A Study on Feature Selection for Gender Detection in Speech Processing for Assamese Language

Kankana Dutta(✉), Rizwan Rehman, Priyakshi Mahanta, and Ankumon Sarmah

Centre For Computer Science and Applications, Dibrugarh University, Dibrugarh, India
{kankanadutta,rizwan,priyakshimahanta,ankumonsarmah}@dibru.ac.in

Abstract. Gender identification is an integral part of a Speech recognition system. Specifically, for the low resource languages, it is a challenging task. For any speech recognition system, finding a suitable feature plays an essential role in the system's performance. In this paper, we have done a comparative analysis of gender identification from formant frequencies F1 and F2 of speech data set collected from the speakers of Assamese language (a low recourse language of North-East India). The objective is to explore different classification techniques for developing a gender identification module for Assamese language. We have used four supervised classification techniques kNN, Logistic Regression, decision tree, and SVM, and found that when F1 and F2 are used together, the methods give the best result. One unsupervised method Gaussian Mixture Model (GMM) is also applied and found that the best result is given by formant frequency F1.

Keywords: Gender identification · Speech processing system · Formant frequency · Supervised learning method · Un-supervised learning method

1 Introduction

Speech processing-related research has gained much attention in the last few decades. Also, people nowadays prefer to use speech recognition systems which are helpful in many ways. A speech signal is an acoustic wave that can provide different information related to the speaker and the language being spoken. The information regarding the speaker may include age, gender, speaker identity, the emotional status of the speaker, and many more. Different parameters or features are collected from the speech signal, and different methods are applied to find the required information. These features collected from speech signals are fundamental frequency, also referred to as pitch and format frequencies. Formant frequency refers to the resonance frequencies of the vocal tract. The different types of formant frequencies are F1, F2, F3, and F4. Vowel formant frequencies are the most used voice features in the field of speech processing research. Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC), Line Spectral Frequencies (LSF), Discrete Wavelet Transform (DWT), and Perceptual Linear Prediction (PLP) are other speech

C. Badica et al. (Eds.): ICICCT 2022, CCIS 1670, pp. 73–82, 2022.
https://doi.org/10.1007/978-3-031-20977-2_6

feature extraction techniques which are also used widely in the field of speech processing to extract information or features from the speech signal. Identifying the most salient features and analyzing and finding the desired information from a large set of data is a challenging task. Different machine learning techniques, which help analyze a large set of data, maybe applied and construct a classification model for a particular task.

Assamese language speaker is found mainly in Assam, which is a state in India, and more than 15 million people speak this language. Assam is a state of multi-lingual and multi-ethnic people. Apart from the Assamese language, tribal language speakers such as Bodo, Mising, Rabha, Karbi, and Dimasa are found in different districts of Assam. Other languages like Nepali, Hindi, etc. are also spoken by a large number of the population which are spread all over the state. In the last few years, some researchers have been working in the field of speech processing and trying to develop different applications for Assamese as well as other regional languages spoken in Assam.

Gender Identification is a part of a speech recognition system that finds the gender of a speaker from the speech information extracted from the speech signal. The information about gender can be helpful in speaker recognition systems which are used in different security systems today.

1.1 Literature Review

This review covers the work related to speech processing research done for Assamese and other tribal languages of Assam and some recent works done for other languages.

Talukdar et al. [1] analyzed Cepstral features and formant frequencies of Bodo and Rabha phonemes and words using LPC and observed that Cepstral coefficients of Bodo and Rabha vowels had shown distinctive characteristics for male and female speakers. The variation of the cepstral coefficients for males is very irregular, and the same for the female is stable.

An automatic Language Identification (LID) system was developed for the languages Bodo, Dimasa, Rabha, and Tiwa, which belong to the same language subfamily, and identifies the language from a short sound file [2]. The system was implemented using a Gaussian mixture model with Mel-Frequency Cepstral Coefficients (MFCCs) as features and showed 92.7% accuracy when the duration of the speech file is 3 s.

An automatic Assamese vowel recognition system from spoken Assamese words [3] was developed by P Sharma *et al.* using Support Vector Machine (SVM), K-Nearest Neighbor (kNN), and Random Forest (RF) classifier and found that Random Forest provides better recognition rate than other two classifiers.

Mridusmita Sharma and Kandarpa Kumar Sarma [4] used Recurrent Neural Network (RNN) based algorithm and k-Nearest Neighbor (kNN) based algorithm for the recognition of the vowels of the Assamese language for four major dialects of the language. The kNN based approach gave a better recognition rate than ANN.

A study on Missing language vowels was performed by Rehman R. *et al.* using the Fisher score algorithm [5] to find the most distinctive feature of speech data for gender classification. The result is cross-validated with a Tree-based algorithm and found that fundamental formant frequencies (F0) are the best parameter among all the other parameters.

A study on different attributes of speech signals like time, pitch, formant frequencies, and speaker type was done on Missing language vowels by Saikia U. *et al.* in 2019 [6] using a regression model and found that fundamental formant frequencies (F0), i.e. pitch, varies for the vowels with the gender of the speaker and this information can be utilized for speaker identification in Mising language. A logistic regression model is built and applied on pitch value (F0) to detect the gender of the speaker [7] and can detect the gender.

An automatic transcription model of the Assamese language [8] was developed by Sarma, H., *et al.* to represent the speech sound with a symbol which is known as phonetic transcription. The experiment was done using Hidden Markov Mode Tool Kit (HTK).

Sarma H. *et al.* worked on different aspects of speech-to-text processing using HTK [9] and developed an automatic syllabification model for the Assamese language to find the syllables of a word automatically and found 12 different syllable patterns where five are found most frequent.

An automatic speech recognition system for Assamese was implemented using Kaldi Toolkit for continuously spoken Assamese [10] by Deka *et al.,* and experiments were conducted to explore the effect of excluding low-quality speech files from the training set on the performance of the system.

Basu J, *et al.* developed a multi-lingual speech corpus for Low-Resource Eastern and North-Eastern Indian Languages for Speaker and Language Identification [11] and found the Best performance of 94.49% (using MFCC (Mel frequency cepstral coefficients) + SDC(shifted delta cepstral) feature and LSTM-RNN) in Speaker identification and 95.69% (using MFCC + SDC feature and LSTM-RNN) in Language identification for short duration speech files.

Analysis of the first three formant frequencies was presented by Dr. Bhargab Medhi using the LPC model and observed that the variation of F1and F2 for a different vowel is quite distinct, formant values of a female speaker are comparatively high than the male speaker, and the third formant frequency F3 does not play a crucial role in the identification of a specific vowel spectrum [12].

Yücesoy, E proposed a model for gender and age group classification using MFCC and delta coefficients of speech values applied in the Gaussian Mixer model [13]. A test has been carried out to find the minimum number of GMM components.

A gender recognition system was developed by Kumar P. *et al.* [14] where the front end extracted speech signal features such as energy, ZCR, MFCC, and Entropy, and the back end classified the speaker according to gender using GMM.

A speaker recognition system was developed by Gupta M. *et al.* [15], with a two-level approach where first the gender will be identified, and then the speaker is recognized. MFCC and pitch values are used for gender recognition with the SVM classifier, and MFCC, Pitch, and RASTA-PLP features are used for speaker recognition using GMM.

Alkhawaldeh R.S discusses and uses three prominent feature selection techniques EA, PSO, and wolf techniques to find the optimal features from MFCCs, Chroma, Mel, Contrast, and Tonnetz for improving the result of classification techniques [16].

1.2 Motivation and Contribution

From the literature review, it is observed that several experiments have been carried out to evaluate experiments on Gender, language, and speech identification in Assamese language and other regional languages of Assam using different classification models. It is also observed that the most used feature type extracted from the audio file was MFCC to identify the speaker. It is also used for the speech recognition process. In the case of gender recognition, mostly fundamental frequency (F0) and formant frequencies (F1–F4) of speech signals were used. Different approaches and classifiers such as SVM, KNN, RNN, Tree-based model, and Regression model were used on formant frequencies in gender recognition. The first two formants i.e. F1 and F2 show distinct features in the case of male and female speakers of the languages Assamese, Bodo, Rabha, and Mising [1, 5–7, 12].

Based on these observations, we are performing a study on Assamese speech data using different prominent classification models. The aim is to study the impact of the formant frequencies F1 and F2 in the case of gender identification. We have only considered F1 and F2 because the main information of a speech signal is concentrated in the low-frequency part of a speech signal. Thus the objective of this study can be defined as

1. To study the performance of different supervised machine learning methods on the formant frequencies F1 and F2.
2. To study the performance of these formant frequency values in the Gaussian Mixture Model (GMM), which is an unsupervised learning method, and
3. To determine which of these formant frequencies show the better result in gender identification for the Assamese language.

2 Methodology

This section describes the dataset and methods used to experiment.

2.1 Dataset

The dataset contains data from both male and female speakers. There are total 23 numbers of speakers, out of which 15 are male, and 8 are female. A speech sample from each speaker is collected speaking the Assamese vowels. In the Assamese language, there is a total of 11 vowels, and all these vowels are spoken by the speakers. There is a total of 4822 speech samples present in the dataset. The data are recorded in a noise-free environment. For the current study, we have used only two formant frequencies, F1 and F2, of the speech samples. These formant frequencies are extracted by using PRAAT software.

The features are plotted in a graph for both males and females, as shown in Fig. 1 and Fig. 2, respectively.

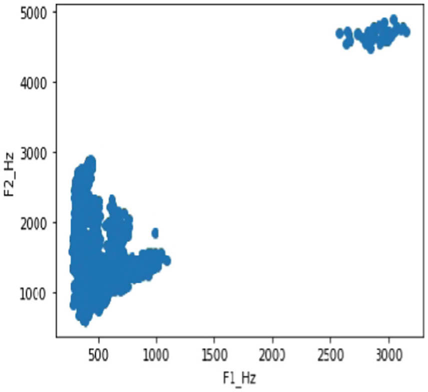


Fig. 1. F1 and F2 features for females

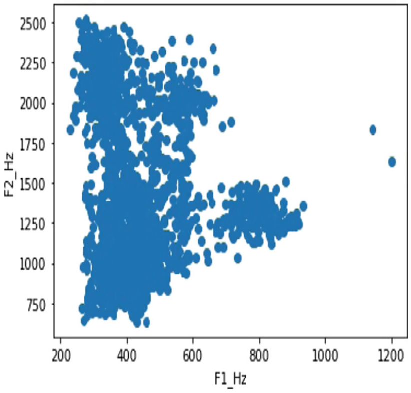


Fig. 2. F1 and F2 features for male

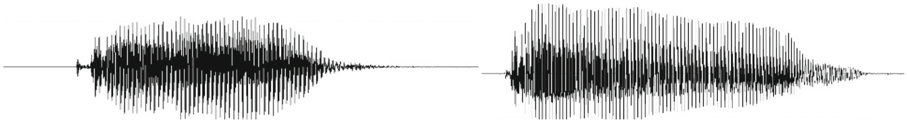
2.2 Methods

For analyzing the data, we are using different prominent machine learning methods. Two types of machine learning methods are used here- supervised and unsupervised methods.

In supervised learning, the machine is trained with some known data. The data used to train the model are already labelled with the correct class. The machine is trained with labelled data, and whenever some new information is obtained, it calculates the answer with the trained data. In this experiment, four prominent supervised learning methods are being used, viz., k Nearest Neighbour (kNN), Support Vector Machine (SVM), Logistic Regression, and Decision Tree method. On the other hand, in the unsupervised learning method, the data does not have any label to help in the classification method; it is designed to find the label on its own. Unsupervised methods are helpful in complex processing tasks compared to supervised methods. In this experiment, one unsupervised method is being used, Gaussian Mixture Model (GMM). GMM is a probabilistic model which represents the clusters as a weighted sum of Gaussian component densities. This GMM model is used in many experiments performed on Assamese as well as other regional languages and found that GMM shows promising results in these languages.

3 Results and Discussion

The speech spectrum contains different information about the speaker. For this study of gender recognition, only f1 and f2 formant frequencies are considered. Figure 3 shows two voice samples of male and female speakers speaking the first Assamese vowel.



a) Male voice sample b)Female voice sample

Fig. 3. Voice sample of male and female speakers

Figure 4 represents F1 and F2 features for male and female speakers both speaking the first Assamese vowel. Green and blue dots represent formant frequency for male-female speakers, respectively. This figure shows a clear separation of F1 and F2 values for male and female speakers.

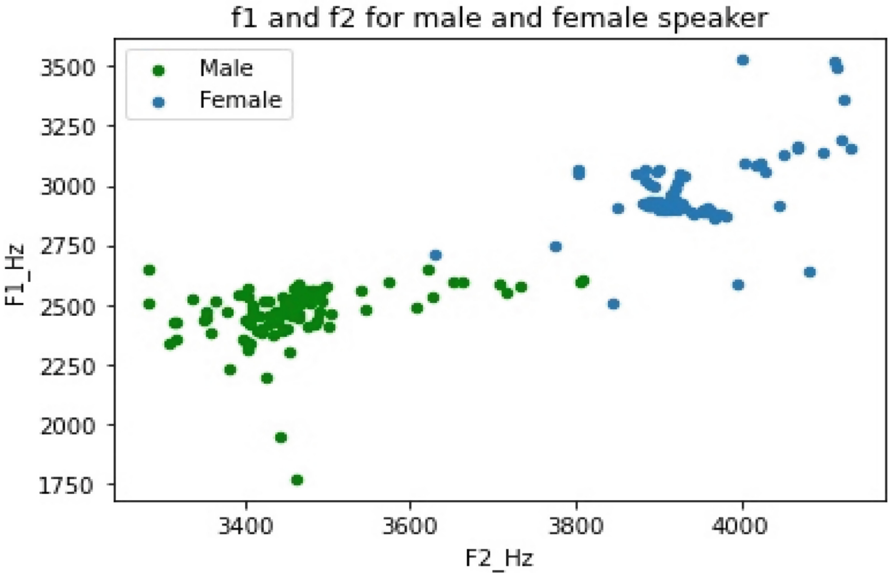


Fig. 4. F1 and F2 representation of male and female speakers

Table 1 represents some data from the dataset. It consists of formant values of F1 and F2 along with the gender.

After extracting the speech features F1 and F2 from the speech signals, KNN, Logistic Regression, SVM, and Decision Tree methods are applied to these features. These methods are supervised machine learning algorithms. Data processing is carried out using Python. First, these classification methods are applied using the formant frequency feature F1 and then using the feature F2. Also, classifiers are applied using both the features F1 and F2 together. The database is divided into 80% training data and 20% test data. Table 2 shows the experiment results.

Table 1. Some example data from the dataset

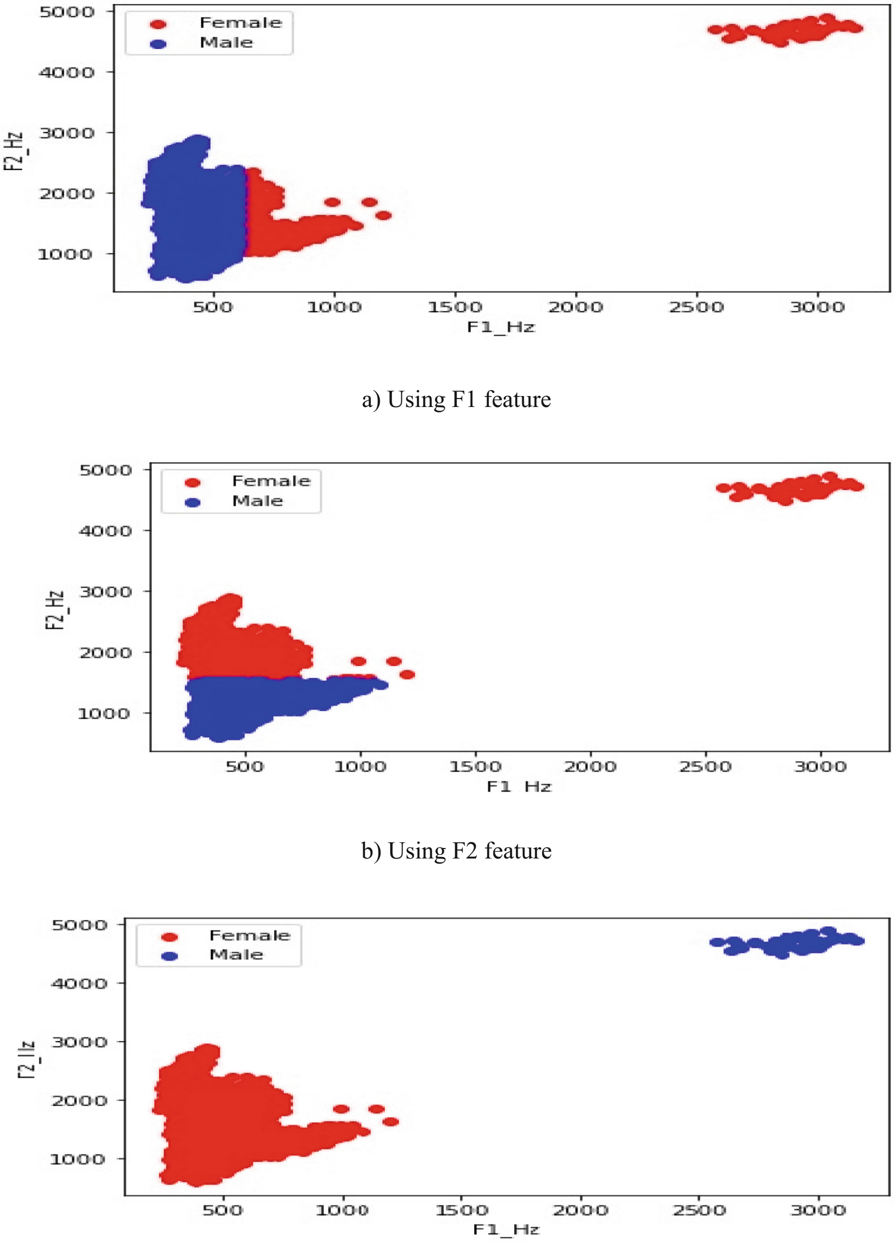
Sl. No	F1_Hz	F2_HZ	Gender
1	904.2219	1435.993	Female
2	895.4748	1396.962	Female
3	855.5742	1386.76	Female
4	910.6128	1423.129	Female
5	884.541	1429.456	Female
6	858.5995	1441.599	Female
7	288.3647	1911.452	Male
8	277.6312	2033.321	Male
9	321.0209	1842.203	Male
10	245.2099	1963.462	Male
11	243.5955	1894.786	Male
12	367.2974	2013.839	Male

Table 2. Rate of recognition using formant frequencies

Formant frequencies	Recognition rate			
	kNN	Logistic regression	SVM	Decision tree
F1	58.34	57.36	54.25	56.59
F2	55.85	52.64	50.77	52.09
F1, F2	69.22	59.79	59.38	62.79

It is evident from the table that formant frequency F2 alone shows the worst result, and the best result is obtained by using both the formants frequencies F1 and F2 together. Among these methods, the kNN method gives the best answer when F1 and F2 are used together.

Next, we have used Gaussian Mixer Model (GMM), which is are unsupervised clustering technique. Since it is an unsupervised method, the model does not have any information about the number of clusters and data labels. Since we are classifying the data into male and female, we are applying the GMM model for creating two clusters to represent males and females. First, we have applied the model for classifying the data using the formant frequencies F1. Then we have applied the model on formant frequency F2, and finally, the model is applied using both F1 and F2 together. The graphical representation of data according to their gender plotted in a graph is shown in Fig. 5.



a) Using F1 feature

b) Using F2 feature

c) Using both F1 and F2

Fig. 5. (a, b, c). Clustering result by GMM

The accuracy achieved for the method is calculated for all the cases, and also precision, recall, and f1-score values are calculated and shown in Table 3.

Table 3. Error analysis of gmm method

Feature name	Precision	Recall	F1-score	Accuracy
F1	52.21	88.68	65.72	54.58%
F2	49.00	66.21	56.2	49.56%
F1, F2	0	88.68	0	50.06%

From the above results, it is seen that the model provides the best F1 score and accuracy value with the feature F1. It is ranked the attributes as F1, (F1, F2), and F3. This shows a difference in results with supervised methods where the attributes are ranked as (F1, F2), F1, and F2 according to their performance. A similar result is found in both supervised and unsupervised methods, where the F2 value gives the worst accuracy result.

4 Conclusion

In this comparative study, we have tried to find a formant frequency value between F1 and F2, which works better for gender recognition using Assamese vowels. For that, we have compared the results of supervised machine learning methods kNN, SVM, Decision Tree, and Logistic regression classifier and found that when F1 and F2 are used together, it gives the best result, and F2 gives the worst result. Again the same experiment is performed using the unsupervised method GMM and found that F1 is giving the best result. F2 is giving the worst result, which is the same in the case of supervised methods also. This experiment is done using only two formant frequency values. The accuracy rate is not very high, which suggests that these two formant frequencies are not sufficient to detect gender. More experiments can be done to find whether the performance improves when F1 and F2 are used with the other two formant frequency features, F3, F4, or with the MFCC features.

References

1. Talukdar, J., Pathak, N.: Acoustic representation of BODO and RABHA phonemes. Int. J. Comput. Commun. Network. **1**(1) (2012)
2. Chakraborty, J., Sarmah, P., Vijaya, S.: Spoken language identification of four Tibet-Burman languages, Oriental COCOSDA (2020)
3. Sarma, P., Mitra, M., Bhuyan, M.P., Deka, V., Sarmah, S., Sarma, S.K.: Automatic vowel recognition from Assamese spoken words. IJITEE **8**(10) (2019)
4. Sharma, M., Sarma, K.K.: Dialectal Assamese vowel speech detection using acoustic-phonetic features, KNN and RNN. In: 2nd International Conference on Signal Processing & Integrated Networks (2015)

5. Rehman, R., Bordoloi, K., Dutta, K., Borah, N., Mahanta, P.: Feature selection and classification of speech dataset for gender identification: a machine learning approach. J. Theoretical Appl. Inf. Technol. **99** (2020)
6. Saikia, U., Rehman, R., Hazarika, J., Hazarika, G.C.: Predictive analysis using regression methods in low resource language "*MISING*". In: 2nd International Conference on Information Systems & Management Science (ISMS) (2019)
7. Saikia, U., Hazarika, J.: Analysis of speech signal data of Mising vowels using logistic regression and k-Means clustering. Int. J. Adv. Comput. Sci. Appl. 12(4) (2021)
8. Sarma, H., Saharia, N., Sharma, U.: Development of Assamese speech corpus and automatic transcription using HTK. Advances in Signal Processing and Intelligent Recognition Systems, Advances in Intelligent Systems and Computing (2014)
9. Sarma, H., Saharia, N., Sharma, U.: Development and analysis of speech recognition systems for Assamese language using HTK. ACM Trans. Asian Low-Resour. Lang. Inf. Process. **17**(1)(Article 7) (2017)
10. Deka, B., Sarmah, P., Vijaya, S.: Assamese database and speech recognition, IEEE (2019)
11. Basu, J., Khan, S., Roy, R., Basu, T.K., Majumder, S.: Multilingual speech corpus in low-resource eastern and northeastern indian languages for speaker and language identification. Circuits Syst. Signal Process. **40**(10), 4986–5013 (2021). https://doi.org/10.1007/s00034-021-01704-x
12. Medhi, B.: Analysis of formant frequency F1, F2, and F3 in Assamese vowel phonemes using LPC model. Int. J. Eng. Res. Technol. **6**(05) (2017)
13. Yücesoy, E.: Speaker age and gender classification using GMM super vector and NAP channel compensation method. J. Ambient Intell. Human. Comput. **13**, 3633–3642 (2020)
14. Kumar, P., Baheti, P., Jha, R.K., Sarmah, P., Sathish, K.: Voice gender detection using Gaussian Mixture Model. J. Network Commun. Emerg. Technol. **8**(4) (2018). www.jncet.org
15. Gupta, M., Bhartiy, S.S., Agarwaly, S.: Gender-Based Speaker Recognition from Speech Signals Using GMM Model. Modern Physics Letters, World Scientific Publishing Company (2019)
16. Alkhawaldeh, R.S.: DGR: Gender Recognition Of Human Speech Using One-Dimensional Conventional Neural Network. Hindawi Scientific Programming, Vol. 2019, Article ID 7213717 (2019)