INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Do Vocal Breath Sounds Encode Gender Cues for Automatic Gender Classification?

Mohammad Shaique Solanki[1], *Ashutosh Bharadwaj*[2], *Jeevan Kylash*[1], *Prasanta Kumar Ghosh*[1]

[1]Indian Institute of Science, Bangalore, India
[2]R V College of Engineering, Bangalore, India

solankishaique@gmail.com, ashutosh.bharadwaj99@gmail.com, jeevankylash@outlook.com, prasantg@iisc.ac.in

Abstract

The acoustic features of continuous speech, such as pitch (F0) and formant frequencies (F1, F2) have been utilized for gender classification. However, non-speech signals including vocal breath sounds have not been explored due to the absence of gender-specific acoustic features. This study investigates if vocal breath sounds carry gender information and if they can be used for automatic gender classification. The study examines the use of data-driven and knowledge-based features from breath sounds, classifier complexity, and the importance of breath signal segment location and duration. Results from experiments on 54 minutes of male and 52 minutes of female breath sounds demonstrate that classifiers with low-complexity and knowledge-based features (MFCC statistics) perform similarly to high-complexity classifiers with data-driven features. Breath segments of around 3 seconds are found to be the most suitable choice regardless of location, eliminating the need for breath cycle boundary marking.

Index Terms: Vocal Breath Sound Signals, Automatic Gender Classification, CNN-LSTM, Mel-Spectrogram, MFCC Statistics

1. Introduction

Humans produce different sounds of which some are used for communication, like speech, whereas others are for non-communication purposes such as cough, breath sounds, etc. The sounds used for communication reveal speaker characteristics such as their emotions, gender, age, etc [1]. Speech signals such as continuous speech are known to have acoustic features such as pitch(F0), and formant frequencies(F1, F2) which can be used for gender classification[2][3][4]. Usually, males have lower pitch and formant frequencies than females during continuous speech[5][6]. A considerable amount of work uses voiced speech signals such as continuous speech to exploit the difference in the acoustic features for gender classification. For example, S Bhukya[7] used articulatory cues such as pitch(F0) and formant frequencies(F1, F2, F3) extracted from continuous speech samples to develop a gender classification model to improve Automatic Speech Recognition(ASR) performance. Similar works are found in the literature where pitch(F0) information is used to develop gender classification models [8][9][10]. Anna V Kuchebo used Mel Frequency Cepstral Coefficients(MFCC) and spectral contrast on continuous speech to classify gender[11]. S Levitan et al [12] used the combination of pitch information(F0) and MFCC to classify gender using continuous speech samples. This work was further used by Kabil et al [13] who used Convolutional Neural Networks (CNN) for gender classification on raw speech using MFCC only. Rangga et al [14] used a Bidirectional Long Short Term Memory network to classify gender using a voiced dataset. All these works use either spontaneous speech and/or other types of voice speech signals.

This study aims to explore whether vocal breath sound signals contain gender-related information and, if so, how this information can be extracted and utilized for the purpose of automated gender classification. It should be noted that the purpose of this study is not to directly compare our approach with existing speech-based gender classification models. Rather, our focus is on investigating the potential of using vocal breath sounds as a source of gender cues for gender classification.

The development of a gender classification model based on vocal breath sounds can be an important prerequisite for the implementation of automatic Pulmonary Function Test (PFT) variable prediction systems that rely solely on breath sounds. As lung size and capacity vary between males and females of the same age and height, PFT parameters such as Forced Vital Capacity (FVC), Total Lung Capacity (TLC), and Forced Expiratory Volume in one second (FEV1) may exhibit higher values in males than in females[15][16][17]. FVC reflects the maximum amount of air that an individual can forcefully and completely exhale after taking a deep breath, while FEV1 corresponds to the amount of air that an individual can forcefully exhale in one second after taking a deep breath. These gender-specific differences in PFT parameters can be accounted for by the vocal breath sound-based gender classification model. These classifications can then be utilized by PFT variable prediction systems where gender information can increase accuracy and reliability. Our hypothesis is that vocal breath sound signals contain gender cues that can be learned by a neural network and used for automatic gender classification.

It should be re-noted that goal of this work is not to compare vocal breath sound-based gender classification with speech-based ones. Rather investigate **five questions**: **1)** Are gender cues present in the mel-spectrogram of the breath cycle? **2)** Do MFCC statistics features encode spectral characteristics of gender which can be used for automatic gender classification with reduced parameters and training time? **3)** What is the role of breath boundaries in gender classification? **4)** What is the effect of the number of frames in a breath chunk on gender classification accuracy? Finally, **5)** What is the effect of taking random frames from the entire breath audio on gender classification?

To investigate the 1st question, we employed a 2-D Convolutional Neural Network (CNN) to learn the spectro-temporal features present in the mel-spectrogram of a breath cycle for the purpose of gender classification. A breath cycle is defined as a single inhalation followed by exhalation by a subject. Due to the complexity of this task and the large number of parameters involved, the resulting model requires significant computational resources and training time.

In order to address the 2nd question, we performed calcula-

10.21437/Interspeech.2023-2190

tions on four 13-feature MFCC statistics - mean, median, mode, and standard deviation, resulting in a total of 52(=13x4) features. These features were computed for a complete breath cycle, utilizing breath boundaries. We employed a 1-D CNN LSTM (Long Short Term Memory) model to acquire information from the 52 MFCC statistics for gender classification. While CNN layer extracts features across a complete breath cycle, LSTM layer aids in sequence prediction and classification. The model effectively assimilated cues from the MFCC statistics, with a reduced number of parameters and training time, representing a knowledge-based model with low complexity.

To investigate 3rd question, we extracted segments of breath sound with varying lengths from the complete audio recording of each subject's breathing (which consisted of multiple breath cycles) and subsequently trained 1-D CNN LSTM models on these segments. Our aim was to compare the performance of these models with the performance of the model trained specifically on breath cycle boundaries.

To address the 4th question we selected a random sample of frames from the most effective chunk identified in the 3rd question, in order to examine accuracy trends. In instances where the model's accuracy decreased for specific frame numbers, it indicated that the model was unable to utilize MFCC statistical features to discern gender cues for classification.

Finally, for the 5th question, the 1-D CNN LSTM model was trained on 100 randomly selected frames from the entire breath audio file. As the model was able to access the MFCC statistics across all frames from the entire breath recording, we hypothesized that the accuracy of this experiment would surpass that of both the entire breath cycle and the continuous chunks utilized in the second and third questions respectively. This experiment aimed to improve model generalization by relaxing experimental conditions across the entire breath recording.

2. Dataset

The data used for this study consists of audio files pertaining to 106 subjects (55 Male, 51 Female). Each audio file has vocal breath, followed by cough, sustained vowels: /ɑː/, /iː/, /uː/, /ɔː/, /eː/, and sustained fricatives: /sː/, /zː/. The audio data was recorded with a sampling rate of 44.1 kHz using a Zoom H6 microphone at a hospital in non-laboratory conditions with ambient and background noise, under the guidance of a pulmonologist. The microphone was placed approximately 10 cm away from the subject's mouth. The data was collected over a span of 4 years, from 2016 to 2019, and was used in the study of breath characteristics of healthy (control) and asthmatic subjects[18][19]. The audio data was labeled to indicate boundaries of inhalation, exhalation, and breath cycles. For this study, the breath segments of the subjects were used.

3. Experiments and Results

Inhale and exhale parts of the breath cycle are boundary labeled along with the entire breath cycle as shown in Fig.1.

Experimental Setup: Mel-spectrogram images and 52 MFCC statistics pertaining to 106 subjects are divided into 5 gender-balanced folds with 21 subjects in each fold except fold 5 which has 22 subjects. Out of 5 folds, one fold is used for testing. The remaining 4 folds contain 85 subjects to be used for training and validation. With the fold boundaries held in place, 4 males and 4 females (randomly chosen) are used for validation after the model trains on the 77 subjects. This is repeated five times in a five-fold cross-validation setting where each fold is tested in a round-robin fashion. There is no overlap between

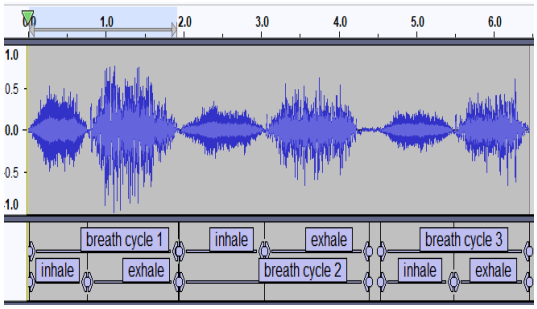

Figure 1: *Breath cycle annotation scheme*

any of the sets. The models, 2-D CNN, and 1-D CNN LSTM are fit on the train set, fold-wise and hyperparameter tuned on the validation set with early stopping criteria on model loss with the patience of 40 epochs and tested on the test set.

Evaluation Metrics : The evaluation metrics used in this study is segment-level accuracy, F1 score, precision, and recall. Predicted gender labels for each segment in the test set are compared with the true label or the ground truth to calculate the above-mentioned metrics.

Experimental road-map : The experimental road-map of this paper is shown in Fig.2.

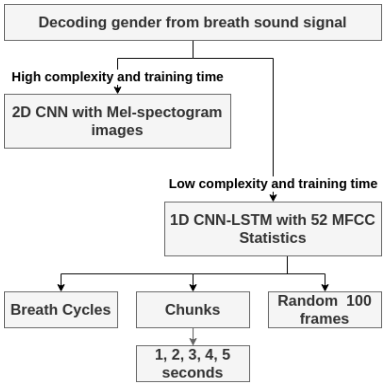

Figure 2: *Experimental pipeline to address the five questions*

3.1. Gender decoding using 2-D CNN on mel-spectrogram

To perform gender classification using mel-spectrogram images, we utilized a fully connected 2-D CNN with the architecture displayed in Fig. 3. We standardized the extracted breath segments of each subject by zero-padding them to the maximum length of breath signal present in the dataset. We then computed the mel-spectrogram of these extracted breath cycles using the Librosa Speech Processing toolkit[20] with 128 mel-filters (n-mels), a window length of 20ms, and a hop length of 10ms. Subsequently, we converted these spectrograms into RGB images with a dimension of 128x128. The confusion matrix was computed fold-wise and average accuracy was calculated to be **0.77±0.07** as shown in Table 1 along with other metrics.

The 2-D CNN model is a complex model in terms of parameters totaling 2,193,729, out of which learnable and non-learnable parameters are 2,192,321 and 1,408 respectively. The average training time for the 2-D CNN model was 5.4 seconds per epoch across all 5 folds.

3.2. Gender decoding using 1-D CNN LSTM Model

As previously mentioned, while the 2-D CNN model can learn the spectro-temporal features of the mel-spectrogram, it is a

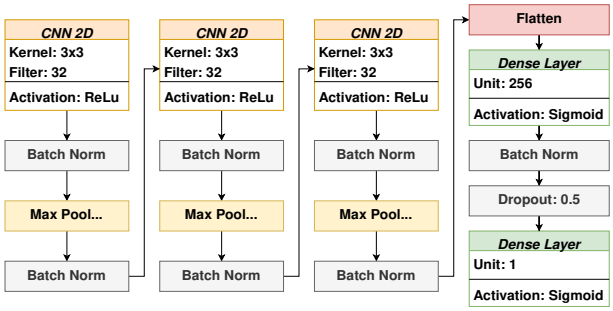

Figure 3: *Architecture of the 2-D CNN model used to address Question 1.*

Table 1: *Confusion Matrices of the 2-D CNN model with mel-spectrogram. Labels in red indicate true labels whereas blue indicate predicted labels.*

	F1		F2		F3		F4		F5	
	M	F	M	F	M	F	M	F	M	F
M	74	12	57	45	72	18	77	18	111	26
F	16	66	23	75	19	72	24	75	36	82
acc	0.83		0.66		0.80		0.78		0.76	
F1	0.83		0.69		0.80		0.78		0.73	
prec	0.85		0.63		0.80		0.81		0.76	
recall	0.80		0.77		0.79		0.76		0.69	
Acc: 0.77 ± 0.6, F1: 0.76 ± 0.05, Precision: 0.77 ± 0.08, Recall: 0.76 ± 0.04										
2D CNN on mel spectrogram Images										

complex classifier. Therefore, we employed a 1-D CNN LSTM model to reduce the classifier complexity in terms of total parameters and training time per epoch, while maintaining classification accuracy. Our fully connected 1-D CNN-LSTM network, architecture of which is shown in Fig.4, comprises CNN layer(s) for feature extraction on the input data, which, in our study, are 52 MFCC statistics calculated across the breath cycle, followed by Long Short Term Memory (LSTM) layers for sequence prediction/classification. To explore the extent to which the vocal breath signals encode gender information, we treated each of the 52 MFCC statistics as individual time stamps, allowing the model to learn hidden spectral characteristics across feature vectors.

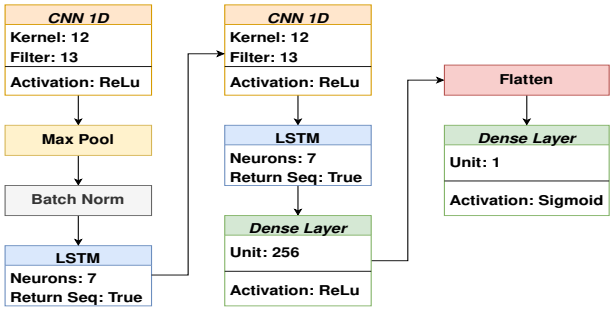

Figure 4: *Architecture of 1-D CNN LSTM model to address Question(s) 2, 3, 4 and 5).*

Confusion matrix was computed fold wise and average accuracy was calculated to be **0.76±0.12** as shown in Table 2 along with F1 Score, Precision, and Recall.

The 1-D CNN LSTM model is significantly lighter compared to the 2-D CNN model discussed in Section 3.1. It comprises only 2,138 trainable parameters out of a total of 2164 parameters and has an average training time of 200ms per epoch across all five folds. By adopting the 1-D CNN LSTM architecture for gender cue learning across 52 MFCC statistical features, we were able to reduce model complexity in terms of total parameters and training time without compromising classification accuracy substantially.

Table 2: *Confusion Matrices of the 1-D CNN LSTM model with 52 MFCC Stats. Labels in red indicate true labels whereas blue indicate predicted labels.*

	F1		F2		F3		F4		F5	
	M	F	M	F	M	F	M	F	M	F
M	65	21	47	55	90	1	80	16	117	20
F	21	62	34	64	25	66	20	79	21	97
acc	0.75		0.555		0.86		0.82		0.84	
F1	0.75		0.59		0.84		0.81		0.83	
prec	0.75		0.54		0.99		0.83		0.83	
recall	0.75		0.65		0.73		0.80		0.82	
Acc: 0.76±0.12, F1: 076±0.10, Precision: 0.79±0.16, Recall: 0.75±0.07										
1-D CNN LSTM on MFCC Stats of entire breath cycle										

3.3. Role of breath boundary in gender classification

To investigate the relative importance of breath boundaries in gender classification, we analyzed chunks of various durations extracted from breath audio. The dataset had an average breath cycle duration of 3.13 seconds, therefore we extracted chunks of ±1 sec and ±2 sec from the average duration, resulting in chunk lengths of 1, 2, 3, 4, and 5 seconds. Across these chunks, we calculated 52 MFCC statistics, taking care to use the same number of segments for each chunk length as there were breath cycles for each subject. We then retrained the 1-D CNN LSTM model with these MFCC statistics. The resulting fold-wise confusion matrices and evaluation metrics are presented in Table 3, ensuring experimental uniformity.

Our results indicate that gender classification using **3-second** breath chunks had the highest chunk-level accuracy of **0.77±0.11**, which is a 0.1 improvement from the breath cycle experiment. The **5-second** breath chunks had a similar performance to the breath cycle experiment, with a chunk-level accuracy of **0.76±0.11**.

3.4. Effect of number of frames in a chunk on classification accuracy

To understand whether there may be any redundancies in the frames of a chunk, we randomly chose 10, 50, 100, and 200 frames for the 3-second breath chunk (best performing) which has 300 frames. The model was re-trained again on these randomly chosen frames. Table 4 summarises the effect of chosen frames from within the 3-sec chunk on model accuracy.

3.5. Random frame-based classification

In this section, in contrast to subsection 3.3, we randomly take 100 frames(discontinuous) from the entire breath recording of the subject. The number of segments of 100 frames taken is equal to the number of breath cycles of the subject to maintain uniformity similar to subsection 3.3. This increases the span of audio frames that are used to train the 1-D CNN LSTM model. The chunk-level confusion matrices and accuracy of each fold are shown in Table 5 along with other metrics.

We conclude that N segments of 100 frames taken randomly from the entire breath recording, where N is the number of breath cycles of the subject, increases the sample space of frames on which MFCC statistics are calculated. This improves the model accuracy compared to the case where MFCC statistics are calculated across continuous frames to **0.80±0.10** from 0.77±0.11(3-sec chunk) and 0.76±0.12(entire breath cycle).

Table 3: *Confusion Matrices of 1-D CNN LSTM model with different chunks. Labels in red indicate true labels whereas blue indicates predicted labels.*

	F1		F2		F3		F4		F5	
	M	F	M	F	M	F	M	F	M	F
M	67	19	41	61	73	17	78	17	113	24
F	16	66	25	73	24	67	22	77	26	92
acc	0.79		0.57		0.77		0.80		0.80	
F1	0.79		0.63		0.77		0.80		0.79	
prec	0.78		0.54		0.80		0.82		0.79	
recall	0.80		0.74		0.74		0.78		0.78	
Acc: 0.75±0.10, F1: 0.75±0.07, Precision: 0.75±0.11, Recall: 0.77±0.03										
1-D CNN LSTM on 1 sec chunk										
	F1		**F2**		**F3**		**F4**		**F5**	
	M	F	M	F	M	F	M	F	M	F
M	74	12	38	64	72	18	79	16	105	32
F	27	55	30	68	18	73	33	66	14	104
acc	0.77		0.53		0.80		0.75		0.82	
F1	0.74		0.59		0.80		0.73		0.82	
prec	0.82		0.52		0.80		0.80		0.76	
recall	0.67		0.69		0.80		0.67		0.88	
Acc: 0.73±0.12, F1: 0.74±0.09, Precision: 0.74+0.13, Recall: 0.74±0.10										
1-D CNN LSTM on 2 sec chunk										
	F1		**F2**		**F3**		**F4**		**F5**	
	M	F	M	F	M	F	M	F	M	F
M	63	23	24	78	89	1	71	24	124	13
F	10	82	5	89	29	61	15	92	35	68
acc	0.81		0.58		0.83		0.81		0.8	
F1	0.83		0.68		0.80		0.83		0.74	
prec	0.78		0.53		0.98		0.79		0.84	
recall	0.89		0.95		0.68		0.86		0.66	
Acc: 0.77±0.11, F1:0.78±0.06, Precision: 0.79±0.16, Recall: 0.81±0.13										
1-D CNN LSTM on 3 sec chunk										
	F1		**F2**		**F3**		**F4**		**F5**	
	M	F	M	F	M	F	M	F	M	F
M	66	20	30	72	78	12	72	23	104	33
F	19	73	17	77	27	63	16	91	25	78
acc	0.78		0.55		0.78		0.81		0.76	
F1	0.79		0.63		0.76		0.82		0.73	
prec	0.78		0.52		0.84		0.80		0.70	
recall	0.79		0.82		0.70		0.85		0.76	
Acc: 0.74±0.11, F1: 0.75±0.07, Precision: 0.73±0.13										
1-D CNN LSTM on 4 sec chunk										
	F1		**F2**		**F3**		**F4**		**F5**	
	M	F	M	F	M	F	M	F	M	F
M	71	15	31	57	90	0	77	18	118	19
F	21	71	27	67	24	66	21	86	32	71
acc	0.79		0.58		0.86		0.80		0.78	
F1	0.80		0.61		0.85		0.82		0.74	
prec	0.83		0.54		1.00		0.83		0.79	
recall	0.77		0.71		0.73		0.80		0.69	
Acc: 0.76±0.11, F1: 0.76±0.09, Precision: 0.80±0.16, Recall: 0.74+0.05										
1-D CNN LSTM on 5 sec chunk										

Table 4: *No. of randomly chosen frames vs accuracy of 1-D CNN LSTM for 3-sec chunk*

Frames	10	50	100	200	300
Model Accuracy	0.73	0.75	0.76	0.77	0.77

4. Discussion

This section analyzes the results of the experiments mentioned in Section 3 to answer the five questions we defined above.

1) Are gender cues present in the mel-spectrogram images of breath cycle?: Yes, the 2-D CNN model trained on mel-spectrogram images of breath cycles has an average chunk-level accuracy of 0.77±0.06 confirming this. Substantiating that the model is able to learn the gender-specific spectro-temporal cues captured by the spectrogram.

2) Do MFCC statistics features encode spectral characteristics of gender which can be used for automatic gender classification with reduced parameters and training time?: Yes, we were able to cut down the number of parameters and training time by using the 1-D CNN LSTM model trained on 52 MFCC statistics without reducing classification accuracy. The average accuracy of the model trained on the breath cycle is 0.76±0.12. This shows that the 1-D CNN LSTM model can learn the gender-specific spectral cues from the 52 MFCC statistics to classify gender using breath cycle audios.

Table 5: *Confusion Matrices of the 1-D CNN LSTM model with MFCC statistics across 100 random frames from the entire breath audio file. Labels in red indicate true labels whereas blue indicates predicted labels.*

	F1		**F2**		**F3**		**F4**		**F5**	
	M	F	M	F	M	F	M	F	M	F
M	71	16	54	57	79	5	84	23	97	23
F	9	67	22	76	24	62	8	97	13	108
acc	0.85		0.62		0.83		0.85		0.85	
F1	0.84		0.66		0.81		0.86		0.86	
prec	0.81		0.57		0.93		0.81		0.82	
recall	0.88		0.78		0.72		0.92		0.89	
Acc: 0.80±0.10, F1: 0.81±0.09, Precision: 0.79±0.13, Recall: 0.84±0.09										
1-D CNN LSTM on Random 100 Frames from entire breath										

3) What is the role of breath boundaries in gender classification?: We used chunks with durations 1sec, 2sec, 3sec, 4sec, and 5sec to compare model accuracy with the one with breath boundary. It is observed that the classification accuracy of the 1-D CNN LSTM model was best for 3-sec chunks(0.77±0.11), followed by 5-sec chunks(0.76±0.11) which is similar to the data-driven model used in the first experiment. The reason behind this may be that 3sec and 5sec chunks have lesser redundancies along the frames compared to other chunks. The fact that the average duration of a breath cycle across the dataset is 3.13-sec supports the observation that 3sec chunks are doing better than other chunks.

4) What is the effect of the number of frames in a breath chunk on gender classification accuracy?: It is observed from Table 4 that decreasing the number of frames within a chunk led to a decrease in the model's accuracy. This is because the model relies on frame-level data to calculate 52 MFCC statistics that capture important spectral cues. Therefore, it is crucial to include as many frames as necessary to obtain an accurate representation of the data within the 3-second chunk, while avoiding any redundant or extraneous frames.

5) What is the effect of taking random frames from the entire breath audio on gender classification?: The model trained on 52 MFCC statistics across random 100 frames from entire breath audio recording outperforms all models trained on continuous frame chunks with segment-level accuracy of 0.80±0.10 as it increases the sample space of frames on which MFCC statistics are calculated.

5. Conclusion

In summary, our study demonstrates that vocal breath sounds do encode gender cues which can be extracted using both data-driven and knowledge-based features. Our findings suggest that using long-term features (features calculated over longer durations of the breath audio) results in a more accurate gender classification model than using short-term features(features calculated over shorter durations). To improve the model, we recommend increasing the size of the breath dataset. Overall, our study provides evidence that vocal breath sounds are a viable and accurate source of gender cues for gender classification.

6. References

[1] N. Gavriely, Y. Palti, and G. Alroy, "Spectral characteristics of normal breath sounds," *Journal of applied physiology*, vol. 50, no. 2, pp. 307–314, 1981.

[2] B. Khalighinejad, G. C. da Silva, and N. Mesgarani, "Dynamic encoding of acoustic features in neural responses to continuous speech," *Journal of Neuroscience*, vol. 37, no. 8, pp. 2176–2185, 2017.

[3] J. Liscombe, J. Venditti, and J. B. Hirschberg, "Classifying subject ratings of emotional speech using acoustic features," 2003.

[4] G. Chen, X. Feng, Y.-L. Shue, and A. Alwan, "On using voice source measures in automatic gender classification of children's speech," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[5] C. M. Sapienza, "Aerodynamic and acoustic characteristics of the adult africanamerican voice," *Journal of Voice*, vol. 11, no. 4, pp. 410–416, 1997.

[6] R. J. Morris, W. Brown Jr, D. M. Hicks, and E. Howell, "Phonational profiles of male trained singers and nonsingers," *Journal of Voice*, vol. 9, no. 2, pp. 142–148, 1995.

[7] S. Bhukya, "Effect of gender on improving speech recognition system," *International Journal of Computer Applications*, vol. 179, no. 14, pp. 22–30, 2018.

[8] B. D. Barkana and J. Zhou, "A new pitch-range based feature set for a speaker's age and gender classification," *Applied Acoustics*, vol. 98, pp. 52–61, 2015.

[9] 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.

[10] Y.-M. Zeng, Z.-Y. Wu, T. Falk, and W.-Y. Chan, "Robust gmm based gender classification using pitch and rasta-plp parameters of speech," in *International conference on machine learning and cybernetics*. IEEE, 2006, pp. 3376–3379.

[11] 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.

[12] S. I. Levitan, T. Mishra, and S. Bangalore, "Automatic identification of gender from speech," in *Proceeding of speech prosody*. Semantic Scholar, 2016, pp. 84–88.

[13] S. H. Kabil, H. Muckenhirn, and M. Magimai-Doss, "On learning to identify genders from raw speech signal using cnns." in *Interspeech*, 2018, pp. 287–291.

[14] R. D. Alamsyah and S. Suyanto, "Speech gender classification using bidirectional long short term memory," in *3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*. IEEE, 2020, pp. 646–649.

[15] S. Miller, J. Hankinson, V. Brusasco, F. Burgos, R. Casaburi, A. Coates, R. Crapo, P. Enright, C. van der Grinten, P. Gustafsson *et al.*, "Normalización de la espirometría," *Eur Respir J*, vol. 26, no. 2, pp. 319–38, 2005.

[16] A. T. Society *et al.*, "European thoracic society. standardisation of lung function testing. standardization of the measurement of lung volumes," *Eur Respir J*, vol. 26, pp. 511–22, 2005.

[17] R. Pellegrino, G. Viegi, V. Brusasco, R. Crapo, F. Burgos, R. Casaburi, A. Coates, C. Van der Grinten, P. Gustafsson, J. Hankinson *et al.*, "Ats/ers task force: standardisation of lung function testing," *Interpretative strategies for lung function tests. Eur Respir J*, vol. 26, pp. 948–968, 2005.

[18] S. Yadav, K. NK, D. Gope, U. M. Krishnaswamy, and P. K. Ghosh, "Comparison of cough, wheeze and sustained phonations for automatic classification between healthy subjects and asthmatic patients," in *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 1400–1403.

[19] S. Yadav, M. Keerthana, D. Gope, U. Maheswari K., and P. Kumar Ghosh, "Analysis of acoustic features for speech sound based classification of asthmatic and healthy subjects," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6789–6793.

[20] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015.