**VOCAL BREATH SOUND BASED GENDER CLASSIFICATION**

Mohammad Shaique Solanki, Ashutosh M Bharadwaj, Jeevan K, Prasanta Kumar Ghosh

SPIRE Laboratory, Indian Institute of Science, Bengaluru, India

**ABSTRACT**

Voiced speech signals such as continuous speech are known to have acoustic features such as pitch (F0), and formant frequencies (F1, F2, F3) which can be used for gender classification. However, gender classification studies using non-speech signals such as vocal breath sounds have not been explored as they lack typical gender-specific acoustic features. In this work, we explore whether vocal breath sounds encode gender information and if so, to what extent it can be used for automatic gender classification. In this study, we explore the use of data-driven and knowledge-based features from vocal breath sounds as well as the classifier complexity for gender classification. We also explore the importance of the location and duration of breath signal segments to be used for automatic classification. Experiments with 54.23 minutes of male and 51.83 minutes of female breath sounds reveal that knowledge-based features, namely MFCC statistics, with low-complexity classifier perform comparably to the data-driven features with classifiers of higher complexity. Breath segments with an average duration of 3 seconds are found to be the best choice irrespective of the location which avoids the need for breath cycle boundary annotation.

**Index Terms**— Speech Processing, Vocal Breath Sound Signals, Gender Classification, CNN-LSTM, Mel-Spectrogram, MFCC statistics

1. INTRODUCTION

Humans produce different sounds, some of them are used for communication, like speech, whereas others are for non-communication purposes such as cough, breath sounds, etc. The sounds used for communications 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, F3) 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) which are gender-dependent features, extracted from continuous speech samples to develop a gender classification model to improve ASR performance. Similar works are found in 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. Gender classification on breath sound signals has many applications especially in the field of forensics, to solve criminal cases if gender is unknown [15]. It can also be used by doctors in the healthcare industry to help diagnose diseases based on gender. However, whether non-speech signals such as vocal breath sound signals encode gender information is not known in the literature. Gender classification using vocal breath sound signals poses a challenge because they lack the presence of any kind of acoustic features such as pitch, and formant frequencies. Our hypothesis is that vocal breath sound signals should contain gender cues that can be learned by a neural network as the source of both speech and vocal breath sound is the same i.e., air passage through vocal tracts, the vibration of which produces the sounds.

In this study, we investigate five questions, 1) Are gender cues present in the mel-spectrogram of a 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 train a 2-D CNN model to learn spectro-temporal characteristics captured by the mel-spectrogram of a breath cycle to classify gender. The breath cycle is defined by: A single breath taken by a subject, which starts with inhale and ends with exhale. This is a data-driven model with high complexity in terms of total parameters and training time.

In order to address 2nd question, we calculate 4 MFCC statistics features to find the effect of taking random frames from the entire breath audio on gender classification accuracy. Finally, 5) 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?

We consider breath waveform chunks to carry experiments related to the 3rd question. We take chunks of breath sound with different lengths from the entire breath audio recording of the subjects (breath audio recording contains multiple breath cycles) and trained the 1-D CNN LSTM model on these chunks to compare their performance with the model trained on breath cycle boundaries.

To address the 4th question we randomly choose a set of frames from the best-performing chunk from the 3rd question to find the trend of accuracy. Lower model accuracy for a particular frame number meant that the model was not able to learn gender cues from...
MFCC statistics features to classify gender.

Finally, for the 5th question, we relaxed the experiment and let the model train on 100 randomly chosen frames from the entire breath audio file. Since the MFCC statistics across frames from the entire breath recording are visible to the model, the accuracy in this experiment should outperform all the continuous chunks chosen for the third question along with breath cycle boundaries.

2. DATASET

The data used for this study consists of an audio file of all 106 subjects. Each audio file has vocal breath, followed by cough, sustained vowels: /æ/, /i:/, /u:/, /ɔ/, and sustained fricatives: /s/, /z/. The audio data was recorded with a sampling rate of 44.1 kHz using a Zoom H6 microphone at St John’s Medical College Hospital in non-laboratory conditions with ambient and background noise, under the guidance of Dr. Uma Maheshwari. 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[16][12][13]. 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. The histogram shown in Fig.1 summarizes the distribution of breath cycle duration across male and female subjects.

![Fig. 1. Breath Count vs Breath Duration histogram.](image)

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.2.

![Fig. 2. Breath Cycle annotation scheme.](image)

**Experimental Setup:** Mel-spectrogram images and 52 MFCC statistics are calculated and divided into 5 gender-balanced folds. Each fold contains 21 subjects, 11 males and 10 females except fold 5 which contains 11 males and 11 females. The folds are further divided into test, train and validation set. There is no overlap between 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.

**Evaluation Metric:** The evaluation metric used in this study is segment-level accuracy. Predicted gender labels for each segment in the test set is compared with the true label or ground truth to calculate model accuracy.

**Experiment road-map:** The experiment road-map of this paper is shown in Fig. 3. The code for the experimental pipeline detailed in this paper is available at - https://github.com/CruelMarco/Breath_Fricative_Gender

![Fig. 3. Experimental pipeline to address the five questions](image)

3.1. Gender decoding using 2D CNN Mel-Spectrogram

For gender classification using mel-Spectrogram images, we use a fully connected 2-D CNN with the architecture shown in Fig 4. The extracted breath segments of each subject are zero-padded to the maximum length of breath signal present in the dataset to maintain uniformity. Mel-Spectrogram of extracted breath cycles is computed using the Librosa Speech Processing toolkit[18] with 128 mel-filters(n-mels), window length of 20ms, and hop length of 10ms which are then converted into RGB images of dimension 128x128.

The confusion matrix was computed fold-wise and average accuracy was calculated to be 0.77±0.07 as shown in Table 1

|   | F1 | F2 | F3 | F4 | F5 |
|---|----|----|----|----|----|
| M | 66 | 16 | 75 | 23 | 72 |
| F | 12 | 74 | 45 | 57 | 18 |
| acc | 0.83 | 0.66 | 0.80 | 0.78 | 0.76 |

![Table 1. Confusion Matrices of the 2-D CNN model with mel-spectrogram. Labels in red indicate true labels whereas blue indicates predicted labels.](image)

The 2-D CNN model as shown above in Fig 4 is a complex model in terms of total parameters. The total parameters of the
model are 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 CNN model was 5.4 seconds per epoch across all 5 folds.

3.2. Gender decoding using 1D CNN LSTM Model

As mentioned above, although the 2-D CNN model is able to learn the spectro-temporal characteristics of the mel-spectrogram, it is a complex classifier. The motivation to use the 1-D CNN LSTM model is to reduce the complexity of the model in terms of total parameters and training time per epoch without compromising classification accuracy. We use a fully connected 1-D CNN-LSTM network which consists of CNN layer(s) for feature extraction on input data, which in our study is 52 MFCC statistics, followed by Long Short Term Memory (LSTM) layers to support sequence prediction/classification.

The 52 MFCC statistics are calculated across the breath cycle of the subject. To explore the extent to which the vocal breath signals encode gender information, our methodology treats each of the 52 MFCC statistics as individual time stamps in order to make the model learn hidden spectral characteristics across feature vectors.

The flow chart of the CNN-LSTM model is shown in Fig 5.

A confusion matrix was computed fold wise and average accuracy was calculated to be 0.76±0.12 as shown in Table 2.

Table 2. Confusion Matrices of the 1-D CNN LSTM model with MFCC statistics across the entire breath cycle. Labels in red indicate true labels whereas blue indicate predicted labels.

|     | F1 | F2 | F3 | F4 | F5 |
|-----|----|----|----|----|----|
| M   | 65 | 21 | 1  | 80 | 117|
| F   | 21 | 62 | 34 | 25 | 20 |
| acc | 0.75 | 0.535 | 0.86 | 0.82 | 0.84 |

Breath Cycle gender classification: 0.76 ± 0.12

3.3. Role of breath boundary in gender classification

Breath boundary is defined as the boundaries of a breath cycle. To explore whether breath boundary is solely important for gender classification, we took out chunks of different lengths from the breath audio. The average duration of a breath cycle across the data set is 3.13 seconds. We took chunks of length ± 1 sec and ±2 sec from the average duration i.e. chunks of length(s) 1, 2, 3, 4, and 5 seconds from the entire breath recording and calculated 52 MFCC statistics across these chunks. To maintain uniformity in the experiments, the number of segments of respective lengths taken is equal to the number of breath cycles of the subject. We re-trained the 1-D CNN LSTM model with MFCC statistics across the above-mentioned chunks. The fold-wise confusion matrices and accuracies are summarized in Table 3.

From Table 3, it is observed that gender classification with 3-sec breath chunks is performing best with a chunk level accuracy of 0.77±0.11, which is an improvement from the breath cycle by 0.1. Followed by 5-sec breath chunks with 0.76±0.11, which is similar to the breath-cycle experiment.

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). The model was re-trained again on these randomly chosen frames. The average accuracy across the folds was calculated as mentioned in the above sections. Table 4 summarises the effect of chosen frames from within a 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 section 3.3. This increases the span of audio frames which are used to train the 1-D CNN LSTM model. The chunk-level confusion matrices and accuracies of each fold are shown in Table 5.

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.
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     | 67   | 19   | 41   | 61   | 73   |
| F     | 16   | 66   | 25   | 73   | 24   |
| acc   | 0.79 | 0.57 | 0.77 | 0.80 | 0.80 |

1 sec chunk gender classification: 0.75 ± 0.10

|       | F1   | F2   | F3   | F4   | F5   |
|-------|------|------|------|------|------|
| M     | 74   | 12   | 38   | 64   | 72   |
| F     | 27   | 55   | 30   | 68   | 18   |
| acc   | 0.77 | 0.53 | 0.80 | 0.75 | 0.82 |

2 sec chunk gender classification: 0.73 ± 0.12

|       | F1   | F2   | F3   | F4   | F5   |
|-------|------|------|------|------|------|
| M     | 63   | 23   | 24   | 89   | 17   |
| F     | 10   | 82   | 5    | 89   | 61   |
| acc   | 0.81 | 0.58 | 0.83 | 0.83 | 0.81 |

3 sec chunk gender classification: 0.77 ± 0.11

|       | F1   | F2   | F3   | F4   | F5   |
|-------|------|------|------|------|------|
| M     | 66   | 20   | 30   | 72   | 78   |
| F     | 19   | 73   | 17   | 77   | 27   |
| acc   | 0.78 | 0.55 | 0.78 | 0.81 | 0.76 |

4 sec chunk gender classification: 0.74 ± 0.11

|       | F1   | F2   | F3   | F4   | F5   |
|-------|------|------|------|------|------|
| M     | 71   | 15   | 31   | 57   | 90   |
| F     | 21   | 71   | 27   | 67   | 24   |
| acc   | 0.79 | 0.58 | 0.86 | 0.80 | 0.78 |

5 sec chunk gender classification: 0.76 ± 0.11

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

| Frames | Accuracy |
|--------|----------|
| 10     | 0.73     |
| 50     | 0.75     |
| 100    | 0.76     |
| 200    | 0.77     |
| 300    | 0.77     |

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

|       | F1   | F2   | F3   | F4   | F5   |
|-------|------|------|------|------|------|
| M     | 71   | 16   | 54   | 57   | 79   |
| F     | 87   | 67   | 22   | 76   | 24   |
| acc   | 0.85 | 0.62 | 0.83 | 0.85 | 0.85 |

Random 100 Frames gender classification: 0.80 ± 0.10

CNN LSTM model can learn the gender-specific spectral cues from the 52 MFCC statistics to classify gender using breath cycle audios.

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 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 and 5sec chunks are doing better than other chunks. This experiment indicates automatic gender classification can be done on chunks other than the breath cycle boundary with substantial accuracy which can motivate the development of an unsupervised model for gender classification using vocal breath sound signals.

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 the accuracy increases as more frames are chosen and then becomes constant after 200 frames. This shows that reducing the number of frames in a chunk reduced the prediction accuracy of the model because frame-level data is important for calculating the 52 MFCC statistics to capture the spectral cues for gender classification, hence close to no redundant frames in the 3-sec chunk.

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 performs 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.

4. DISCUSSION

In this section, we analyze the results of the experiments mentioned in Section 3 to answer the five questions we defined at the start of the study.

1) Are gender cues present in the mel-spectrogram image 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. This substantiates 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

5. CONCLUSIONS

To conclude the studies presented in this paper, we can say that gender information can be extracted from the vocal breath sound signals using data-driven and knowledge-based features with considerable accuracy. We can also conclude that the gender classification model trained on long-term features(features calculated over longer durations of the breath audio) performs better than short-term features extracted from breath sound signals. The accuracy of the work presented in this paper can be further improved by increasing the breath dataset. An unsupervised gender classification framework can be built using the features presented in the paper.

6. ACKNOWLEDGMENTS

We would like to thank the Department of Science and Technology(DST), Govt of India for sponsoring the data collection drive and the project and Dr. Shivani Yadav for her efforts on data collection without which this study could not have been conducted.
7. REFERENCES

[1] Noam Gavriely, YORAM Palti, and Gideon Alroy, “Spectral characteristics of normal breath sounds,” Journal of applied physiology, vol. 50, no. 2, pp. 307–314, 1981.

[2] Bahar Khalighinejad, Guilherme Cruzatto da Silva, and Nima Mesgarani, “Dynamic encoding of acoustic features in neural responses to continuous speech,” Journal of Neuroscience, vol. 37, no. 8, pp. 2176–2185, 2017.

[3] Jackson Liscombe, Jennifer Venditti, and Julia Bell Hirschberg, “Classifying subject ratings of emotional speech using acoustic features,” 2003.

[4] Gang Chen, Xue Feng, Yen-Liang Shue, and Abeer 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] Christine M Sapienza, “Aerodynamic and acoustic characteristics of the adult africanamerican voice,” Journal of Voice, vol. 11, no. 4, pp. 410–416, 1997.

[6] Richard J Morris, WS Brown Jr, Douglas M Hicks, and Elizabeth Howell, “Phonational profiles of male trained singers and nonsingers,” Journal of Voice, vol. 9, no. 2, pp. 142–148, 1995.

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

[8] Buket D Barkana and Jingcheng 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] Yakun Hu, Daopeng Wu, and Antonio Nucci, “Pitch-based gender identification with two-stage classification,” Security and Communication Networks, vol. 5, no. 2, pp. 211–225, 2012.

[10] Yu-Min Zeng, Zhen-Yang Wu, Tiago Falk, and Wai-Yip Chan, “Robust gmm based gender classification using pitch and rastapl parameters of speech,” in International conference on machine learning and cybernetics. IEEE, 2006, pp. 3376–3379.

[11] Zakariya Qawaqneh, Arafat Abu Mallouh, and Buket 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] Sarah Ita Levitan, Taniya Mishra, and Srinivas Bangalore, “Automatic identification of gender from speech,” in Proceeding of speech prosody, Semantic Scholar, 2016, pp. 84–88.

[13] Selen Hande Kabil, Hannah Muckenhirn, and Mathew Magimai-Doss, “On learning to identify genders from raw speech signal using cnns.,” in Interspeech, 2018, pp. 287–291.

[14] Rangga Dwi Alamsyah and Suyanto 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] E. Priya, Janani Priyadharshini S, Padam Satya Reshma, and Sashank S, “Temporal and spectral features based gender recognition from audio signals,” in 2022 International Conference on Communication, Computing and Internet of Things (IC3IoT), 2022, pp. 1–5.

[16] Shivani Yadav, Kausthubha NK, Dipanjan Gope, Uma Maheswari Krishnaswamy, and Prasanta Kumar 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.

[17] Shivani Yadav, Merugu Keerthana, Dipanjan Gope, Uma Maheswari K., and Prasanta 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.

[18] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, “librosa: Audio and music signal analysis in python,” in Proceedings of the 14th python in science conference, 2015, vol. 8.