEEE Access **8**, 152423–152438 (2020). https://doi.org/10.1109/ACCESS.2020.3017462
32. E. Tanuar, E. Abdurachman, F.L. Gaol, Analysis of Gender Identification in Bahasa Indonesia using Supervised Machine Learning Algorithm, in *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, (IEEE, Yogyakarta, Indonesia, 2020), pp. 421–424. https://doi.org/10.1109/ICOIACT50329.2020.9332145
33. G.R. Nitisara, S. Suyanto, K.N. Ramadhani, Speech Age-Gender Classification Using Long Short-Term Memory, in *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, (IEEE, Yogyakarta, Indonesia, 2020), pp. 358–361. https://doi.org/10.1109/ICOIACT50329.2020.9331995
34. R.D. Alamsyah, S. Suyanto, Speech Gender Classification Using Bidirectional Long Short Term Memory, in *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, (IEEE, Yogyakarta, Indonesia, 2020), pp. 646–649. https://doi.org/10.1109/ISRITI51436.2020.9315380
35. S.A. Kokatnoor, B. Krishnan, Twitter Hate Speech Detection using Stacked Weighted Ensemble (SWE) Model, in *2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, (IEEE, Bangalore, India, 2020), pp. 87–92. https://doi.org/10.1109/ICRCICN50933.2020.9296199

36. E.E. Kalaycı, B. Doğan, Gender Recognition by Using Acoustic Features of Sound With Deep Learning and Data Mining Methods, in *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, (IEEE, Istanbul, Turkey, 2020), pp. 1–4. https://doi.org/10.1109/ASYU50717.2020.9259824
37. P. Mishra, R. Sharma, Gender Differentiated Convolutional Neural Networks for Speech Emotion Recognition, in *2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, (IEEE, Brno, Czech Republic, 2020), pp. 142–148. https://doi.org/10.1109/ICUMT51630.2020.9222412
38. V. Kadyan, S. Bala, P. Bawa, Training augmentation with TANDEM acoustic modelling in Punjabi adult speech recognition system. Int. J. Speech Technol. **24**, 473–481 (2021). https://doi.org/10.1007/s10772-021-09797-0
39. A. Varga, H.J. Steeneken, Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Comm. **12**(3), 247–251 (1993). https://doi.org/10.1016/0167-6393(93)90095-3